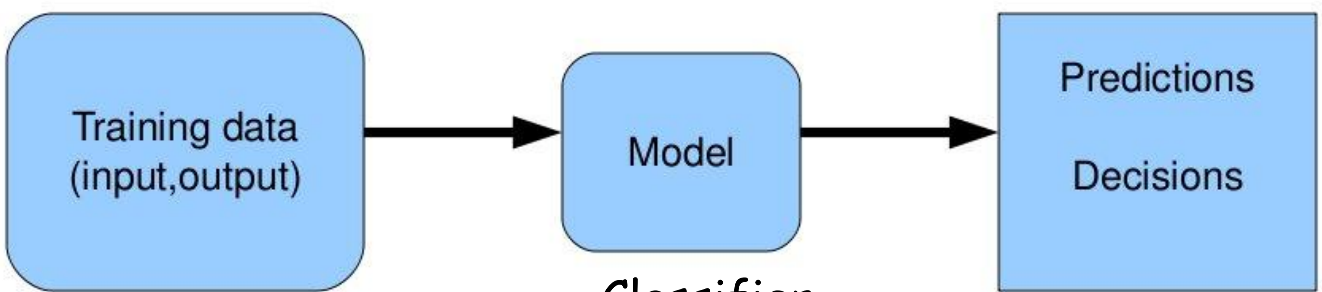
Transfer Learning from a Machine Learning Perspective

Adriana Bîrluțiu
Universitatea „1 Decembrie 1918"
Alba Iulia

Machine Learning

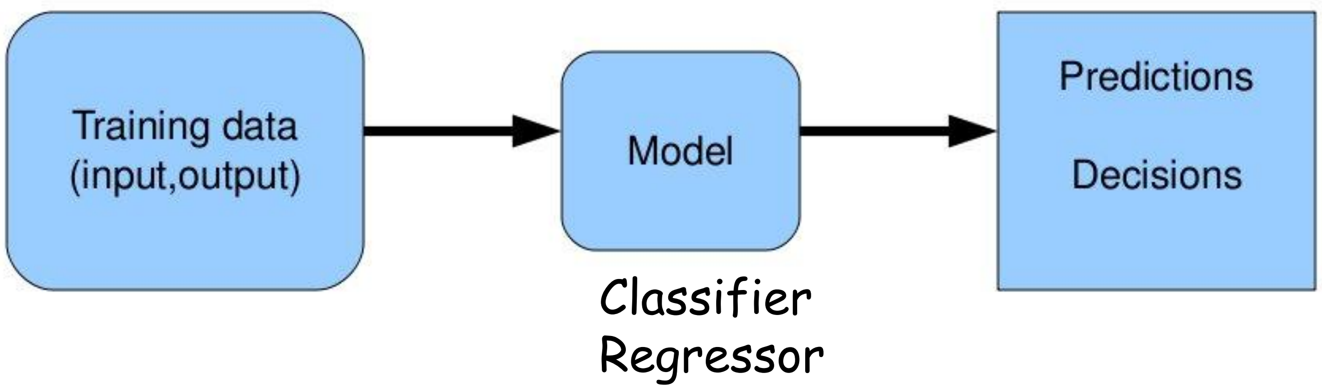
- Branch of AI focused on the design and development of methods that allow machines to learn based on observations
- Various applications
- Success due to increasing availability of empirical data and computational power



Classifier
Regressor

Machine Learning

- Branch of AI focused on the design and development of methods that allow machines to learn based on observations
- Varioua applications
- Success due to increasing availability of empirical data and computational power



- Obtaining labeled data to train the algorithms is expensive!

Efficient machine learning

Characteristics of (human) learning:

- Based on prior experience
  - transfer learning (e.g. C++ -> Java)
- Selects the most useful information
  - active learning

Efficient machine learning

Characteristics of (human) learning:

- Based on prior experience
  - transfer learning (e.g. C++ -> Java)
- Selects the most useful information
  - active learning

Transfer learning

- Fundamental assumption in machine learning:
  - Data is i.i.d. (independent and identically distributed)
  - Training and test data stem from the same distribution
- Often, this assumption does not hold
- Transfer learning addresses the mismatch between training and test data

Traditional vs transfer learning

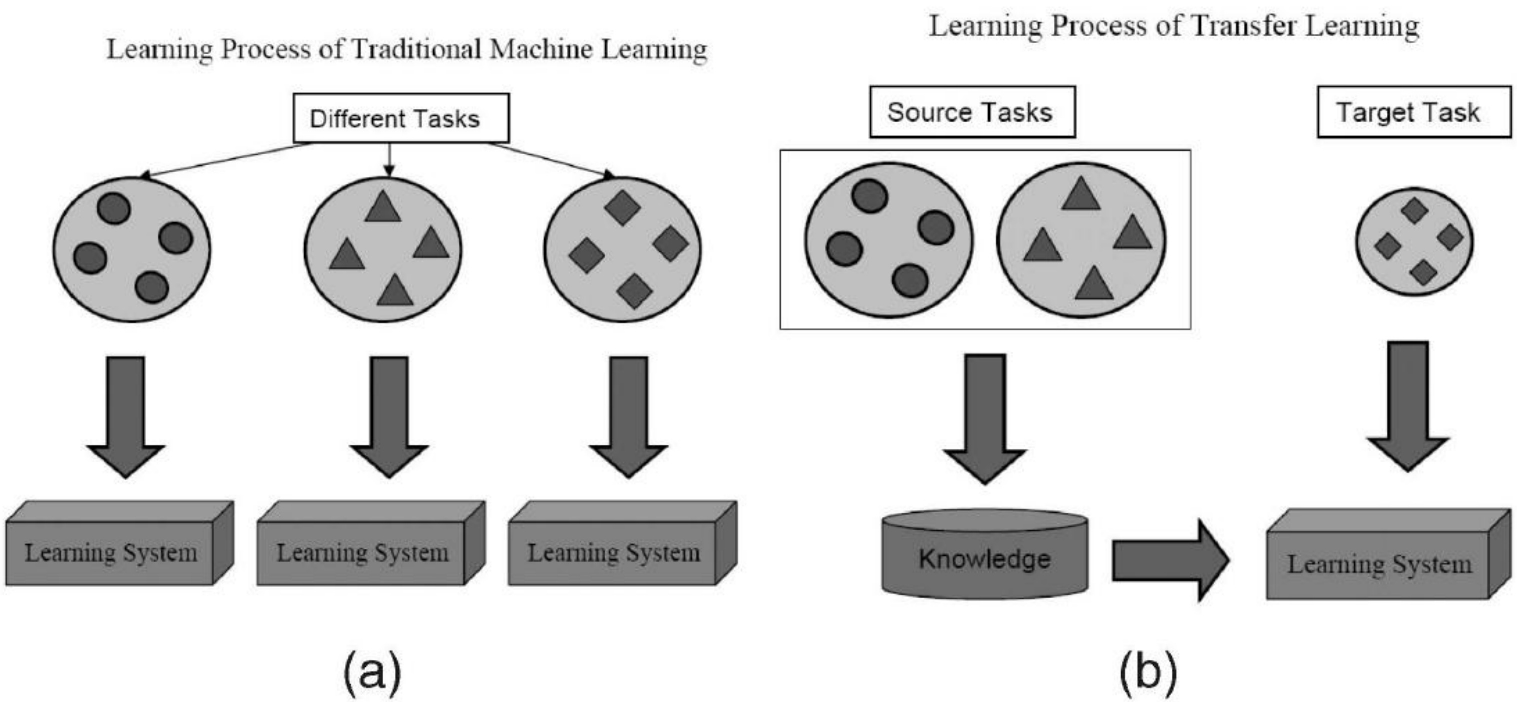


Fig. 1. Different learning processes between (a) traditional machine learning and (b) transfer learning.

(figure adapted from Pan et. al, IEEE TNN 2011)

Questions

Given a target task and some previous source tasks, the questions are:

- How to identify the commonality between the target task and the previous tasks?
- How to transfer knowledge from the previous tasks to the target one?

Approaches

- Instance-based: reweighted source data are used for learning in the target space
- Parameter-based: source and target model share some common parameters or a prior distribution
- Feature-based: source knowledge is used for learning a good feature representation in the target space

Outline

- Overview on transfer learning
- Approaches
  - Feature-based methods
    - Transfer component analysis (TCA) for a chemometric application
  - Parameter-based methods
    - Application: Personalization of hearing-aids based on hierarchical modeling

TL for chemometric application

- Application: control the polymerization process of melamine based on spectroscopic data, measured in-line at an industrial partner

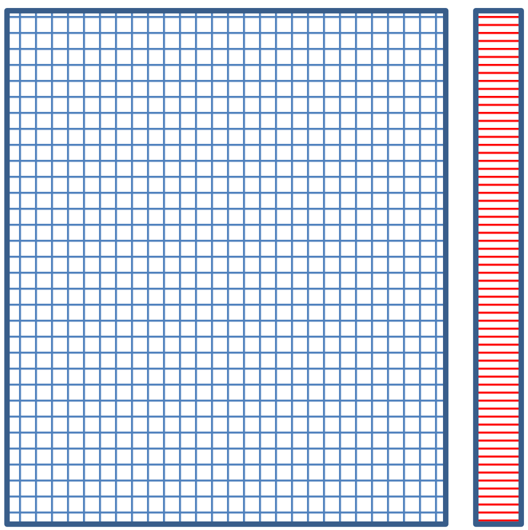
- Regression problem: predict the temperature of a sample based on spectroscopic data

- Transfer learning settings:
  1. Change of lamp
  2. Change of reactor + optical fibre
  3. Recipe change
  4. ....

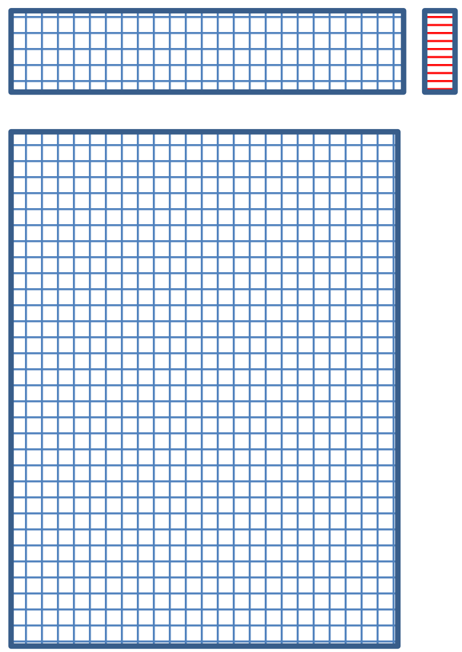
- as very often:
  - Unlabelled data (spectra) easy to obtain
  - Labels (reference values) cumbersome / expensive to measure

Data

Source data:
spectra + reference values



Target data:
A few spectra + reference values
A lot of spectra without reference values
(unlabelled data)



Question:
How do we combine these data?

Approach

- Use only target labelled data and ignore any other source data (no transfer)
- Use source data + target labelled data (all labelled data pulled together)
- Combine all data using a more "sophisticated" method

Transfer component analysis (TCA)

- Source domain (S), target domain (T)
- Assumption: $\mathrm{P}(\mathrm{X}_S) \neq P(\mathrm{X}_T)$
  - holds for the chemometric application since the conditions under which the spectra were obtained are different between domains

- Intuition: discover a good feature representation across domains
- Idea: maps data in a shared subspace s.t.
  - distance between distributions is minimized
  - data properties are preserved

- Goal: find a feature map $\phi: X \rightarrow H$ where H is a RKHS such that

  $P\big(\phi(X_S)\big) \approx P\big(\phi(X_T)\big)$ s.t. Data properties are preserved
- Key assumption:

  $P(X_S) \neq P(X_T)$ but $P(Y_S|\phi(X_S)) = P(Y_T|\phi(X_T))$

[Pan et. al, 2011, IEEE TNN]

Multi-TCA

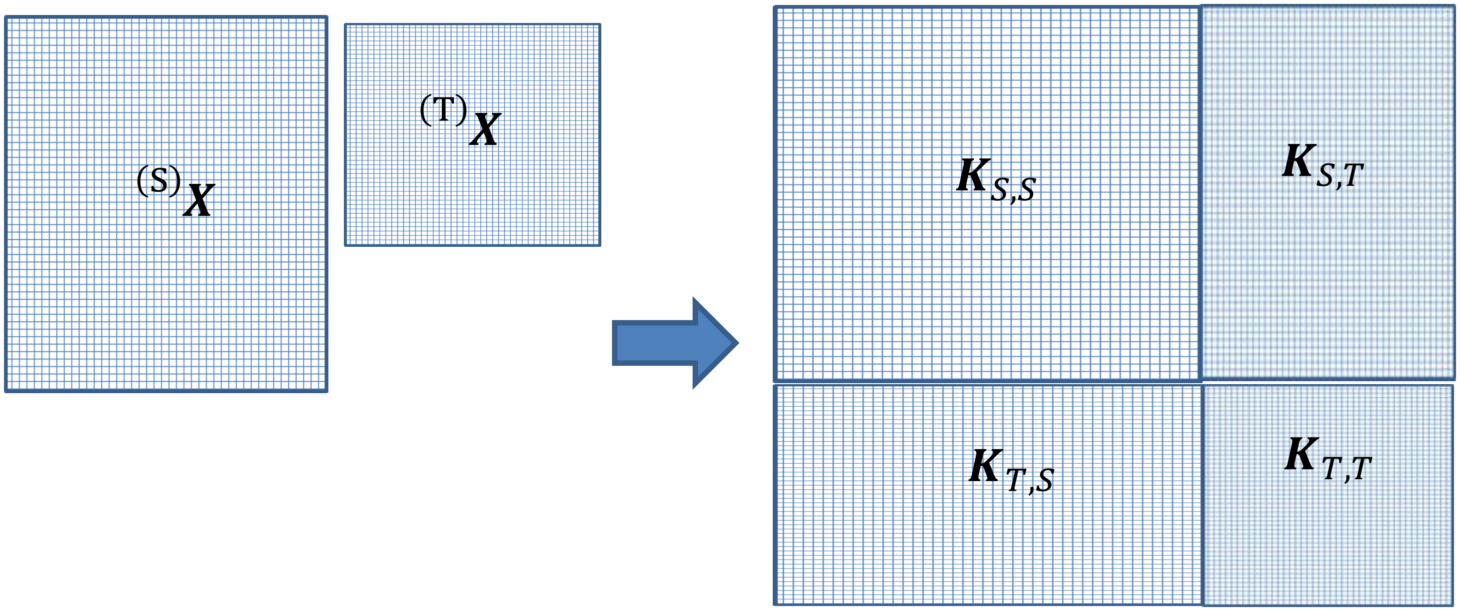
- Multi-TCA an extension of TCA to multiple source and target domains

  [Grubinger, Birlutiu et al. 2015, IWANN]
- Domain generalization: no input data from target domains but the characteristics of target data are sufficiently captured by X1, X2,…XS
- Goal: find $\phi: X \rightarrow H$ a feature map and H a RKHS

  $P\big(\phi(X_1)\big) \approx \cdots \approx P\big(\phi(X_S)\big)$ under some constraints

- Distance between distributions, e.g.:
  - Kullback-Leibler divergence,
  - Maximum Mean Discrepancy [Gretton et. al 2007]: distance between distributions = distance between the means of the two samples mapped in a RKHS

Use kernels for finding ϕ



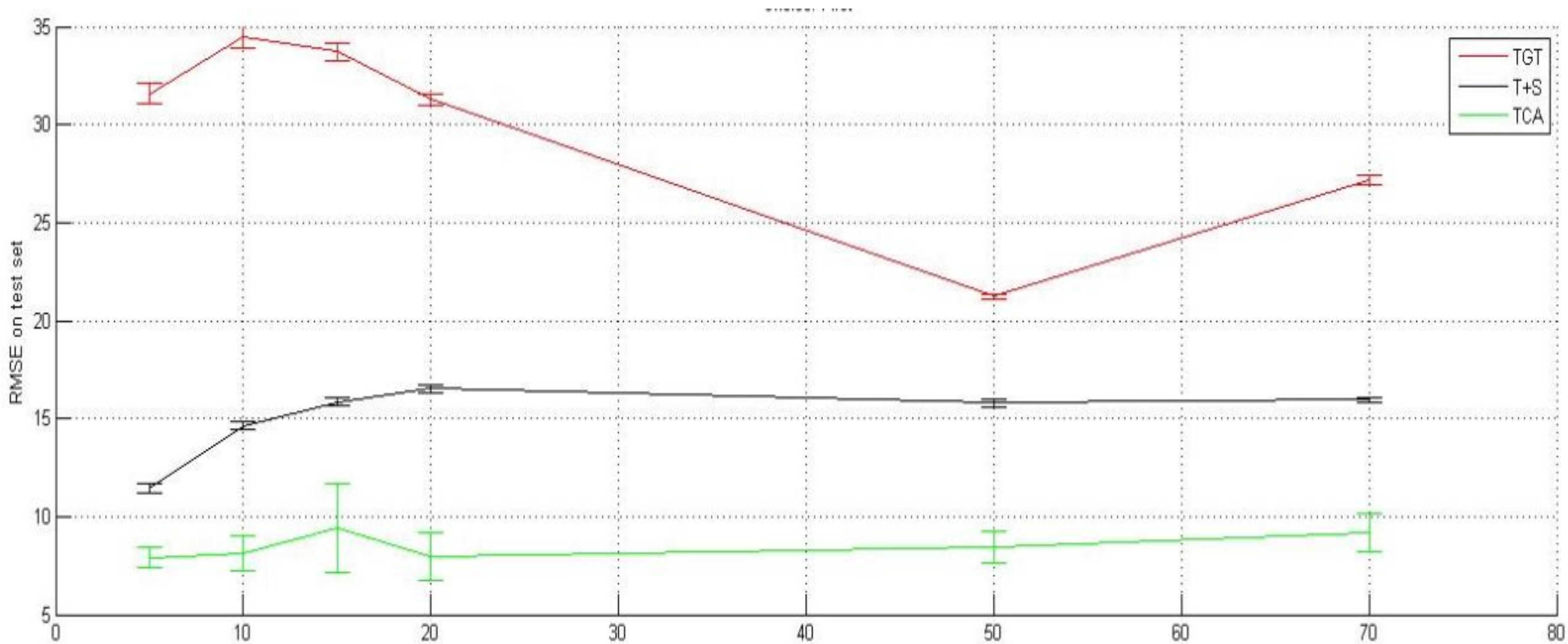
Kernel trick: $k(x_i, x_j) = \phi(x_i)'\phi(x_j)$

$$\mathrm{Dist}(X'_S, X'_T) = \mathrm{tr}(KL),$$

$$K = \begin{bmatrix} K_{S,S} & K_{S,T} \\ K_{T,S} & K_{T,T} \end{bmatrix}$$

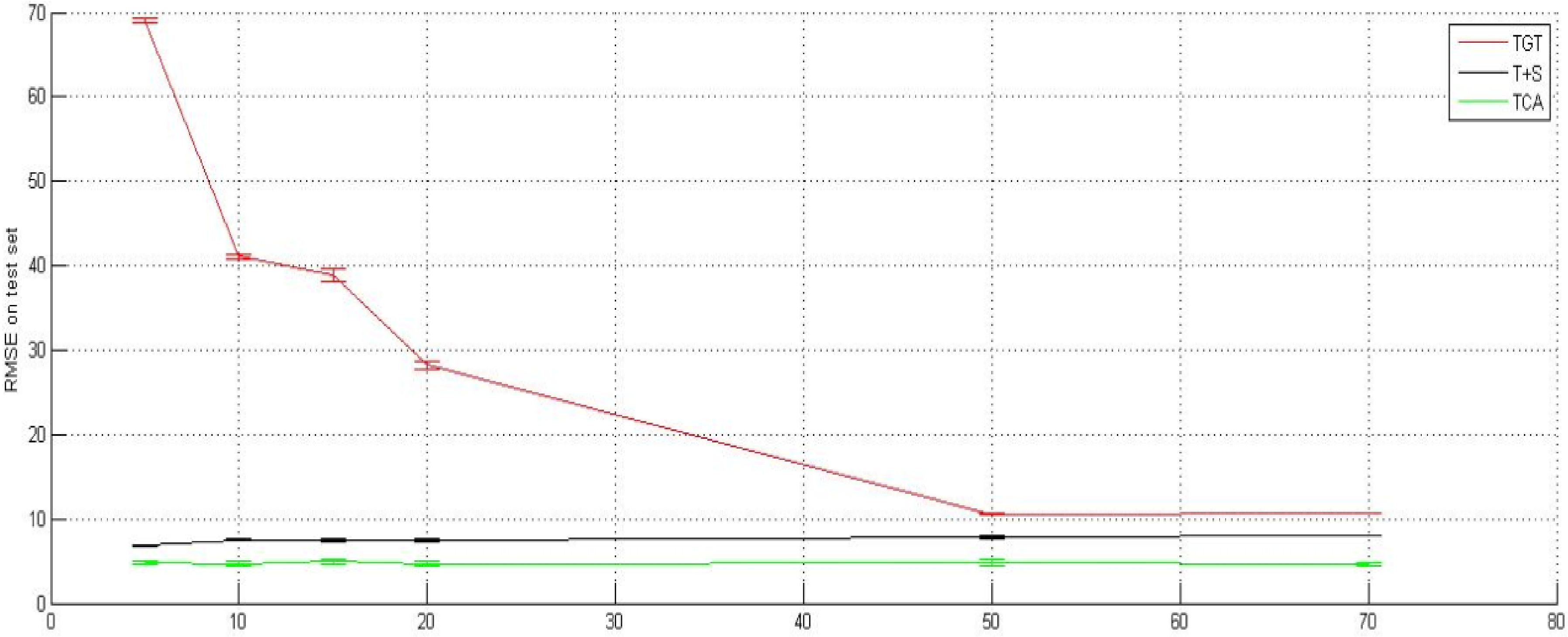
and $L = [L_{ij}] \succeq 0$ with $L_{ij} = \frac{1}{n_1^2}$ if $x_i, x_j \in X_S$; $L_{ij} = \frac{1}{n_2^2}$ if $x_i, x_j \in X_T$; otherwise, $-\frac{1}{n_1 n_2}$.

Experimental evaluation



Difference between source and target domains: Change of reactor

Experimental evaluation



Difference between source and target domains: Change of lamp

Outline

- Overview on transfer learning
- Approaches
  - Feature-based methods
    - Transfer component analysis (TCA) for a chemometric application
  - Parameter-based methods
    - Application: Personalization of hearing-aids based on hierarchical modeling

Personalization of hearing-aids

- People are not quite satisfied with hearing-aids
- Hundred parameters for hearing-aids

Goal: tune the parameters to maximize user satisfaction

Problems:

- Large dimensionality of the parameter space
- Determinants of hearing-impaired user satisfaction are unknown
- Listening tests are costly and unreliable

=> Personalized fitting based on a probabilistic framework

Approach to Hearing Aid Personalization

- Experimental setup: Patient listens to sounds, each processed with 2 hearing-aid parameters and says which of the 2 he prefers
- Learning: Based on these preferences a model is learned
- Decision making: Given this model we can compute the optimal setting of the hearing-aid parameters

Use Bayesian theory to perform less listening experiments and to compute the optimal setting of the parameters

- Knowledge transfer: use the audiological similarities between patients to learn a joint prior probability
  [Birlutiu et. al, 2010, Neurocomputing]
- Experiment selection: select the listening experiments that in expectation give the most information about the patient's preferences
  [Birlutiu et. al. 2013, Machine Learning]

Probabilistic choice models

Qualitative preference observations: $X = \{x_1, \ldots, x_n\}$ a set of inputs, D a set of observed preference comparisons over instances in X corresponding to a user

$$D = \{(a_j, c_j) | 1 \leq j \leq J, c_j \in \{1, \ldots, A\}\}$$

with $a_j = (\mathbf{x}_{i_1(j)}, \ldots, \mathbf{x}_{i_A(j)})$ the alternatives presented and c_j the choice made

A latent utility function value $U(x_i)$ associated with each input x_i captures the individual preference of a subject for x_i

The probability that the cth alternative is chosen by the subject in the jth comparison follows a multinomial logistic model (Bradley-Terry model)

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

Utility model

$U : X \to \mathbb{R}$, where each input x is characterized by a set of features $\phi(x) \in R^p$

$$U(x) = \sum_{k=1}^{p} w_k \phi_k(x)$$

$w = (w_1, \ldots, w_p)$ is a vector of weights which captures the importance of each feature of the input x when evaluating the utility U for a specific user, $\phi_k(x)$ are the components of the vector $\phi(x)$

The preferences of a user are encoded in the vector w and learning the utility function for a user reduces to learning w

Bayesian updating

Bayesian framework in which *the vector of parameters w is treated as a random variable*

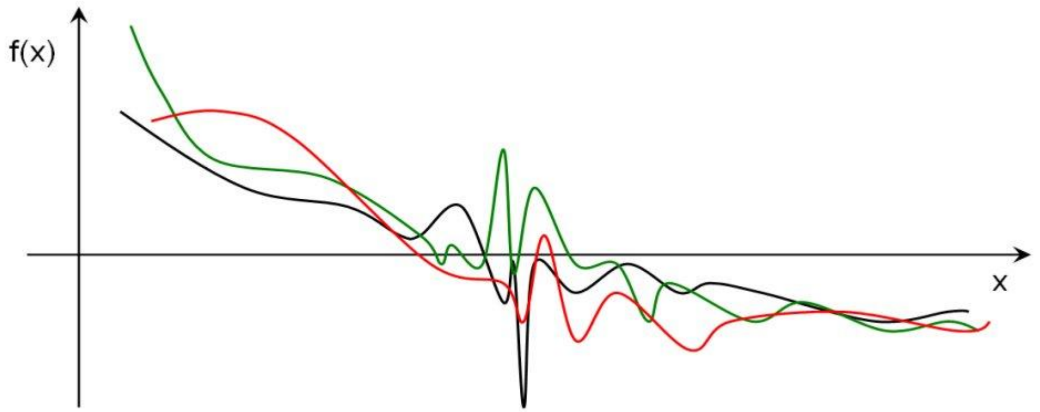
We consider a Gaussian prior distribution over w which is updated based on the observations from the preference comparisons using Bayes' rule

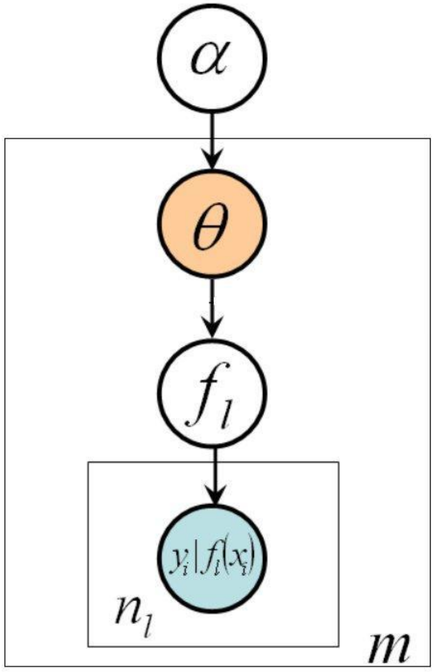
$$p(w|D, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto p(w|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{j=1}^{J} p(c_j|a_j, w)$$

- Likelihood is the probabilistic choice model
- The posterior distribution obtained is approximated to a Gaussian
- *Incremental Bayesian updating* of the utility model: the prior is the posterior distribution from the previous step

Multi-task learning
Hierarchical modeling

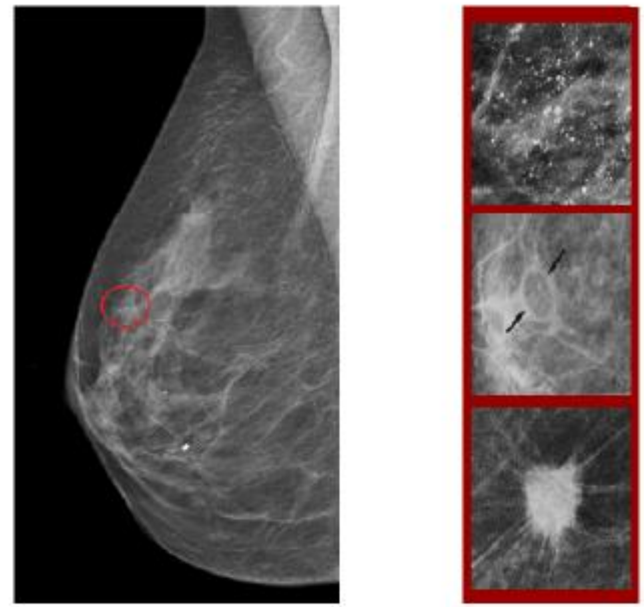
Learning multiple related functions





Current work: Transfer learning in medical image analysis

- How to combine synthetic with real data?
- How to combine image data obtained with different modalities?



Imbunătățirea calității imaginii folosind învățarea automată

Some Research Issues

- How to avoid negative transfer? Given a target domain/task, how to find source domains/tasks to ensure positive transfer
- Transfer learning meets active learning
- Given a specific application, which kind of transfer learning methods should be used?

Thank you!