Author Query Form

Dear Author,

During the preparation of your manuscript for typesetting, some questions have arisen. These are listed below.

Queries and/or remarks

Location in article (page/line)	Query / remark	Response
27/1317	Please cite or delete the uncited references. You will find a special note in the margin by each uncited reference item.	

Many thanks for your assistance

Metadata of the article that will be visualized in Online First

Please note: Images will appear in color online but will be printed in black and white.

Journal Name	Machine Learning	
Article Title	Efficiently learning the preferences of people	
Copyright holder	The Author(s)	
	This will be the copyright line in the final PDF.	

Corresponding Author	Family name	Birlutiu
	Particle	
	Given Name	Adriana
	Suffix	
	Division	Institute for Computing and Information Sciences
	Organization	Radboud University Nijmegen
	Address	Heyendaalseweg 135, 6525, AJ Nijmegen, The Netherlands
	Division	
	Organization	Romanian Institute of Science and Technology
	Address	Str. Ciresilor 29, 400487, Cluj-Napoca, Romania
	E-mail	A.Birlutiu@science.ru.nl
Author	Family name	Groot
	Particle	
	Given Name	Perry
	Suffix	
	Division	Institute for Computing and Information Sciences
	Organization	Radboud University Nijmegen
	Address	Heyendaalseweg 135, 6525, AJ Nijmegen, The Netherlands
	Division	Dept. of Electrical Engineering, Control Systems
	Organization	Technical University Eindhoven
	Address	Potentiaal 4.28, 5600, MB Eindhoven, The Netherlands
	E-mail	Perry.Groot@science.ru.nl
Author	Family name	Heskes
	Particle	
	Given Name	Tom
	Suffix	
	Division	Institute for Computing and Information Sciences
	Organization	Radboud University Nijmegen

Abstract	This paper presents a framework for optimizing the preference learning process. In many real-world applications in which preference learning is involved the available training data is scarce and obtaining labeled training data is expensive. Fortunately in many of the preference learning situations data is available from multiple subjects. We use the multi-task formalism to enhance the individual training data by making use of the preference information learned from other subjects. Furthermore, since obtaining labels is expensive, we optimally choose which data to ask a subject for labelling to obtain the most of information about her/his preferences. This paradigm—called active learning—has hardly been studied in a multi-task formalism. We propose an alternative for the standard criteria in active learning which actively chooses queries by making use of the available preference data from other subjects. The advantage of this alternative is the reduced computation costs and reduced time subjects are involved. We validate empirically our approach on three real-world data sets involving the preferences of people.

Efficiently learning the preferences of people

Adriana Birlutiu · Perry Groot · Tom Heskes

Abstract This paper presents a framework for optimizing the preference learning process. In many real-world applications in which preference learning is involved the available training data is scarce and obtaining labeled training data is expensive. Fortunately in many of the preference learning situations data is available from multiple subjects. We use the multi-task formalism to enhance the individual training data by making use of the preference information learned from other subjects. Furthermore, since obtaining labels is expensive, we optimally choose which data to ask a subject for labelling to obtain the most of information about her/his preferences. This paradigm—called active learning—has hardly been studied in a multi-task formalism. We propose an alternative for the standard criteria in active learning which actively chooses queries by making use of the available preference data from other subjects. The advantage of this alternative is the reduced computation costs and reduced time subjects are involved. We validate empirically our approach on three real-world data sets involving the preferences of people.

Keywords Learning preferences · Active learning · Experimental design · Multi-task learning · Hierarchical modeling

Editor: James Cussens.

A. Birlutiu (✉) · P. Groot · T. Heskes
Institute for Computing and Information Sciences, Radboud University Nijmegen,
Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
e-mail: A.Birlutiu@science.ru.nl

P. Groot
e-mail: Perry.Groot@science.ru.nl

T. Heskes
e-mail: T.Heskes@science.ru.nl

A. Birlutiu
Romanian Institute of Science and Technology, Str. Ciresilor 29, 400487 Cluj-Napoca, Romania

P. Groot
Dept. of Electrical Engineering, Control Systems, Technical University Eindhoven, Potentiaal 4.28,
5600 MB Eindhoven, The Netherlands

Ⓐ Springer

PDF-OUTPUT

1 Introduction

There has been an increasing interest recently in learning the preferences of people within artificial intelligence research (Doyle 2004). Preference learning provides the means for modeling and predicting people's desires and this makes it a crucial aspect in modern applications such as decision support systems (Chajewska et al. 2000), recommender systems (Blythe 2002; Blei et al. 2003), and personalized devices (Clyde et al. 1993; Heskes and de Vries 2005).

A prototypical example for an application of preference learning that we will use in this paper is fitting hearing-aids, i.e., tuning of hearing-aid parameters so as to maximize user satisfaction. This is a complex task, due to three reasons: (1) high dimensionality of the parameter space, (2) the determinants of hearing-impaired user satisfaction are unknown, and (3) the evaluation of this satisfaction through listening tests is costly (in terms of patient burden and clinical time investment) and unreliable (due to inconsistent responses). The last point illustrates an important issue that preference learning has to address, which is the limited availability of labeled data used for model training. Obtaining appropriate training data in preference learning applications requires time and effort from the modeled user. This shortcoming can be addressed by taking advantage of two characteristics of the settings in which preference learning is usually applied. First, the training data is mostly acquired through interactions with the modeled user; and, second, preferences are modeled for multiple users, as a result multiple training data sets are available. In order for the preference learning methods to be implemented in real-world systems, they must be capable of exploiting all possible sources of information and in the most efficient way.

In most of the situations in which preference learning is involved data is available from multiple subjects. Thus, even though individual data is scarce and difficult to obtain, we can optimize the learning of preferences of a new subject by making use of the available data from other subjects. Learning in this setting is well-known as multi-task or hierarchical learning and has been studied extensively in recent years in machine learning. By using the multi-task formalism, the preference data collected for other subjects can be gathered and used as prior information when learning the preferences of a new subject. Furthermore, to deal with the fact that obtaining labeled data is expensive we can speed up learning by optimally choosing the examples to be queried. At each learning step we can decide which example gives the most information about the subject's preferences. This paradigm, called active learning in the machine learning literature and related to sequential experimental design in statistics, has been studied extensively, but hardly in the multi-task setting.

The aim of this work is to present an efficient framework for optimizing the preference learning process. This framework considers the combination between active learning and multi-task learning in the preference learning context. The contribution of this work is a criterion for active learning designed for the multi-task setting. The advantages of this criterion are in its interpretation and the ease in computability.

The structure of this paper is as follows. First, this section ends by presenting related work on preference learning, multi-task learning and active learning. In Sect. 2 we describe the learning framework. We consider learning from qualitative preference observations in which the subject makes a choice for one of the presented alternatives. This can be modeled using the probabilistic choice models introduced in Sect. 2.1. Learning a utility function representing the preferences of a subject from this type of preference observations is described in Sect. 2.2. Learning the utility function in a multi-task setting by making use of the data available from other subjects is considered in Sect. 2.3. In Sect. 3 we present several criteria for selecting the most informative experiments with respect to a subject's preferences. After

 Springer

Mach Learn

reviewing some of the standard criteria from experimental design, we propose an alternative criterion which makes use of the preference observations collected already from a community of subjects. We show that this alternative criterion is connected to the standard criteria from experimental design. In Sect. 4 we demonstrate experimentally the usefulness of our framework on three data sets, a subset of the Letor data set, an audiological data set and a data set about people's preferences for art images. In Sect. 5 we present several conclusions and discuss directions for future research.

1.1 Background and related work

In this section we review some studies from preference learning, multi-task learning, and active learning related to the work presented in this paper.

1.1.1 Preference learning

Preference learning is the creation of a model from collected data which can be used to model and predict people's desires. A very recent and complete overview about preference learning is given in (Fürnkranz and Hüllermeier 2010). There are different approaches to preference learning which can be categorized according to the learning task, the learning technique and the application area. We will briefly enumerate them and state in which category our current work can be included (for more details we refer the interested reader to Fürnkranz and Hüllermeier 2010).

1. Based on the application area, preference learning approaches can be divided into the following main groups: (i) applied to the field of information retrieval, e.g., learning to rank search results of a query or a search engine, (ii) applied to recommender systems, e.g., used by online stores to recommend products to their customers, or for personalized devices, and (iii) bipartite ranking and label ranking, which find applications in disciplines such as medicine and biology. The application scenarios that we consider in the experimental evaluation in Sect. 4 belong to information retrieval (the Letor data set) and recommender systems (the audiological and art data sets).
2. The learning technique divides the preference learning approaches into four categories: (i) learn a binary preference relation that compares pairs of alternatives, (ii) model-based approach that aims at identifying the preference relation by making sufficiently restrictive model assumptions, (iii) local estimation techniques which lead to aggregating preferences, and (iv) learning utility functions by using regression to map instances to target valuations for direct ranking. We focus on the latter approach and use a utility function in order to model a subject's preferences. The utility function is learned in a Bayesian framework.
3. The learning task includes label, instance, and object ranking. Label ranking can be seen as a generalization of classification where a complete ranking of labels is associated with an instance instead of only a class label. Instance ranking can be seen as a generalization of ordinal classification where an instance belongs to one among a finite set of classes and the classes have an order. The setting of object ranking has the peculiarity of having no supervision in the sense that no class label is associated with an object. Instead, a finite set of pairwise preferences or other ordering between objects is given. The setting that we consider in this work belongs to the last category.

In many preference learning settings it is important to take into account the context, i.e., context-aware preference learning (Adomavicius et al. 2005). The motivation for context

« MACH 10994   layout: Small Condensed v.1.1   file: mach5297.tex (Ramune)   class: spr-small-v1.3 v.2012/04/18   Prn:2012/05/21; 15:39   p. 4/28»
«   doctopic: OriginalPaper   numbering style: ContentOnly   reference style: apa»

Mach Learn

aware preference learning is that the same subject/user/consumer may use different decision-making strategies and prefer different products under different contexts. For hearing-aid fitting, which is one of the application scenarios that we use in this work, this means that a user would prefer a certain setting of the hearing-aid parameters if he is listening to a concert and another setting if he is in a discussion. In general for context aware preferences bigger data sets are needed, as preferences would have to be learned for all contextual situations. The approach that we present in this paper can be applicable in this setting as well. While this is an interesting, related topic, it is beyond the scope of the current work.

1.1.2 Multi-task learning

The idea behind multi-task learning is to utilize labeled data from other "similar" learning tasks in order to improve the performance on a target task. It is inspired by the research on transfer of learning in psychology, more specifically on the dependency of human learning on prior experience. For example, the abilities acquired while learning to walk presumably apply when one learns to run, and knowledge gained while learning to recognize cars could apply when recognizing trucks. The initial foundations for multi-task learning were laid by (Thrun 1995; Caruana 1997). The psychological theory of transfer of learning implies the similarity between tasks. In a related way, the multi-task learning assumes similarity between models of different tasks. For example, (Evgeniou et al. 2005; Argyriou et al. 2008) exploit similarity between the deterministic parts of the models by means of regularization, with the effect of improvement in performance. In this work we implement multi-task learning using a Bayesian approach. The Bayesian approach to multi-task learning assumes the parameters of individual models to be drawn from the same prior distribution. Examples of the Bayesian approach to multi-task learning are (Bakker and Heskes 2003) where a mixture of Gaussians is used for the top of the hierarchy. This leads to clustering the tasks, one cluster for each Gaussian in the mixture. In (Yu et al. 2005; Birlutiu et al. 2009) a hierarchical Gaussian Process is derived with a normal-inverse Wishart distribution used at the top of the hierarchy.

1.1.3 Active learning

Active learning, also known in the statistics literature as sequential experimental design, is suitable for situations in which labeling points is difficult, time-consuming, and expensive. The idea behind active learning is that by optimal selection of the training points a better performance can be achieved instead of random selection. The scenarios in which active learning can be applied belong to one of the following three categories: (i) generating *de novo* points for labeling; (ii) stream-based active learning where the learner decides whether to request the label of a given instance or not; (iii) pool-based active learning where queries are selected from a large pool of unlabeled data. In this work we consider the pool-based active learning setting.

Methods for active learning can be roughly divided into two categories: those with and without an explicitly defined objective function. Uncertainty sampling (Lewis and Gale 1994), Query-by-Committee (Seung et al. 1992; Freund et al. 1997) and variants thereof belong to the latter category. They are based on the idea of selecting the most uncertain data given the previously trained models. The methods with an explicit objective function are often motivated by the theory of experimental design (Fedorov 1972; Chaloner and Verdinelli 1995; Schein and Ungar 2007; Lewi et al. 2009; Dror and Steinberg 2008). The objective function quantifies the expected gain of labeling a particular input, for example

🙋 Springer

Mach Learn

in terms of the expected reduction in the entropy of the model parameters (MacKay 1992; Cohn et al. 1996). With respect to the performance of the two categories of methods, Schein and Ungar (2007) show that the methods from the second approach perform better but are computationally more expensive due to re-training the models for each candidate point. A trend is to improve the performance of the active learning methods by combining them with heuristics designed either for the context in which they are applied or by the models they use, e.g., making use of the unlabeled data available (McCallum and Nigam 1998; Yu et al. 2006), exploiting the clusters in the data (Dasgupta and Hsu 2008), diversifying the set of hypotheses (Melville and Mooney 2004), or adapting the active learning to Gaussian processes (Chu and Ghahramani 2005a; Brochu et al. 2008).

Preference learning can benefit from the active learning paradigm. In most of the preference learning settings labels are given by people in an explicit way. This means that for acquiring training preference data, the subjects have to interact with the system, and they need to express their preferences explicitly. These situations appear when it is impossible or insufficient to implicitly collect training preference data. For example, when learning preferences in a live system where subjects choose electronically their favorite movie, labelling is done automatically by the selection, but, for other scenarios, like for example, fitting hearing-aids, this implicit way of collecting training data cannot be applied. In these situations, it makes sense to use active learning in order to collect the most informative data. There are severals studies in the literature that use active learning in a preference learning setting. Brinker (2004) presents some extensions of pool-based active learning to label ranking problems; Xu et al. (2010) address the problem of preference learning using relational models between items; Guo and Sanner (2010) investigate active preference learning for real-time systems; Brochu et al. (2008) propose a criterion for active learning that maximizes the expected improvement at each query without accurately modeling the entire valuation surface. Furthermore, there are several studies which investigate active preference learning for practical applications such as, collaborative filtering (Jin and Si 2004; Harpale and Yang 2008; Boutilier et al. 2003), personalized calendar scheduling (Gervasio et al. 2005), or for optimizing search results for biomedical documents (Arens 2008). The difference between our work and the other studies for active preference learning mentioned above is that we consider active preference learning in a multi-task setting, i.e., we are interested about settings with multiple learning tasks and how active learning can be implemented in an efficient way in this case. We propose a criterion for active learning designed for this multi-task learning setting. This criterion, which we call the Committee criterion, will make use of the preference observations collected already from a community of subjects. The idea behind the Committee criterion is related to the Query-by-Committee method from active learning which selects those queries that have maximum disagreement amongst an ensemble of hypotheses. The difference in our case is that the group of subjects, for which the preferences were already learned, plays the role of the ensemble of hypotheses instead of an ensemble of models learned on the same task.

1.2 Notation

Boldface notation is used for vectors and matrices and normal fonts for their components. Upperscripts are used to distinguish between different vectors or matrices and lowerscripts to address their components. The notation $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is used for a multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The transpose of a matrix $\boldsymbol{M}$ is denoted by $\boldsymbol{M}^T$. Capital letters are used for constants and small letters for indices, e.g., $i = 1, \ldots, I$.

🖄 Springer

2 Learning framework

The idea of using the preference observations from other subjects in order to optimize the process of learning the preferences of a new subject can be basically applied in any preference learning context. In this work, we consider the case of qualitative preference observations which can be modeled using the probabilistic choice models described in this section.

2.1 Probabilistic choice models

In many real-world applications preferences are learned from experiments in which the subject makes a choice for one of the presented alternatives. The motivation for this is that people are very good at making comparisons between alternatives and expressing a preference for one of them, which results in qualitative preference observations. This is in contrast to quantitative preference observations where the people have to assign an absolute rating to each alternative independently. Let $X = \{x_1, \ldots, x_I\}$ be a set of inputs. Let $\mathcal{D}$ be a set of J observed preference comparisons over instances in X corresponding to a subject,

$$\mathcal{D} = \big\{(a_j, c_j) \mid 1 \le j \le J, \ c_j \in \{1, \ldots, A\}\big\} \tag{1}$$

with $a_j = (x_{i_1(j)}, \ldots, x_{i_A(j)})$ representing the alternatives presented to the subject and c_j the choice made, $i_1, \ldots, i_A : \{1, \ldots, J\} \to \{1, \ldots, I\}$ index functions such that $i_1(j)$ represents the input presented first in the jth preference comparison and $c_j = c$ means that $x_{i_c(j)}$ is chosen from the A alternatives presented in the jth comparison. For $A = 2$ this setup reduces to pairwise comparisons between two alternatives.

The main idea behind probabilistic choice models is to assume a latent utility function value $U(x_i)$ associated with each input x_i which captures the individual preference of a subject for x_i (the utility function will be formally defined in the next section). In the ideal case the latent function values are consistent with the preference observations. This means that alternative c is preferred over the other alternatives c' in the jth comparison whenever the utility for c exceeds the utilities for the other alternatives c', i.e., $U(x_{i_c(j)}) > U(x_{i_{c'}(j)})$. In practice, however, subjects are often inconsistent in their responses. A very inconsistent subject will have a high uncertainty associated with the utility function; this uncertainty is directly taken into account in the probabilistic framework. We define this probabilistic framework using the Bradley-Terry model (Bradley and Terry 1952; Kanninen 2002; Glickman and Jensen 2005) by making a standard modeling assumption that the probability that the cth alternative is chosen by the subject in the jth comparison follows a multinomial logistic model, which is defined as

$$p(c_j = c|a_j, U) = \frac{\exp[U(x_{i_c(j)})]}{\sum_{c'=1}^{A} \exp[U(x_{i_{c'}(j)})]}, \tag{2}$$

where "exp" is the exponential function and the other terms are as defined before. Efficiently learning preferences reduces to learning the unknown utility function U as accurately and with as few comparisons as possible.

One important drawback of the Bradley-Terry model is that it assumes very strong transitivity conditions of preference relations, while some psychological experiments have shown that human preference judgments can violate transitivity (Anand 1993; Tversky 1998). In most situations transitivity violations can be considered as noise. When this is not applicable, specific probabilistic models for human preference judgements which preserve intransitive reciprocal relation have to be designed. This was recently investigated in (Pahikkala et al. 2009) which introduced a new kernel function in the framework of regularized least squares which is capable of inferring intransitive reciprocal relations.

Mach Learn

2.2 The utility function

The utility function U is a real-valued function, $U : X \rightarrow \mathbb{R}$, which associates with every input $x \in X$ a real number $U(x)$. Each input $x \in X$ is characterized by a set of features, $\phi(x) \in \mathbb{R}^D$. One possible choice for the utility function is to express it as a linear combination of the features,

$$U(x) = \sum_{i=1}^{D} \alpha_i \phi_i(x), \tag{3}$$

where $\alpha = (\alpha_1, \ldots, \alpha_D)$ is a vector of weights which captures the importance of each feature of x when evaluating the utility U for a specific subject, $\phi_i(x)$ are the components of the vector $\phi(x)$. The preferences of a subject are thus encoded in the vector α and learning the utility function reduces to learning α.

In order to make the definition of the utility function more flexible, we can use a semi-parametric model in which the utility function is defined as a linear combination of basis functions. The basis functions are defined by a kernel function κ centered on the data points,

$$U(x) = \sum_{i=1}^{I} \alpha_i \kappa(x, x_i), \tag{4}$$

where the vector α with dimension I—the number of data points (the size of the set of inputs X)—captures the preferences of the subject. A non-linear utility function can be obtained by using, for example, a Gaussian kernel,

$$\kappa_{\text{Gauss}}(x, x') = \exp\left(-\frac{\ell}{2} \sum_{j=1}^{D} (\phi_j(x) - \phi_j(x'))^2\right), \tag{5}$$

where ℓ is a length-scale parameter. The two definitions of the utility function from Eqs. (3) and (4) are similar in the sense that they are both linear in the parameter. Equation (4) is suited when the number of features is larger than the number of data points, i.e., $D > I$ and for introducing non-linearity in the utility model.

In order to learn the utility function, we use a Bayesian framework in which we treat the vector of parameters α as a random variable. We consider a Gaussian prior distribution over $\alpha \sim \mathcal{N}(\mu, \Sigma)$, which is updated based on the observations from the preference comparisons using Bayes' rule (where we omitted the normalization constant),

$$p(\alpha|\mathcal{D}, \mu, \Sigma) \propto p(\alpha) \prod_{j=1}^{J} p(c_j|a_j, \alpha), \tag{6}$$

with the likelihood terms of the form given in Eq. (2). The choice of the prior will be discussed in the next section. The posterior distribution obtained is approximated to a Gaussian. The Gaussian approximation of the posterior is a good approximation because with few data points the posterior is close to the prior which is a Gaussian, and with many data points the posterior approaches again a Gaussian as a consequence of the central limit theorem (Bishop 2006). To perform the approximation of the posterior a good choice is to use deterministic methods (e.g., Laplace's method (MacKay 2002), Expectation Propagation (Minka 2001)) since they are computationally cheaper than the non-deterministic ones (sampling) and because they are known to be accurate for these types of models (Glickman and Jensen 2005).

« MACH 10994   layout: Small Condensed v.1.1   file: mach5297.tex (Ramune)   class: spr-small-v1.3 v.2012/04/18   Prn:2012/05/21; 15:39   p. 8/28»
«   doctopic: OriginalPaper   numbering style: ContentOnly   reference style: apa»

Mach Learn

2.3 Multi-task preference learning

One property which distinguishes preference learning from other learning settings is that in most of the cases in which preference learning is involved, observations are available from multiple subjects. For example, in the case of product recommendation, preferences are learned from multiple consumers. Furthermore, in situations where the training preference data is collected in an explicit way, by interactions with the subject, the individual training data is usually small. As a consequence, it is reasonable to make use of all the data available. In this direction, we make the assumption that each subject is not an isolated case, but belongs to a group of people sharing a common underlying rationale in their way of making preference decisions. This property combined with the Bayesian framework allows the transfer of information about preferences between different subjects. Basically, we use the preference data previously seen from some other subjects to learn an informed prior which will be used as the starting prior when learning the preferences of a new subject. For learning this informed prior, we use Bayesian hierarchical modeling which assumes that the parameters for individual models are drawn from the same hierarchical prior distribution. Let us assume that we already have preference data available from a group of M subjects. We make the common assumption of a Gaussian prior distribution, $p(\boldsymbol{\alpha}^m) = \mathcal{N}(\boldsymbol{\alpha}^m; \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$, $m = 1, \ldots, M$ with the same $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\Sigma}}$ for the preference models of all subjects. This prior is updated using Bayes' rule based on the observations from each subject, resulting in a posterior distribution for each individual subject. The common prior over all task parameters controls the general part of the model. This common prior is learned from the data belonging to a group of tasks, other than the current (new) task for which the learning is performed. Starting from this general-model given by the common prior, the model is updated using the observation (data) seen in the current task. These task-specific observations control the task-specific part of the model. The hierarchical prior is obtained by maximizing the penalized log-likelihood of all data in a so-called type-II maximum likelihood approach. This optimization is performed by applying the EM algorithm (Gelman et al. 2003), which reduces to the iteration (until convergence) of the following steps:

E-step: Estimate the sufficient statistics (mean $\boldsymbol{\mu}^m$ and covariance matrix $\boldsymbol{\Sigma}^m$) of the posterior distribution corresponding to each subject m, given the current estimates at step t ($\bar{\boldsymbol{\mu}}^{(t)}$ and $\bar{\boldsymbol{\Sigma}}^{(t)}$) of the hierarchical prior.

M-step: Re-estimate the parameters of the hierarchical prior:

$$\bar{\boldsymbol{\mu}}^{(t+1)} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\mu}^m, \tag{7}$$

$$\bar{\boldsymbol{\Sigma}}^{(t+1)} = \frac{1}{M} \sum_{m=1}^{M} (\boldsymbol{\mu}^m - \bar{\boldsymbol{\mu}}^{(t+1)})(\boldsymbol{\mu}^m - \bar{\boldsymbol{\mu}}^{(t+1)})^T + \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\Sigma}^m. \tag{8}$$

The details of the derivations of Eqs. (7) and (8) can be found in Appendix C of our paper (Birlutiu et al. 2009). The hierarchical prior is set to $p(\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{\alpha}; \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ where $\bar{\boldsymbol{\mu}} = \bar{\boldsymbol{\mu}}^T$ and $\bar{\boldsymbol{\Sigma}} = \bar{\boldsymbol{\Sigma}}^T$, at step T when iterations were stopped (at convergence). Once we have learned the hierarchical prior we can use it as an informative prior for the preference model of a new subject in Eq. (6).

Mach Learn

3 Active preference learning

In this section we discuss methods for active preference learning. We start from Query-by-Committee (QBC) (Seung et al. 1992) method for active learning and based on it we propose some variants of QBC adapted to the setting of preference learning for multiple subjects (Sect. 3.1). Furthermore, we show how these variants of QBC can be naturally linked to the hierarchical Bayesian modeling for reducing the computations (Sect. 3.2). Finally, we show connections between the variants of QBC proposed and other active learning criteria (Sect. 3.4).

3.1 QBC for preference learning

In this section we will discuss how to adapt QBC to our preference learning setting.

3.1.1 The committee members

For the QBC approach to be effective it is important that the committee is made of consistent and representative models. The main idea in this work is to exploit the preference learning setting with multiple subjects and use the learned models of other subjects $\mathcal{M}_1, \ldots, \mathcal{M}_M$ as committee members when learning the preferences of a new subject.

After choosing the committee we still have to decide upon a suitable criterion for selecting the next examples. Some measures of disagreement among the committee members appear to be most obvious, and in the following we will consider two alternatives.

3.1.2 Vote criterion

A simple and straightforward way is to consider the labels assigned by the other subjects, e.g., through the Vote criterion defined as

$$\text{Vote}(a) = \max_c \sum_{m=1}^{M} \delta(a, c; m), \tag{9}$$

where $\delta(a, c; m) = 1$ if $(a, c) \in \mathcal{D}_m$, and $\delta(a, c; m) = 0$ otherwise. The score $\text{Vote}(a)$ is minimal when the labels assigned by the committee members are equally distributed (total disagreement) and maximal when all members fully agree. There are two problems with this criterion. First, a comparison a may not be labeled by a subject m. This can be overcome if we consider the predictions computed based on the learned model of subject m and allow each committee member to 'vote' for its winning class. This same idea is implemented in the so-called vote entropy method (Dagan and Engelson 1995). The entropy is measured over the final classes assigned to an example by possible models, and not over class probabilities given by possible models. Second, in practical applications just scoring votes turns out to be suboptimal. The reason, as also suggested in (McCallum and Nigam 1998), is that the Vote criterion does not take into account the confidences of the committee members' predictions.

3.1.3 Committee criterion

We will use the following notation for the predictive probability corresponding to a subject $m = 1, \ldots, M$,

$$p_m(c|a) \equiv p(c|a, \mathcal{M}_m).$$

🖄 Springer

AUTHOR'S PROOF

The predictive probability can be computed either by taking into account the entire distribution $\mathcal{M}_m = \mathcal{N}(\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m)$

$$p_m(c|a) = \int p(c|a, \boldsymbol{\alpha}) \mathcal{N}(\boldsymbol{\alpha}|\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m) d\boldsymbol{\alpha},$$

or, for computational reasons, we can consider only a point estimate for $\mathcal{M}_m$, for example, the mean of the Gaussian distribution, and use it to compute the predictive probabilities using Eq. (2)

$$p_m(c|a) \approx p(c|a, \boldsymbol{\mu}^m). \tag{10}$$

Inspired by (McCallum and Nigam 1998), we propose to measure disagreement by taking the average prediction of the entire committee and computing the average Kullback-Leibler (KL) divergence of the individual predictions from the average:

$$\text{Committee}(a) = \sum_{m=1}^{M} \frac{1}{M} \text{KL}[\bar{p}(\cdot|a)||p_m(\cdot|a)], \tag{11}$$

with $\bar{p}(\cdot|a)$ the average predictive probability of the entire committee, which will be more precisely defined in Sect. 3.2.

The KL divergence for discrete probabilities is defined as

$$\text{KL}[p_1(\cdot|a)||p_2(\cdot|a)] = \sum_c p_1(c|a) \log\left(\frac{p_1(c|a)}{p_2(c|a)}\right).$$

The KL divergence can be seen as a distance between probabilities, where we abused the notion of distance, since the KL-divergence is not symmetric, i.e., $\text{KL}[p_1||p_2] \neq \text{KL}[p_2||p_1]$. This drawback of the KL-divergence can be overcome by considering a symmetric measure, for example, $\text{KL}[p_1||p_2] + \text{KL}[p_2||p_1]$. In (McCallum and Nigam 1998), the disagreement is computed between committee members constructed based on the current model, i.e., the committee changes with every update and the criterion has to be recomputed with every update. A committee of models learned on different tasks is fixed and thus selecting examples solely based on it leads to a fixed instead of an active design: all examples can be ranked beforehand (the same applies to the Vote criterion defined above).

To arrive at an active design and take into account the current model, we propose a small modification based on the following intuition. Querying examples on which the committee members disagree makes sense, because it will force the current model to make a choice between options that, according to the committee members, are reasonably plausible. However, when the current model on a particular example already "made up its mind", i.e., deviates substantially from the average prediction of the committee based on what it learned from other input/output pairs, it makes no sense to still query that example, even though the committee members might disagree. Taking into account this consideration, we propose the Committee criterion which assigns a score to a candidate query comparison a through

$$\text{Committee}(a) = \frac{1}{M} \sum_{m=1}^{M} \text{KL}[\bar{p}(\cdot|a)||p_m(\cdot|a)] - \gamma \text{KL}[\bar{p}(\cdot|a)||p(\cdot|a)], \tag{12}$$

with $p(\cdot|a)$ the current model's predictive probability based on the data seen so far and γ a parameter that accounts for the degree of similarity between subjects. According to the

⚡ Springer

Mach Learn

Committee criterion, the most interesting experiments are those on which the other models disagree (the first term on the right-hand side of Eq. (12)), with the current model (still) undecided (the second term on the right-hand side of Eq. (12)).

An advantage of the Committee criterion is its computational efficiency: the first term on the right-hand side of Eq. (12) as well as the average predictive probability can be computed beforehand. The Committee criterion does require computation of the predictive probabilities corresponding to the current model, but this is the least one could expect from an active design. This is to be compared with the QBC criterion (any of the two variants considered), which requires constructing new committee members with each update, and D-optimal experimental design, which calls for keeping track of variances.

Note that we have not made any restriction so far with respect to the probabilistic models used in the active learning design. In the following we will consider only the log-linear models introduced in Sect. 2. They have some nice properties, which simplify the computation of the Committee criterion (Sect. 3.2), and provide a natural link to hierarchical Bayesian modeling (Sect. 2.3). The general idea, of using the already learned models from the other tasks as the committee members in a QBC-like approach, is of course also applicable to other models.

3.2 Average probability

In this section we discuss how to efficiently compute the average probability used for computing the committee criterion in Eq. (12) in the case of log-linear models (Christensen 1997). For linear utility functions the likelihood function defined in Eq. (2) is a log-linear model. The log-odds of the model are linear in the parameter.

Let $p_m(c|a)$ be the predictive probability defined in Eq. (10). We define the average predictive probability of the committee, $\bar{p}(c|a)$, as the prediction probability that is closest to the prediction probabilities of the members:

$$\bar{p}(c|a) \equiv \operatorname{argmin}_{p(c|a)} \sum_{m=1}^{M} \frac{1}{M} \operatorname{KL}\big[p(c|a)||p_m(c|a)\big]. \tag{13}$$

The solution of the optimization from above is the so-called logarithmic opinion pool (Bordley 1982)

$$\bar{p}(c|a) = \frac{1}{Z(a)} \prod_{m=1}^{M} \big[p_m(c|a)\big]^{\frac{1}{M}} = \frac{1}{Z(a)} \exp\left(\frac{1}{M} \sum_{m=1}^{M} \log p_m(c|a)\right), \tag{14}$$

with $Z(a)$ a normalization constant

$$Z(a) = \sum_{c} \prod_{m=1}^{M} \big[p_m(c|a)\big]^{\frac{1}{M}}.$$

For log-linear models, the logarithmic opinion pool boils down to a simple averaging of model parameters:

$$\bar{p}(c|a) = p(c|a, \bar{\boldsymbol{\mu}}) \quad \text{with } \bar{\boldsymbol{\mu}} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\mu}^m. \tag{15}$$

« MACH 10994   layout: Small Condensed v.1.1   file: mach5297.tex (Ramune)   class: spr-small-v1.3 v.2012/04/18   Prn:2012/05/21; 15:39   p. 12/28»
«   doctopic: OriginalPaper   numbering style: ContentOnly   reference style: apa»

Mach Learn

This natural combination between log-linear models and logarithmic opinion pools is the advantage of using the logarithmic opinion pool instead of the linear opinion pool used in (McCallum and Nigam 1998).

As can be seen from the EM updates in Eq. (7), the average $\bar{\mu}$ in the logarithmic opinion pool is then precisely the mean of the learned hierarchical prior. Summarizing, once we have learned a hierarchical prior from the data available for subjects 1 through M using the EM algorithm, we can start off the new model $M + 1$ from this prior (as is normally done in hierarchical Bayesian learning). On top of this, the same EM algorithm gives us the information we need to compute the Committee criterion that can be used subsequently to select new inputs to label.

3.3 Other criteria for active learning

In this section we discuss how several strategies for active learning can be implemented in the learning framework considered here. All the strategies are being concerned with evaluating the informativeness of the unlabeled points. Let the new model obtained after incorporating an observation (a, c) be $\mathcal{M}_{(a,c)} \approx \mathcal{N}(\boldsymbol{\mu}_{(a,c)}, \boldsymbol{\Sigma}_{(a,c)})$.

1. *Uncertainty sampling* (Lewis and Gale 1994). In this strategy an active learner chooses for labeling the example for which the model's predictions are most uncertain. The uncertainty of the predictions can be measured, for example, using Shannon entropy

$$\text{Uncertainty}(a) = - \sum_c p(c|a, \mathcal{M}) \log p(c|a, \mathcal{M}). \tag{16}$$

For a binary classifier this strategy reduces to querying points whose prediction probabilities are close to 0.5. Intuitively this strategy aims at finding as fast as possible the decision boundary since this is indicated by the regions where the model is most uncertain.

2. *Variance reduction* (MacKay 1992). This strategy, also known in experimental design as D-optimality (Fedorov 1972; Chaloner and Verdinelli 1995; Berger 1994; Ford and Silvey 1980), chooses as the most informative experiments the ones that give the most reduction in the model's uncertainty. The motivation behind this strategy is a result of (Geman et al. 1992) which shows that the generalization error can be decomposed into three components: (i) noise (which is independent of the model or training data); (ii) bias (due to the model); (iii) model's uncertainty. Since the model cannot influence the noise and the bias components, the future generalization error can only be influenced via the model's variance. Formally, this criterion can be written as

$$\text{Variance}(a) = \sum_c p(c|a, \mathcal{M}) \text{variance}[\mathcal{M}_{(a,c)}] - \text{variance}[\mathcal{M}]. \tag{17}$$

In the setting considered in this work the variance of the model is expressed in the covariance of the Gaussian distribution. In order to use Eq. (17) we need to choose a measure for the variance. We can consider, for example, the log-determinant of the covariance matrix

$$\text{Variance-logdet}(a) = \sum_c p(c|a, \mathcal{M}) \log \det(\boldsymbol{\Sigma}_{(a,c)}) - \log \det(\boldsymbol{\Sigma}), \tag{18}$$

🖄 Springer

Mach Learn

which is actually minimizing the entropy of the Gaussian random variable representing the current model, or the trace of the covariance matrix

$$\text{Variance-trace}(a) = \sum_c p(c|a, \mathcal{M}) \, \text{Tr}(\boldsymbol{\Sigma}_{(a,c)}) - \text{Tr}(\boldsymbol{\Sigma}). \tag{19}$$

3. *Expected model change* (Cohn et al. 1996). This strategy chooses as the most informative query the one which when added to the training set would yield the greatest model change. Quantifying the model change depends on the learning framework. For gradient-based optimization the change can be measured via the training gradient, i.e., the vector used to re-estimate parameter values (Settles and Craven 2008). In the Bayesian framework, the model change can be quantified via a distance measure between the current distribution and the posterior distribution obtained after incorporating the candidate point

$$\text{Change}(a) = \sum_c p(c|a, \mathcal{M}) \text{distance}[\mathcal{M}, \mathcal{M}_{(a,c)}].$$

A suitable distance for our setting is the Kullback-Leibler divergence between distributions, which for two Gaussians has a closed form solution and can be written as follows

$$\text{Change-KL}(a) = \sum_c p(c|a, \mathcal{M}) KL[\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) || \mathcal{N}(\boldsymbol{\mu}_{(a,c)}, \boldsymbol{\Sigma}_{(a,c)})]$$

$$= \sum_c p(c|a, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \left[\log\left(\frac{\det \boldsymbol{\Sigma}_{(a,c)}}{\det \boldsymbol{\Sigma}}\right) + \text{Tr}(\boldsymbol{\Sigma}_{(a,c)}^{-1} \boldsymbol{\Sigma}) \right.$$

$$\left. + (\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_{(a,c)}^{-1} (\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu}) - n \right]. \tag{20}$$

The KL divergence between Gaussians is used by Seeger (2008) to design an efficient sequential experimental design in a setting similar to the one used in this work.

Uncertainty sampling, and QBC and its variants are attractive due to their applicability in various machine learning settings. Variance reduction and expected model change are robust and in many situations they have proved to be the best one can do (Schein and Ungar 2007). Although more robust, the variance reduction and expected model change strategies are computationally more demanding since for each candidate comparison query and each possible label the posterior distribution induced has to be computed. For the learning setting considered in this study, the posterior distribution computed using Bayes' theorem (Eq. (6)) does not have an analytical expression and approximations are needed for it; these approximations are usually costly. In contrast, the variants of QBC proposed in this paper are computationally efficient

3.4 Similarities between criteria

In this section we consider the following active learning criteria: Variance-logdet, Committee, Variance-trace and Change-KL. We investigate how similar the active learning criteria are and how they can be related. We analyze the modifications induced to the model by the criteria after updating the probability model to incorporate the information from new training points. A single update induces a small change in the posterior distribution, and this allows for Taylor expansions, keeping only the lowest non-zero contribution. In the following we present the main results of the approximations while some of the details can be found in the Appendix.

⁂ Springer

As we will show below and in the Appendix, under the assumption that the updates of the posterior distribution for each alternative a and choice c lead to small changes in the model $\mathcal{M}$, we can approximate the active criteria to the form

$$\sum_c p(c|a, \boldsymbol{\alpha}) \boldsymbol{g}(c|a, \boldsymbol{\alpha})^T \boldsymbol{Q} \boldsymbol{g}(c|a, \boldsymbol{\alpha}), \tag{21}$$

for some vector $\boldsymbol{\alpha}$ and matrix $\boldsymbol{Q}$ and with $\boldsymbol{g}$ the gradient of the log-probabilities

$$\boldsymbol{g}(c|a, \boldsymbol{\alpha}) \equiv \frac{\partial \log p(c|a, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}.$$

The following lemma approximates the Variance-logdet criterion to the form from Eq. (21).

Lemma 1 *In a first order approximation, assuming that $\boldsymbol{\Sigma}_{(a,c)}$ is close to $\boldsymbol{\Sigma}$, we can simplify*

$$\text{Variance-logdet}(a) \approx \sum_c p(c|a, \boldsymbol{\mu}) \boldsymbol{g}(c|a, \boldsymbol{\mu})^T \boldsymbol{\Sigma} \boldsymbol{g}(c|a, \boldsymbol{\mu}), \tag{22}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ represent the mean and covariance of the Gaussian posterior distribution.

Proof In a first order approximation we have

$$\boldsymbol{\Sigma}_{(a,c)}^{-1} \approx \boldsymbol{\Sigma}^{-1} - \frac{\partial^2 \log p(c|a, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} \bigg|_{\boldsymbol{\alpha}=\boldsymbol{\mu}} \tag{23}$$

where we ignored the change from the old $\boldsymbol{\alpha}$ to a new MAP solution depending on c and a. We will use the notation

$$\boldsymbol{H}(c|a, \boldsymbol{\alpha}) \equiv \frac{\partial^2 \log p(c|a, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T}.$$

For a matrix $\boldsymbol{A}$ and ϵ small compared to $\boldsymbol{A}$, the following holds (see, for example, Boyd and Vandenberghe 2004, p. 642)

$$\log \det(\boldsymbol{A} + \epsilon \boldsymbol{I}) \approx \log \det(\boldsymbol{A}) + \text{Tr}\big[\boldsymbol{A}^{-1}\epsilon\big]. \tag{24}$$

Assuming $\boldsymbol{\Sigma}_{(a,c)}^{-1}$ is close to $\boldsymbol{\Sigma}^{-1}$ which makes $\boldsymbol{H}(c|a, \boldsymbol{\mu})$ small, we can use Eq. (23) in Eq. (24) with the following substitutions $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{H}(c|a, \boldsymbol{\mu}) = \epsilon \boldsymbol{I}$ to obtain

$$\log \det \boldsymbol{\Sigma}_{(a,c)}^{-1} \approx \log \det \boldsymbol{\Sigma}^{-1} - \text{Tr}\big[\boldsymbol{\Sigma} \boldsymbol{H}(c|a, \boldsymbol{\mu})\big]. \tag{25}$$

The probability that the subject gives the response c when presented the alternatives a follows by integrating $p(c|a, \boldsymbol{\alpha})$ over the current posterior. We make a second order Taylor expansion of $p(c|a, \boldsymbol{\alpha})$ around the point $\boldsymbol{\mu}$:

$$p(c|a) = \int d\boldsymbol{\alpha} \, p(c|a, \boldsymbol{\alpha}) \mathcal{N}(\boldsymbol{\alpha}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\approx \int d\boldsymbol{\alpha} \, p(c|a, \boldsymbol{\alpha}) + (\boldsymbol{\alpha} - \boldsymbol{\mu})^T \frac{\partial p(c|a, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \bigg|_{\boldsymbol{\alpha}=\boldsymbol{\mu}} \mathcal{N}(\boldsymbol{\alpha}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Mach Learn

$$+ \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\mu})^T \frac{\partial^2 p(c|a, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T}\bigg|_{\boldsymbol{\alpha}=\boldsymbol{\mu}} (\boldsymbol{\alpha} - \boldsymbol{\mu}) \mathcal{N}(\boldsymbol{\alpha}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$= p(c|a, \boldsymbol{\mu}) + \frac{1}{2} \mathrm{Tr}\left[\boldsymbol{\Sigma} \frac{\partial^2 p(c|a, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T}\bigg|_{\boldsymbol{\alpha}=\boldsymbol{\mu}}\right].$$

The first order term cancels since the gradient is zero at the maximum solution $\boldsymbol{\alpha} = \boldsymbol{\mu}$. In a lowest order approximation we can ignore the correction upon $p(c|a, \boldsymbol{\alpha})$ to obtain

$$\text{Variance-logdet}(a) = -\sum_c p(c|a, \boldsymbol{\mu}) \log \det \boldsymbol{\Sigma}_{(a,c)} + \log \det \boldsymbol{\Sigma}$$

$$= -\sum_c p(c|a, \boldsymbol{\mu})\left[-\log \det\left(\boldsymbol{\Sigma}_{(a,c)}^{-1}\right) + \log \det\left(\boldsymbol{\Sigma}^{-1}\right)\right]$$

$$\approx -\sum_c p(c|a, \boldsymbol{\mu}) \mathrm{Tr}\left[\boldsymbol{\Sigma} \boldsymbol{H}(c|a, \boldsymbol{\mu})\right],$$

where for the last approximation we used the approximation from Eq. (25). To obtain the proof of this lemma we use Lemma 3 in the Appendix at the end of the paper which states a relationship between Hessian and Fisher matrices. □

Using the same type of approximation, the Committee criterion can be approximated to the same form given in Eq. (21).

Lemma 2 *In a lowest order approximation the Committee criterion can be written as*

$$\text{Committee}(a) \approx \frac{1}{2}\sum_c p(c|a, \bar{\boldsymbol{\mu}}) \boldsymbol{g}(c|a, \bar{\boldsymbol{\mu}})^T \tilde{\boldsymbol{\Sigma}} \boldsymbol{g}(c|a, \bar{\boldsymbol{\mu}}),$$

where $\bar{\boldsymbol{\mu}}$ is the mean of the hierarchical prior learned from the other subjects and

$$\tilde{\boldsymbol{\Sigma}} \equiv \frac{1}{M}\sum_{m=1}^{M}\left(\boldsymbol{\mu}^m - \bar{\boldsymbol{\mu}}\right)\left(\boldsymbol{\mu}^m - \bar{\boldsymbol{\mu}}\right)^T - (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T.$$

We make a second order Taylor expansion of the KL divergences from the definition of the Committee criterion in Eq. (12)

$$\mathrm{KL}\left[\bar{p}(\cdot|a)||p(\cdot|a)\right] = \sum_c p(c|a, \bar{\boldsymbol{\mu}}) \log\left[\frac{p(c|a, \bar{\boldsymbol{\mu}})}{p(c|a, \boldsymbol{\mu})}\right],$$

around the point $\bar{\boldsymbol{\mu}}$.

The first order term of the Taylor expansion is:

$$-\sum_c p(c|a, \bar{\boldsymbol{\mu}}) \frac{\partial \log p(c|a, \bar{\boldsymbol{\mu}})}{\partial \boldsymbol{\mu}}\bigg|_{\boldsymbol{\mu}=\bar{\boldsymbol{\mu}}} (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T = \sum_c p(c|a, \bar{\boldsymbol{\mu}}) \boldsymbol{g}(c|a, \bar{\boldsymbol{\mu}})(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T$$

which cancels since based on Eq. (28) from the Appendix $\sum_c p(c|a, \bar{\boldsymbol{\mu}}) \boldsymbol{g}(c|a, \bar{\boldsymbol{\mu}})$ is the vector with every component 0.

The second order term can be rewritten using Lemma 3 as:

$$-\frac{1}{2}\sum_c p(c|a,\bar{\boldsymbol{\mu}})(\boldsymbol{\mu}-\bar{\boldsymbol{\mu}})^T \boldsymbol{H}(c|a,\bar{\boldsymbol{\mu}})(\boldsymbol{\mu}-\bar{\boldsymbol{\mu}})$$

$$=\frac{1}{2}(\boldsymbol{\mu}-\bar{\boldsymbol{\mu}})^T \sum_c p(c|a,\bar{\boldsymbol{\mu}})\boldsymbol{g}(c|a,\bar{\boldsymbol{\mu}})\boldsymbol{g}(c|a,\bar{\boldsymbol{\mu}})^T(\boldsymbol{\mu}-\bar{\boldsymbol{\mu}})$$

$$=\frac{1}{2}\sum_c p(c|a,\bar{\boldsymbol{\mu}})\boldsymbol{g}(c|a,\bar{\boldsymbol{\mu}})^T(\boldsymbol{\mu}-\bar{\boldsymbol{\mu}})(\boldsymbol{\mu}-\bar{\boldsymbol{\mu}})^T \boldsymbol{g}(c|a,\bar{\boldsymbol{\mu}}).$$

Since the other terms cancel, we obtain that the KL-divergence between the predictive probabilities can be approximated as

$$\text{KL}\big[\bar{p}(\cdot|a)||p(\cdot|a)\big]=\frac{1}{2}\sum_c p(c|a,\bar{\boldsymbol{\mu}})\boldsymbol{g}(c|a,\bar{\boldsymbol{\mu}})^T(\boldsymbol{\mu}-\bar{\boldsymbol{\mu}})(\boldsymbol{\mu}-\bar{\boldsymbol{\mu}})^T \boldsymbol{g}(c|a,\bar{\boldsymbol{\mu}}).$$

Making this approximation for all the KL-divergences in the definition of the Committee criterion from Eq. (12) and computing the sum we obtain the result stated in this lemma.

Furthermore, it can also be showed that Variance-trace and Change-KL can be approximated to the same form given in Eq. (21), namely

$$\text{Variance-trace}(a) \approx \sum_c p(c|a,\boldsymbol{\mu})\boldsymbol{g}(c|a,\boldsymbol{\mu})^T \boldsymbol{\Sigma}^2 \boldsymbol{g}(c|a,\boldsymbol{\mu}), \tag{26}$$

and

$$\text{Change-KL}(a) \approx \text{Variance-logdet}(a). \tag{27}$$

For the derivations of these approximations see Lemmas 4 and 5 in the Appendix.

We will focus on the differences between the Variance-logdet criterion (considered as the reference) and the Committee criterion. The differences between their approximations are as follows.

1. The gradients $\boldsymbol{g}(c|a,\cdot)$ are evaluated at different points: the prior hierarchical mean $\bar{\boldsymbol{\mu}}$ and the current posterior mean $\boldsymbol{\mu}$. This effect is small since $\boldsymbol{\mu}$ is still close enough to $\bar{\boldsymbol{\mu}}$ for a sufficiently accurate approximation of the gradients, in particular at the start of the learning when selecting the right points to label is more important.
2. The current posterior variance $\boldsymbol{\Sigma}$ is replaced by $\tilde{\boldsymbol{\Sigma}}$. The effect of the precise weighting of the gradients is not so important, and again, at the beginning of learning $\tilde{\boldsymbol{\Sigma}}$ is close to $\boldsymbol{\Sigma}$.

The way in which experiments are selected is more important at the beginning of the learning process, when $\boldsymbol{\mu}$ is still close to the prior mean $\bar{\boldsymbol{\mu}}$, and $\tilde{\boldsymbol{\Sigma}}$ to $\boldsymbol{\Sigma}$.

4 Experimental evaluation

This section presents the experimental evaluation of the framework proposed in this paper. We will use pairwise comparisons data. The main goal of the experimental evaluation section is to show that optimally selecting data for labeling using the Committee criterion achieves higher accuracy than random selection. Furthermore, we also show that the Committee criterion performs in practice similarly to other standard active learning criteria, but in addition has a computational advantage.

Mach Learn

4.1 Data sets

The following data sets related to the preferences of people were used in the experimental evaluation.

4.1.1 Letor

This data set consists of relevance levels assigned to documents with respect to a given textual query (the OHSUMED data set from Letor 3.0, Qin et al. 2010). The relevances were assessed by human experts, using three rank scales: definitely relevant, partially relevant, and not relevant. We used a subset of this data related to Query 1. It contains 138 references with the following labels: 24 definitely relevant, 26 partially relevant, and 88 not relevant. Each of the samples is characterized by a 45-dimensional vector consisting of text features extracted from the titles and abstracts of the documents. The features were normalized. Based on this data set we constructed pairwise preferences belonging to 50 subjects in a way that we describe below. We followed a procedure similar to (Xu et al. 2010) to turn the relevance levels into pairwise preference comparisons. Since such coarse relevance judgements are considered unrealistic in many real-world applications, Xu et al. (2010) proposed to add uniform noise in the range $[-0.5, 0.5]$ to the true relevance levels. This addition preserves the relative order between definitely relevant (respectively partially relevant) documents and partially relevant (respectively not relevant) ones, but randomly breaks ties within each relevance level. To introduce a hierarchical component, we replaced the random tie-breaking of Xu et al. (2010) by a subject-specific one. We do this by changing the uniform noise by a subject (and feature) dependent term as follows. For subject m, a weight vector $\boldsymbol{\alpha}_m$ is drawn from a zero mean fully factorized Gaussian with unit variance, $\boldsymbol{\alpha}_m \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Given features $\boldsymbol{x}_i$, noise terms are then the inner products $\boldsymbol{\alpha}_m^T \boldsymbol{x}_i$, linearly scaled back to the interval $[-0.5, 0.5]$ (not to destroy the relative order of the true relevance levels), and the relevance levels are taken to be the true relevance levels plus these noise terms.

4.1.2 Audio

This data set consists of evaluations of sound quality from 32 subjects (this data set was borrowed from Arehart et al. 2007). Each subject performed 576 pairwise comparison listening experiments. Each listening experiment represents one sound sample processed with two different settings of the hearing-aid parameters, and the choice for one of the two. The processed sound sample is represented by a 3-dimensional feature vector.

4.1.3 Art

This data set consists of evaluations of art images from 190 subjects (this data set was borrowed from Yu et al. 2003). Each subject was presented a number of images from a total of 642 images and asked to rate each of them: like/dislike (on average, each subject rated 90 images). We considered the 32 subjects who rated more than 120 images. Each image is described by a 275-dimensional feature vector, with features which characterize the image such as, color, shape, texture, etc. For computational reasons, we reduced the high-dimensional feature vector (275 dimensions) to a lower dimension. We noticed that most of the features where not very informative for predicting the outcomes, this is why we used only the 10 most informative features (the informativeness of the features was measured by averaging the correlations between features and observations). Note that this data set

$\underline{\textcircled{\tiny 2}}$ Springer

does not contain pairwise comparisons like the other two data sets. With each instance, a binary label is associated: like or dislike, which makes the learning task on this data set to be a binary classification task. The combination of multi-task and active learning that we propose in this work can be still applied in this case in the same framework which was introduced in Sect. 2 by using the logistic regression model instead of the Bradley-Terry model as likelihood terms in the model from Eq. (6).

4.2 Protocol

Our experiments use a leave-one-out scheme in which each subject was considered once as the current/test subject for which the preferences need to be learned. For each test subject the learning started with the hierarchical prior learned from the data of the other remaining subjects. The data for the test subject was split into 5 folds, 1 fold was used for training and the rest was used for testing. The training data was used as a pool out of which points were selected for labeling either randomly or actively using one of the active learning criteria. The hierarchical prior was updated based on these data points. After every update predictions were made on the test set using the current model. We used accuracy (percentage of correct predictions among all the predictions) as a measure of performance. The accuracy of the predictions on the test data measures how much we learned about the subject preferences. The results were averaged over the 5 splits and over the subjects.

4.3 Performance

The framework that we propose in this work for optimizing preference learning consists of combining the multi-task formalism together with active learning. The multi-task ideas in preference learing are especially useful when the training preference data from a subject is very small. In this situation it makes sense to use the preference data from other subjects as additional information.

4.3.1 Letor

The pairwise comparisons from Letor data set were generated by adding noise in the interval $[-0.5, 0.5]$ such that the relative order between the three relevance levels is preserved, but ties within each relevance level are broken. As a result, different subjects do agree on comparisons between different relevance levels. Thus, the data was constructed to have an underlying common structure in the preference of different subjects. Because of this reason we expect that the multi-task learning would improve the performance. In order to validate this hypothesis, we checked whether the preferences of a new subject can be learned more accurately by using the available preference data from other subjects. We compared the hierarchical model with the method of Chu and Ghahramani (2005b) Gaussian processes for preference learning which assumes no prior information. The hierarchical/community prior was obtained by applying the EM algorithm described in Sect. 2.3 in combination with the semiparametric utility function from Eq. (4); the hierarchical prior was learned from 20 samples from each of the other subjects. The method of Chu and Ghahramani (2005b) was applied with a Gaussian kernel in which the kernel parameters were tuned using cross-validation.

Figure 1, left panel, compares the accuracy obtained using the hierarchical model to the Gaussian process method for preference learning of (Chu and Ghahramani 2005b) in a non-active setting, i.e., for both models the updates are done with training points randomly
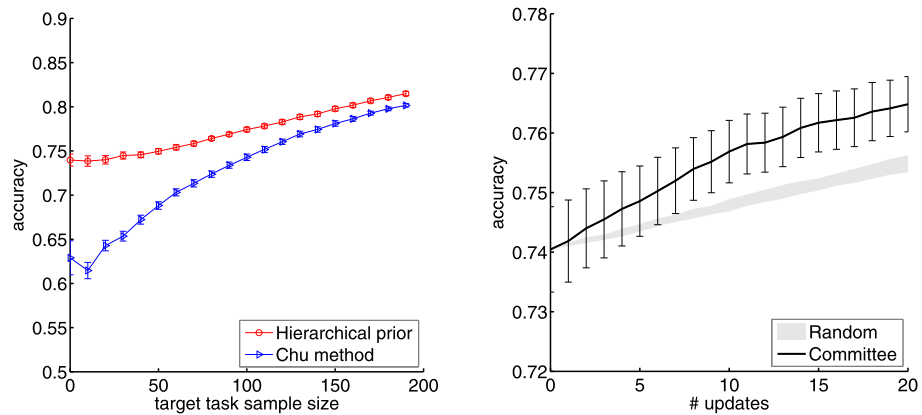
Mach Learn

Fig. 1 *Left*: Comparison between the hierarchical model that we discussed in Sect. 2.3 and the Gaussian processes for preference learning model (Chu and Ghahramani 2005b). The setting is non-active, the updates are done using training points randomly selected. *Right*: Random vs active selection of training points. The performance is evaluated as a function of the number of data points included in the training set. The active selection was implemented using the Committee criterion. The *shaded region* shows the range of 10 random strategies. The *error bars* give the standard deviation of the mean accuracy, averaged over the subjects. Please note the change in scaling of the *y*-axis between the left and right plots

selected. The prediction accuracy is shown as a function of the number of data points included in the training set for the test subject. The plots show that the improvement obtained with the hierarchical model depends on the size of the training data. This is in accordance with the expectation that the multi-task formalism is suited for situations in which the available training data is small. The results obtained using a hierarchical prior on the audio and art data sets are similar to the ones obtained in the case of the Letor data set. Figure 1, right panel, compares the accuracy obtained with random and active selection and starting from a hierarchical prior. Please note the change in scaling of the *y*-axis. The active selection was implemented using the Committee criterion. These plots show that the combination between multi-task and active learning indeed improves the performance.

4.3.2 Audio

Figure 2 shows the performance of the Committee criterion on the left and Variance-logdet criterion on the right versus random selection on the audio data set. The plots show the prediction accuracy (on the *y*-axis) as a function of the number of updates from the hierarchical prior (on the *x*-axis). The shaded region indicates the accuracy of 10 random selection runs. The error bars give the standard deviation of the mean accuracy, averaged over the 32 subjects. We used the Committee criterion with $\gamma = 0$ since the subjects in the committee are quite similar between each other, which is also suggested by the small error bars. The informative prior improves the predictions at the beginning when no preference observations have been observed for the new subject. The hierarchical prior already gives an accuracy of almost 0.7 for the audio data at the beginning of learning. The hierarchical prior was learned from 20 randomly selected data points per subject. Committee and Variance-logdet strongly overlap and are considerably better than a random strategy. The audio data set contains a few very informative data points and some which are not informative. In some cases the difference between the two sound samples presented in an experiment is so small that the
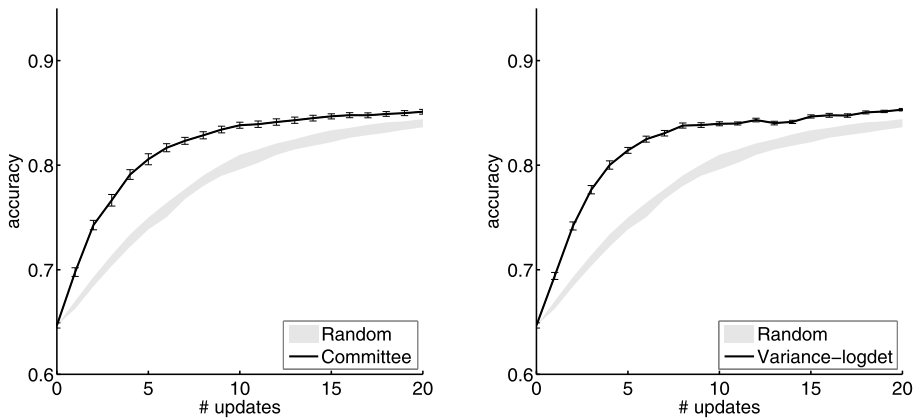
🖄 Springer

Mach Learn

Fig. 2 Performance of the Committee criterion on the left and Variance-logdet criterion on the right versus random selection for the audio data set. The *plots* show the prediction accuracy (on the *y*-axis) as a function of the number of updates from the hierarchical prior (on the *x*-axis). The *error bars* give the standard deviation of the mean accuracy, averaged over the 32 subjects. The *shaded region* shows the range of 10 random strategies

subject cannot hear any difference. Such experiments are not informative because a subject's answer is close to random and does not provide any information with respect to the subject's preferences. The active learning criteria avoid selecting these type of experiments and obtain better performance than random selection. The performance of the other active learning strategies (not shown) is comparable to the active learning strategies shown in Fig. 2, except for the Vote criterion which does not seem to perform better than random. We refer to Sect. 4.5 for an empirical evaluation of the similarities between the active learning criteria considered in Sect. 3.

4.3.3 Art

Figure 3 shows the performance of the Committee criterion on the left and Variance-logdet criterion on the right versus random selection on the art data set. The plots show the prediction accuracy (on the *y*-axis) as a function of the number of updates from the hierarchical prior (on the *x*-axis). The shaded region indicates the accuracy of 10 random selection runs. The error bars give the standard deviation of the mean accuracy, averaged over the subjects. For the art data which has a higher variability between subjects the Committee criterion with $\gamma = 1$ performs slightly better than the Committee criterion with $\gamma = 0$. The preferences of people for art images are more difficult to predict, since preferences do not depend on some low-level characteristics of the image, like texture, color, etc. This is why the accuracy obtained on the art data is less than the accuracy obtained, for example, on the audio data. The Variance-logdet criterion appears to perform slightly better than the Committee criterion. Furthermore, the benefit of active learning over random selection is much smaller. Like in the case of audio data set, the performance of the other active learning strategies (not shown) is comparable to the active learning strategies shown in Fig. 3.

4.4 Computational complexity

One of the main advantages of the Committee criterion is its computational simplicity in comparison with the standard criteria used in experimental design and exemplified here by

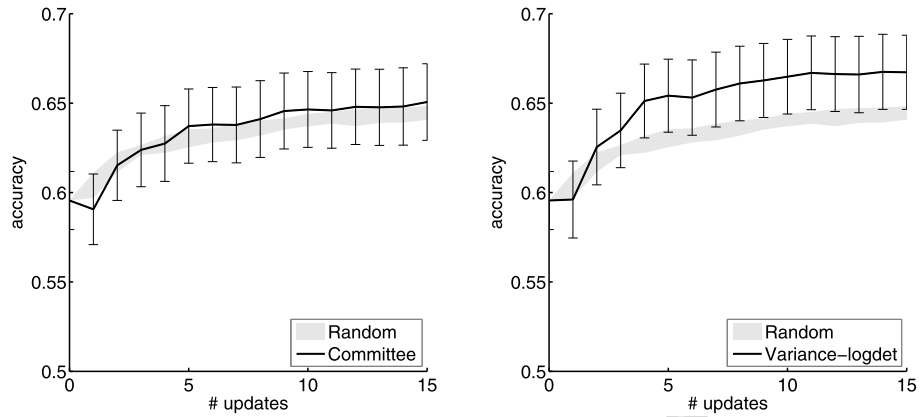♻ Springer

Mach Learn

Fig. 3 Performance of the Committee criterion on the left and Variance-logdet criterion on the right versus random selection for the art data set. The *plots* show the prediction accuracy (on the *y*-axis) as a function of the number of updates from the hierarchical prior (on the *x*-axis). The *error bars* give the standard deviation of the mean accuracy, averaged over the subjects. The *shaded region* shows the range of 10 random strategies

Table 1 Execution time (in seconds) for Variance-logdet and Committee criterion a function of feature dimension

Feature dimension	Variance-logdet (s)	Committee (s)
10	2.894	0.014
50	13.543	0.010
100	37.926	0.009
200	172.661	0.010

the Variance-logdet criterion. For every candidate data point to be included in the training set, the Variance-logdet criterion needs to infer the posterior distribution induced. A standard method for performing this approximation is Laplace's method which has a cubic complexity in the dimension of the data features because it involves inversions of the covariance matrices. For more details about other inference methods suited to pairwise comparison data we refer to (Birlutiu and Heskes 2007). All the algorithms presented here are linear in the number of data points which makes them scalable to large data sets. Table 1 shows a comparison between the execution times for Variance-logdet and Committee criterion as a function of feature dimension. Since these execution times for a fixed number of updates (one in our case) do not depend on the actual nature of the data, we randomly generated data in order to be able to change the number of input dimensions. The data was randomly generated to have dimension 10, 50, 100 and 200. The time was evaluated for 100 candidate data points and for 1 update step. The Committee criterion was computed from data belonging to 20 users. The simulations were performed using Matlab on an Intel Xeon processor with 16 Gb of memory which runs Fedora release 9 with Linux kernel 2.6.27. In the case of the Committee criterion, only KL divergences between predictive probabilities need to be computed. The Committee criterion is clearly much faster than the Variance-logdet, furthermore, contrary to the Variance-logdet criterion, the computational complexity of the Committee criterion is independent of the dimension of the features.
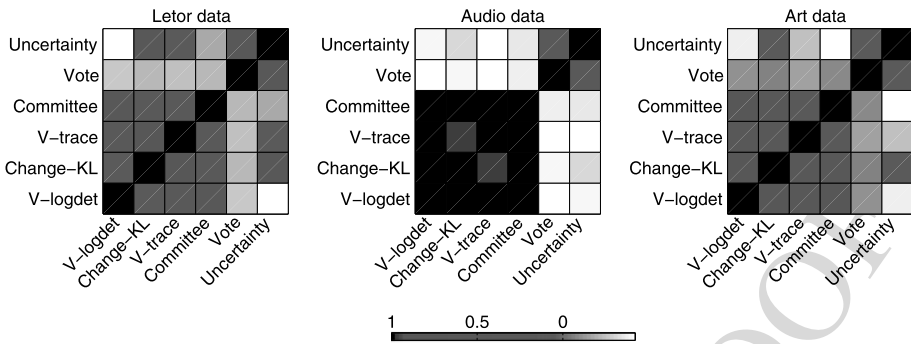
Mach Learn

1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067

Fig. 4 Spearman rank correlation coefficients for the scores assigned by the active learning criteria on the three data sets: letor data (*left*) audio data (*center*) and on the art data (*right*). The *darker the color*, the higher the correlations, suggesting that the criteria are very similar. The *lighter the color*, the lower the correlations, suggesting that the criteria are not similar. The Variance-logdet and Variance-trace criteria are referred in the plots as V-logdet and V-trace. The correlations between the Variance-logdet and Committee criterion are about 0.7 for the Letor data set, 0.95 for the audio data set and around 0.85 for the art data set

1068
1069
1070

4.5 Similarities between criteria

1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095

In order to test empirically the approximations and similarities from Sect. 3.4, we computed the Spearman rank correlations (Hollander and Wolfe 1973) between the scores assigned by the criteria when evaluating the informativeness of the data points. The correlations were computed for all the data sets considered in the experimental evaluation, Letor, audio and art data set. For each subject the learning started with the hierarchical prior learned from the data of the other subjects. This prior was updated by taking into account the information from 20 randomly selected data points for both data sets. After these updates, we computed the scores assigned by each of the active learning criteria to 50 randomly chosen data points. Figure 4 shows these Spearman rank correlation coefficients for each pair of criteria; the darker the color the closer to 1 the correlations are and the stronger the similarity between the two criteria. There are several observations to be made from this figure: (i) One can notice a darker square on the left-down part of the figures, for all three data sets. This square involves the Variance-logdet, Change-KL, Variance-trace, and Committee criteria. The correlations between each pair of them are very close to 1 which suggests that these criteria perform in practice very similarly. This is also what the theory from Sect. 3.4 suggests by approximating these criteria to a similar form. (ii) The Variance-logdet and Change-KL criterion have the Spearman rank correlation extremely close to 1. Their approximations are proven to be equivalent in Lemma 5 in the Appendix. These two observations also suggest that the approximations of the Variance-logdet and Change-KL are very accurate. (iii) The Vote criterion performs in some situations randomly since when the number of subjects is much smaller than the number of data points considered, the scores assigned by the Vote criterion are the same for most of the experiments. (iv) The Uncertainty criterion is most different from the others.

1096
1097

5 Conclusions and discussions

1098
1099
1100

This work studied how to exploit models learned on other scenarios to actively learn a model for a new scenario in an efficient way. Our approach to active learning in a multi-scenario

⁂ Springer

Mach Learn

setting combines a hierarchical Bayesian prior (to learn from related scenarios) with active learning (to learn efficiently by selecting informative examples). Our new Committee criterion inspired by the Query-by-Committee method is very similar to the standard criteria from experimental design, in particular in the early stages of active learning, but computationally more efficient. Aside from the computational advantage, the Committee criterion introduces the idea to have the data, available from other users, collaborate in order to select the most informative experiments to perform with a new user. The same idea is already implicit in the Query-by-Committee algorithm. We show, theoretically and through experiments, that this conceptual idea also works with a committee of people. This can be interpreted as another way of using people as the elements of a machine learning algorithm, which is a very promising research area, as suggested also by (Sanborn and Griffiths 2008).

There are several aspects related to the approach proposed here that require further attention: (i) The design is myopic in the sense that the active learning criteria look one step ahead at a time when evaluating the informativeness of a data point. A non-myopic design "looks" more than just one step and it is theoretically closer to the best possible design but computationally much more expensive. Due to the computational complexity involving a non-myopic design, we discussed all the active learning criteria from a myopic perspective, however, a non-myopic perspective can be applied to all of them, similar to the one proposed by Boutilier (2002). (ii) In this work we used log-linear models and Gaussian distributions to model the preference data. The same idea, of using models learned on data from different subjects (or scenarios) to actively select examples for a new subject, can be applied to other models and starting from different priors as well, although the mathematics will be a bit more involved and less intuitive. In particular, considering a mixture of Gaussians as the prior may still be feasible and may lead to an active learning strategy that tries to find those examples that can best discriminate to which mixture component the current model belongs.

Appendix

In this Appendix we prove the equivalences between the active learning criteria stated in Sect. 3.4. We show that these criteria can be approximated to the same form, namely

$$\sum_c p(c|a, \boldsymbol{\alpha}) \boldsymbol{g}(c|a, \boldsymbol{\alpha})^T \boldsymbol{Q} \boldsymbol{g}(c|a, \boldsymbol{\alpha}),$$

for some vector $\boldsymbol{\alpha}$ and matrix $\boldsymbol{Q}$. The difference between the approximations for different criteria is the point $\boldsymbol{\alpha}$ in which the gradients and the probabilities are evaluated and the weighting matrix of the gradients $\boldsymbol{Q}$.

We consider probabilistic choice models of the form given in Eq. (2), which by using the definition of the utility function from Eqs. (3) and (4) can be rewritten as

$$p(c|a, \boldsymbol{\alpha}) = \frac{\exp[\sum_{i=1}^D \phi_i(\boldsymbol{x}_c)\alpha_i]}{Z(\boldsymbol{\alpha}, a)} \quad \text{with } Z(\boldsymbol{\alpha}, a) \equiv \sum_{c'} \exp\left[\sum_i \phi_i(\boldsymbol{x}_{c'})\alpha_i\right].$$

Mach Learn

We define the derivatives of the log probabilities

$$\boldsymbol{g}(c|a,\boldsymbol{\alpha}) \equiv \frac{\partial \log p(c|a,\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}, \qquad \boldsymbol{H}(c|a,\boldsymbol{\alpha}) \equiv \frac{\partial^2 \log p(c|a,\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T}.$$

We first prove a lemma which states a relationship between the Hessian and the Fisher matrices which will be used in further proofs.

Lemma 3 *For any input a and vector $\boldsymbol{\alpha}$ we have the following relationship between the Hessian and the Fisher matrices*:

$$\sum_c p(c|a,\boldsymbol{\alpha})\boldsymbol{H}(c|a,\boldsymbol{\alpha}) = -\sum_c p(c|a,\boldsymbol{\alpha})\boldsymbol{g}(c|a,\boldsymbol{\alpha})\boldsymbol{g}(c|a,\boldsymbol{\alpha})^T.$$

Proof We use shorthand notation $p_c = p(c|a,\boldsymbol{\alpha})$, $g_{cj} = g_j(c|a,\boldsymbol{\alpha})$, $\phi_{ci} = \phi_i(\boldsymbol{x}_c)$, omitting the dependencies on a and $\boldsymbol{\alpha}$.

From $\log p_c = \sum_j \phi_{cj}\alpha_j - \log Z$, it is easy to see that

$$g_{cj} = \phi_{cj} - \frac{\partial \log Z}{\partial \alpha_j}, \qquad H_{c,ij} = -\frac{\partial^2 \log Z}{\partial \alpha_i \partial \alpha_j}.$$

Furthermore,

$$\frac{\partial \log Z}{\partial \alpha_j} = \frac{1}{Z}\frac{\partial Z}{\partial \alpha_j} = \frac{1}{Z}\sum_c \exp\left[\sum_{j'}\phi_{cj'}\alpha_{j'}\right]\phi_{cj} = \sum_c p_c\phi_{cj},$$

$$\frac{\partial^2 \log Z}{\partial \alpha_i \partial \alpha_j} = \sum_c \phi_{cj}\left[\frac{\exp(\sum_{j'}\phi_{cj'}\alpha_{j'})\phi_{ci}Z - \frac{\partial Z}{\partial \alpha_i}\exp(\sum_{j'}\phi_{cj'}\alpha_{j'})}{Z^2}\right]$$

$$= \sum_c \phi_{cj}\left[\phi_{ci}p_c - \sum_{c'}(p_{c'}\phi_{c'i})p_c\right] = \sum_c p_c\phi_{cj}\phi_{ci} - \sum_c p_c\phi_{cj}\sum_{c'}p_{c'}\phi_{c'i},$$

and thus

$$g_{cj} = \phi_{cj} - \sum_{c'}p_{c'}\phi_{c'j},$$

$$H_{c,ij} = -\sum_{c'}p_{c'}\phi_{c'j}\phi_{c'i} + \sum_{c'}p_{c'}\phi_{c'j}\sum_{c''}p_{c''}\phi_{c''i} = H_{ij}. \qquad (28)$$

Note that the second derivative is in fact independent of c. We then have

$$\sum_c p_c H_{c,ij} = \sum_c p_c H_{ij} = H_{ij} = -\sum_c p_c\phi_{ci}\phi_{cj} + \sum_c p_c\phi_{ci}\sum_{c'}p_{c'}\phi_{c'j}$$

$$= -\sum_c p_c\left(\phi_{ci} - \sum_{c'}p_{c'}\phi_{c'i}\right)\left(\phi_{cj} - \sum_{c'}p_{c'}\phi_{c'j}\right) = -\sum_c p_c g_{ci}g_{cj}. \qquad \square$$

The following lemma proves the approximation of the Variance-trace criterion from Eq. (26).

🖄 Springer

Mach Learn

Lemma 4 *In a first order approximation, the Variance-trace criterion boils down to*

$$\text{Variance-trace}(a) = \sum_c p(c|a, \boldsymbol{\mu}) \boldsymbol{g}(c|a, \boldsymbol{\mu})^T \boldsymbol{\Sigma}^2 \boldsymbol{g}(c|a, \boldsymbol{\mu}).$$

Proof We have

$$\boldsymbol{\Sigma}_{(a,c)} = \left(\boldsymbol{\Sigma}_{(a,c)}^{-1}\right)^{-1} = \left(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_{(a,c)}^{-1} - \boldsymbol{\Sigma}^{-1}\right)^{-1} \approx -\boldsymbol{\Sigma}\left[\boldsymbol{\Sigma}_{(a,c)}^{-1} - \boldsymbol{\Sigma}^{-1}\right]\boldsymbol{\Sigma}.$$

Use of Eq. (23) and Lemma 3 gives the result. □

The following lemma proves the approximation of the Change-KL criterion from Eq. (27).

Lemma 5 *In a first order approximation, assuming that* $\boldsymbol{\Sigma}_{(a,c)}$ *is close to* $\boldsymbol{\Sigma}$, *we have*

$$\text{Change-KL}(a) \approx \text{Variance-logdet}(a),$$

i.e., the two criteria are indistinguishable.

Proof We evaluate the terms of the Change-KL criterion one by one,

$$\text{Change-KL}(a) = \sum_c p(c|a)\left[\log\left(\frac{\det \boldsymbol{\Sigma}_{(a,c)}}{\det \boldsymbol{\Sigma}}\right) + \text{Tr}\left(\boldsymbol{\Sigma}_{(a,c)}^{-1} \boldsymbol{\Sigma}\right)\right]$$
$$+ \sum_c p(c|a)\left[(\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_{(a,c)}^{-1}(\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu}) - n\right].$$

The first term gives

$$\sum_c p(c|a) \log\left(\frac{\det \boldsymbol{\Sigma}_{(a,c)}}{\det \boldsymbol{\Sigma}}\right) = -\text{Variance-logdet}(a).$$

The second term

$$\sum_c p(c|a) \text{Tr}\left[\boldsymbol{\Sigma}_{(a,c)}^{-1} \boldsymbol{\Sigma}\right] = \sum_c p(c|a) \text{Tr}\left[\left(\boldsymbol{\Sigma}^{-1} - \boldsymbol{H}(c|a, \boldsymbol{\mu})\right)\boldsymbol{\Sigma}\right]$$
$$= n - \sum_c p(c|a) \text{Tr}\left[H(c|a, \boldsymbol{\mu})\boldsymbol{\Sigma}\right]$$
$$= n - \sum_c p(c|a, \boldsymbol{\mu}) \text{Tr}\left[\boldsymbol{\Sigma} H(c|a, \boldsymbol{\mu})\right]$$
$$= n + \text{Variance-logdet}(a).$$

In the same lowest order, we obtain for the third term

$$\sum_c p(c|a)(\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_{(a,c)}^{-1}(\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu})$$
$$\approx \sum_c p(c|a, \boldsymbol{\mu}) \boldsymbol{g}(c|a, \boldsymbol{\mu})^T \boldsymbol{\Sigma} \boldsymbol{g}(c|a, \boldsymbol{\mu}) = \text{Variance-logdet}(a).$$

Collection of all the terms then gives the result. □

Mach Learn

References

Adomavicius, A., Sankaranarayanan, R., Sen, S., & Tuzhilin, A. (2005). Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems*, *23*(1), 103–145.

Anand, P. (1993). The philosophy of intransitive preferences. *The Economic Journal*, *103*(417), 337–346.

Arehart, K. H., Kates, J. M., Anderson, C. A., & Harvey, L. O. Jr. (2007). Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, *122*(2), 1150–1164.

Arens, R. (2008). Learning SVM ranking function from user feedback using document metadata and active learning in the biomedical domain. In *Proceedings of the ECML/PKDD workshop on preference learning* (pp. 363–383).

Argyriou, A., Micchelli, C., & Pontil, M. (2008). When is there a representer theorem? Vector versus matrix regularizers. *Journal of Machine Learning Research*.

Bakker, B., & Heskes, T. (2003). Task clustering and gating for Bayesian multitask learning. *Journal of Machine Learning Research*, *4*, 83–99.

Berger, M. P. F. (1994). D-optimal sequential sampling designs for item response theory models. *Journal of Educational and Behavioral Statistics*, *19*, 43–56.

Birlutiu, A., & Heskes, T. (2007). Expectation propagation for rating players in sports competitions. In *Proceedings of the 11th European conference on principles and practice of knowledge discovery in databases* (pp. 374–381).

Birlutiu, A., Groot, P., & Heskes, T. (2009). Multi-task preference learning with an application to hearing aid personalization. *Neurocomputing*, *73*(7–9), 1177–1185.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Blythe, J. (2002). Visual exploration and incremental utility elicitation. In *Proceedings of the 18th national conference on artificial intelligence* (pp. 526–532).

Bordley, R. F. (1982). A multiplicative formula for aggregating probability assessments. *Management Science*, *28*(10), 1137–1148.

Boutilier, C. (2002). A POMDP formulation of preference elicitation problems. In *Proceedings of the 18th national conference on artificial intelligence* (pp. 239–246).

Boutilier, C., Zemel, R. S., & Marlin, B. (2003). Active collaborative filtering. In *Proceedings of the 19th annual conference on uncertainty in artificial intelligence* (pp. 98–106).

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs, I: the method of paired comparisons. *Biometrika*, *39*, 324–345.

Brinker, K. (2004). Active learning of label ranking functions. In *Proceedings of the 27th international conference on machine learning*.

Brochu, E., de Freitas, N., & Ghosh, A. (2008). Active preference learning with discrete choice data. In J. C. Platt, Y. Koller, D. Singer & S. Roweis (Eds.), *Advances in neural information processing systems* (Vol. 20, pp. 409–416). Cambridge: MIT Press.

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*(1), 41–75.

Chajewska, U., Koller, D., & Parr, R. (2000). Making rational decisions using adaptive utility elicitation. In *Proceedings of the 17th national conference on artificial intelligence* (pp. 363–369).

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: a review. *Statistical Science*, *10*, 273–304.

Christensen, R. (1997). *Log-linear models and logistic regression*. Berlin: Springer.

Chu, W., & Ghahramani, Z. (2005a). Extensions of Gaussian processes for ranking: semi-supervised and active learning. In *NIPS workshop on learning to rank*.

Chu, W., & Ghahramani, Z. (2005b). Preference learning with Gaussian processes. In *Proceedings of the 22nd international conference on machine learning*.

Clyde, M., Muller, P., & Parmigiani, G. (1993). Optimal design for heart defibrillators. In *Case studies in Bayesian statistics* (Vol. II, pp. 278–292).

Cohn, D. A., Ghahramani, A., & Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, *4*(1), 129–145.

Dagan, I., & Engelson, S. P. (1995). Committee-based sampling for training probabilistic classifiers. In *Proceedings of the 12th international conference on machine learning* (pp. 150–157).

Dasgupta, S., & Hsu, D. (2008). Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on machine learning* (pp. 208–215).

Doyle, J. (2004). Prospects for preferences. *Computational Intelligence*, *20*(2), 111–136.

Springer

Mach Learn

Dror, H. A., & Steinberg, D. M. (2008). Sequential experimental designs for generalized linear models. *Journal of the American Statistical Association*, *103*(481), 288–298.

Evgeniou, T., Micchelli, A., & Pontil, M. (2005). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, *6*, 615–637.

Fedorov, V. V. (1972). *Theory of optimal experiments*. New York: Academic Press.

Ford, I., & Silvey, S. D. (1980). A sequentially constructed design for estimating a nonlinear parametric function. *Biometrika*, *67*, 381–388.

Freund, Y., Shamir, E., & Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Machine Learning*, *28*(2–3), 133–168.

Fürnkranz, J., & Hüllermeier, E. (2010). *Preference learning*. Berlin: Springer.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). London: Chapman & Hall/CRC.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58.

Gervasio, M. T., Moffitt, M. D., & Pollack, M. E. (2005). Active preference learning for personalized calender scheduling assistance. In *Proceedings of the 10th international conference on intelligent user interfaces* (pp. 90–97).

Glickman, M., & Jensen, S. (2005). Adaptive paired comparison design. *Journal of Statistical Planning and Inference*, *127*, 279–293.

Groot, P. C., Birlutiu, A., & Heskes, T. (2010). Bayesian Monte Carlo for the global optimization of expensive functions. In *Proceedings of the 19th European conference on artificial intelligence* (pp. 249–254).

Guo, S., & Sanner, S. (2010). Real-time multiattribute Bayesian preference elicitation with pairwise comparison queries. In *Proceedings of the 13th international conference on artificial intelligence and statistics* (pp. 289–296).

Harpale, S., & Yang, Y. (2008). Personalized active learning for collaborative filtering. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 91–98).

Heskes, T., & de Vries, B. (2005). Incremental utility elicitation for adaptive personalization. In *Proceedings of the 17th Belgium-Netherlands conference on artificial intelligence* (pp. 127–134).

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods*. New York: Wiley.

Jin, R., & Si, L. (2004). A Bayesian approach toward active learning for collaborative filtering. In *Proceedings of the 20th conference on uncertainty in artificial intelligence* (pp. 278–285).

Kanninen, B. (2002). Optimal design for multinomial choice experiments. *Journal of Marketing Research*, *39*, 307–317.

Lewi, J., Butera, R., & Paninski, L. (2009). Sequential optimal design of neurophysiology experiments. *Neural Computation*, *21*(3), 619–687.

Lewis, D., & Gale, W. (1994). A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 3–12).

MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Computation*, *4*, 590–604.

MacKay, D. J. C. (2002). *Information theory, inference & learning algorithms*. Cambridge: Cambridge University Press.

McCallum, A., & Nigam, K. (1998). Employing EM and pool-based active learning for text classification. In *Proceedings of the 15th international conference on machine learning* (pp. 350–358).

Melville, P., & Mooney, R. (2004). Diverse ensembles for active learning. In *Proceedings of the 21st international conference on machine learning* (pp. 584–591).

Minka, T. (2001). *A family of approximation methods for approximate Bayesian inference*. PhD thesis, MIT.

Pahikkala, T., Waegeman, W., Tsivtsivadze, E., De Baets, B., & Salakoski, T. (2009). From ranking to intransitive preference learning: rock-paper-scissors and beyond. In *Proceedings of the ECML/PKDD workshop on preference learning* (pp. 84–100).

Qin, T., Liu, T. Y., Xu, J., & Li, H. (2010). LETOR: a benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, *13*(4), 346–374.

Sanborn, A. N., & Griffiths, T. L. (2008). Markov chain Monte Carlo with people. In: *Advances in neural information processing systems*.

Schein, A., & Ungar, L. (2007). Active learning for logistic regression: an evaluation. *Machine Learning*, *68*(3), 235–265.

Seeger, M. W. (2008). Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, *9*, 759–813.

Settles, B., & Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1070–1079).

<uncited>

🖄 Springer

Seung, H. S., Opper, M., & Sompolinsky, H. (1992). Query by Committee. In *Proceedings of the 5th annual workshop on computational learning theory* (pp. 287–294).

Thrun, S. (1995). Is learning the *n*th thing any easier than learning the first? In *Advances in neural information processing systems* (pp. 640–646).

Tversky, A. (1998). *Preference, belief, and similarity*. Cambridge: MIT Press.

Xu, Z., Kersting, K., & Joachims, T. (2010). Fast active exploration for link-based preference learning using Gaussian processes. In *Proceedings of the 2010 European conference on machine learning and knowledge discovery in databases: part III* (pp. 499–514).

Yu, K., Schwaighofer, A., Tresp, V., Ma, W. Y., & Zhang, H. J. (2003). Collaborative ensemble learning: combining collaborative and content-based information filtering. In *Proceedings of the 19th conference on uncertainty in artificial intelligence* (pp. 616–623).

Yu, K., Tresp, V., & Schwaighofer, A. (2005). Learning Gaussian processes from multiple tasks. In *Proceedings of the 22nd international conference on machine learning*.

Yu, K., Bi, J., & Tresp, V. (2006). Active learning via transductive experimental design. In *Proceedings of the 23rd international conference on machine learning* (pp. 1081–1088).

🍏 Springer