Learning from Multiple Annotators with Gaussian Processes

Perry Groot[1], Adriana Birlutiu[2], and Tom Heskes[2]

[1] Technical University Eindhoven - Dept. of Electrical Engineering - Control Systems
Potentiaal 4.28, 5600 MB Eindhoven - the Netherlands
[2] Radboud University Nijmegen - Intelligent Systems
Heyendaalseweg 135, 6525 AJ Nijmegen - the Netherlands
pcgroot@gmail.com, {A.Birlutiu,t.heskes}@science.ru.nl

Abstract. In many supervised learning tasks it can be costly or infeasible to obtain objective, reliable labels. We may, however, be able to obtain a large number of subjective, possibly noisy, labels from multiple annotators. Typically, annotators have different levels of expertise (i.e., novice, expert) and there is considerable diagreement among annotators. We present a Gaussian process (GP) approach to regression with multiple labels but no absolute gold standard. The GP framework provides a principled non-parametric framework that can automatically estimate the reliability of individual annotators from data without the need of prior knowledge. Experimental results show that the proposed GP multi-annotator model outperforms models that either average the training data or weigh individually learned single-annotator models.

1 Introduction

In most learning settings a function is learned from inputs to outputs and it is assumed that outputs are available for training. In contrast to this, there are many real-world scenarios in which the true values of the outputs used for training are unknown or very expensive to obtain. Instead, multiple annotators are available to provide subjective, noisy estimates of these outputs. For example, annotators can be radiologists [1, 2] who provide a subjective (possible noisy) opinion about a suspicious region on a medical image as being malignant or benign. The actual gold standard (whether it is cancer or not) can be obtained from biopsies, which is an expensive and invasive procedure. Text and image classification are other learning scenarios where multiple human annotators subjectively assign inputs to some categories [3–8]. The amount of noise in the annotators' estimates of the true output can range from very few (annotators which are domain experts) to very much (annotators which are only novice).

Learning with multiple annotators is a special case of supervised learning in which a function $f : \mathcal{X} \to \mathcal{Y}$ is learned given a training data set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$. It can be considered in various learning settings, i.e., regression, binary and multi-class classification, and ranking. Previous work on learning with multiple annotators has focused on learning f using parametric models [1, 2, 9]. In

this paper we present a non-parametric approach based on Gaussian processes (GPs) [10]. Gaussian processes provide a rich, principled, and well-established alternative to parametric models. We provide details on predictive equations and hyperparameter optimization for regression with multiple annotators.

The rest of this paper is structured as follows. Section 2 gives background on Gaussian process regression. Section 3 describes regression with multiple annotators. Section 4 presents an experimental evaluation of our approach. Section 5 concludes and discusses some directions for future work.

2 Gaussian Process Regression

We denote vectors $\boldsymbol{x}$ and matrices $\boldsymbol{K}$ with bold-face type and their components with regular type, i.e., x_i, K_{ij}. With $\boldsymbol{x}^T$ we denote the transpose of the vector $\boldsymbol{x}$. Let $\boldsymbol{x} \in \mathbb{R}^D$ be an input, $y \in \mathbb{R}$ an output, and let $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ a set of N observed input-output pairs. Denote with $\boldsymbol{X} = \{\boldsymbol{x}_1 \ldots, \boldsymbol{x}_N\}$ and $\boldsymbol{Y} = \{y_1, \ldots, y_N\}$ the inputs and outputs occurring in $\mathcal{D}$. We assume that $\mathcal{D}$ is generated by an unknown function $f : \mathbb{R}^D \to \mathbb{R}$ where the observations are possibly corrupted with Gaussian noise, i.e., $y_i = f(\boldsymbol{x}_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. Usually, the noise is taken uniform $\sigma_i^2 = \sigma^2$ for every input $\boldsymbol{x}_i$, but this notation is consistent with the multi-annotator model described in Section 3 in which the noise model will be input dependent and the annotator identities involved.

A Gaussian process (GP) is a collection of random variables, here $\{f(\boldsymbol{x}_i)\}_{i \in I}$ for some index set I, any finite number of which have a joint Gaussian distribution [10]. A GP $\boldsymbol{f} \sim \mathcal{GP}(\boldsymbol{m}, \boldsymbol{K})$ specifies a prior distribution on functions and is completely specified by its mean function and kernel function. The kernel function k (also called (co)variance function) is a function mapping two arguments into $\mathbb{R}$ that is symmetric, i.e., $k(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}', \boldsymbol{x})$, and positive definite, i.e., $\boldsymbol{z}^T \boldsymbol{K} \boldsymbol{z} > 0$ for all nonzero vectors $\boldsymbol{z}$ with real entries ($z_d \in \mathbb{R}$) and $\boldsymbol{K}$ as defined below. The kernel function can be used to construct a covariance matrix $\boldsymbol{K}$ with respect to a set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ by defining $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. An often used kernel is the Gaussian kernel:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{A}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)\right), \tag{1}$$

where $\boldsymbol{A} = \mathrm{diag}(\ell_1^2, \ldots, \ell_D^2)$ with hyperparameters specifying the signal variance (σ_f^2) and specifying the length-scales, i.e., how much the function can vary for each input dimension (ℓ_d for $d = 1, \ldots, D$). A kernel function k leads to a multivariate Gaussian prior distribution on any finite subset $\boldsymbol{f} \subseteq \{f(\boldsymbol{x}_i)\}_{i \in I}$ of function values

$$p(\boldsymbol{f}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\boldsymbol{K}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \boldsymbol{f}^T \boldsymbol{K}^{-1} \boldsymbol{f}\right). \tag{2}$$

Given a GP prior over functions and a set of observations $\mathcal{D}$, a posterior distribution $p(\boldsymbol{f}|\mathcal{D})$ can be computed that can be used to make predictions at new

test points $\boldsymbol{x}, \boldsymbol{x}'$. Let $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_N^2)$. The standard predictive equations for regression with a zero-mean GP are given by [10]

$$\begin{aligned} \overline{f}_{\mathcal{D}}(\boldsymbol{x}) &= k(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{K} + \boldsymbol{\Sigma})^{-1}\boldsymbol{Y}, \\ \text{cov}_{\mathcal{D}}(f(\boldsymbol{x}), f(\boldsymbol{x}')) &= k(\boldsymbol{x}, \boldsymbol{x}') - k(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{K} + \boldsymbol{\Sigma})^{-1}k(\boldsymbol{X}, \boldsymbol{x}'). \end{aligned} \tag{3}$$

Maximum likelihood estimates for the hyperparameters can be obtained by minimizing the negative log marginal likelihood (e.g., using gradient descent), which can be evaluated exactly in the case of GP regression and is given by [10]

$$-\log p(\boldsymbol{Y}|\boldsymbol{X}) = \frac{1}{2}\boldsymbol{Y}^T(\boldsymbol{K} + \boldsymbol{\Sigma})^{-1}\boldsymbol{Y} + \frac{1}{2}\log|\boldsymbol{K} + \boldsymbol{\Sigma}| + \frac{N}{2}\log(2\pi). \tag{4}$$

3 Multi-Annotator Regression

Let $\mathcal{D}_m = \{(\boldsymbol{x}_i^m, y_i^m)\}_{i=1}^{N_m}$ be the data set of the mth annotator. We assume there are M annotators, each annotator m providing noisy labels that follow a Gaussian distribution $\mathcal{N}(0, \sigma_m^2)$ with unknown noise-level σ_m. Let $N = \sum_m N_m$ be the total number of annotations. Let $\boldsymbol{X}_m, \boldsymbol{Y}_m$ be the inputs and outputs occurring in $\mathcal{D}_m$. Define $\boldsymbol{X} = \cup_{m=1}^M \boldsymbol{X}_m$, $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_M\}$. Define I to be the number of (unique) inputs in $\boldsymbol{X}$. Furthermore, we define for $i = 1, \ldots, I$

$$\frac{1}{\hat{\sigma}_i^2} = \sum_{m \sim i} \frac{1}{\sigma_m^2}, \qquad \hat{y}_i = \hat{\sigma}_i^2 \sum_{m \sim i} \frac{y_i^m}{\sigma_m^2}, \qquad \hat{\boldsymbol{\Sigma}} = \text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_I^2), \tag{5}$$

with $m \sim i$ denoting the sum over annotators m that annotated sample $\boldsymbol{x}_i$.

We assume that annotators provide labels independently of each other. Therefore, the multi-annotator likelihood factorizes over cases in the training set and can, up to a constant, be rewritten in terms of the single-annotator model

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{f}) &= \prod_m \prod_{i \sim m} \mathcal{N}(y_i^m | f_i, \sigma_m^2) \propto \exp\left(-\frac{1}{2}\sum_i \sum_{m \sim i} \frac{(y_i^m - f_i)^2}{\sigma_m^2}\right) \\ &= \exp\left(-\frac{1}{2}\sum_i \frac{(\hat{y}_i - f_i)^2}{\hat{\sigma}_i^2} - \frac{1}{2}\sum_i \sum_{m \sim i} \frac{(y_i^m)^2}{\sigma_m^2} + \frac{1}{2}\sum_i \frac{\hat{y}_i^2}{\hat{\sigma}_i^2}\right) \\ &= \exp\left(-\frac{1}{2}\sum_i \frac{(\hat{y}_i - f_i)^2}{\hat{\sigma}_i^2} + c\right), \end{aligned} \tag{6}$$

with c independent of $\boldsymbol{f}$. The posterior process

$$p(\boldsymbol{f}|\boldsymbol{Y}) \propto p(\boldsymbol{f})p(\boldsymbol{Y}|\boldsymbol{f}) \propto \exp\left(-\frac{1}{2}\boldsymbol{f}^T\boldsymbol{K}^{-1}\boldsymbol{f} - \frac{1}{2}(\hat{\boldsymbol{Y}} - \boldsymbol{f})^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{Y}} - \boldsymbol{f})\right), \tag{7}$$

is thus Gaussian $\mathcal{N}(\boldsymbol{m}, \boldsymbol{V}) \propto \exp(-\frac{1}{2}\boldsymbol{f}^T\boldsymbol{V}^{-1}\boldsymbol{f} + \boldsymbol{f}^T\boldsymbol{V}^{-1}\boldsymbol{m})$ with mean and covariance given by

$$\begin{aligned} \boldsymbol{m} &= (\boldsymbol{K}^{-1} + \hat{\boldsymbol{\Sigma}}^{-1})^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{Y}}, \\ \boldsymbol{V} &= (\boldsymbol{K}^{-1} + \hat{\boldsymbol{\Sigma}}^{-1})^{-1}. \end{aligned} \tag{8}$$

From the posterior distribution we can derive the predictive equations for the multi-annotator model. The equations closely follow the predictive equations of the single-annotator model, but now with weighted output $\hat{\boldsymbol{Y}}$ and covariance $\hat{\boldsymbol{\Sigma}}$ which is no longer homogeneous as it depends on the data sample:

$$\begin{aligned} \overline{f}_{\mathcal{D}}(\boldsymbol{x}) &= k(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{K} + \hat{\boldsymbol{\Sigma}})^{-1}\hat{\boldsymbol{Y}}, \\ \mathrm{cov}_{\mathcal{D}}(f(\boldsymbol{x}), f(\boldsymbol{x}')) &= k(\boldsymbol{x}, \boldsymbol{x}') - k(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{K} + \hat{\boldsymbol{\Sigma}})^{-1}k(\boldsymbol{X}, \boldsymbol{x}'), \end{aligned} \tag{9}$$

where we used the fact that for a GP $\mathbb{E}[f_*|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{x}_*] = k(\boldsymbol{x}_*, \boldsymbol{X})\boldsymbol{K}^{-1}\mathbb{E}[\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{Y}]$, with $\mathbb{E}[\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{Y}]$ denoting the posterior mean of $\boldsymbol{f}$ given $\boldsymbol{X}$ and $\boldsymbol{Y}$ [10].

From Eq. (6) follows that the evidence for the multiple-annotator regression model with observations $\boldsymbol{Y}$ is very similar to that of a single-annotator model with observations $\hat{\boldsymbol{Y}}$. Some bookkeeping to account for the constant c gives

$$\begin{aligned} -\log p(\boldsymbol{Y}) = \frac{1}{2}\log|\boldsymbol{K} + \hat{\boldsymbol{\Sigma}}| + \frac{1}{2}\hat{\boldsymbol{Y}}^T(\boldsymbol{K} + \hat{\boldsymbol{\Sigma}})^{-1}\hat{\boldsymbol{Y}} + \frac{N}{2}\log(2\pi) + \\ - \frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}| - \sum_i \sum_{m\sim i} \log\frac{1}{\sigma_m} + \frac{1}{2}\sum_i \sum_{m\sim i} \frac{(y_i^m)^2}{\sigma_m^2} - \frac{1}{2}\sum_i \frac{\hat{y}_i^2}{\hat{\sigma}_i^2}. \end{aligned} \tag{10}$$

It can be checked that with one annotator the last four terms indeed cancel out because $\hat{\boldsymbol{\Sigma}} = \sigma^2\boldsymbol{I}$, $\hat{\boldsymbol{Y}} = \boldsymbol{Y}$, and $\hat{\sigma}_i^2 = \sigma^2$.

4 Experiments

We tested the GP multi-annotator model on the 'housing' benchmark dataset from the UCI machine learning repository. The target labels in the dataset were taken as ground truth from which we generated annotations for each annotator by adding Gaussian noise to the target label using their individual noise level. Inspired by [2, 9], we choose the following set-up. First, we randomly splitted the dataset into a training dataset (70%) and test dataset (30%), which were normalized using the training data. Second, we generated annotations for each annotator. We used three annotators with a variance of 0.25, 0.5, and 0.75. We did not use all the training data but only annotated a portion of it. For each annotator we selected A% of the training data at random for annotation, with $A \in \{10, 20, \ldots, 100\}$. Third, we report the root mean squared error ($RMSE(x, y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2}$) on the test data, averaged over 50 runs, for both the prediction of the targets/outputs and the hyperparameter prediction of the annotator noise-levels.

We show in Fig. 1 the RMSE results of six different models, the GP multi-annotator model, a GP fitted to the averaged training data provided by the annotators, three GP models fitted to each annotator individually, and a model that weighs the individual GP models. The latter model takes the mean prediction of each individual GP model and weighs it with the inverse predicted variance of that model. Note that the GP fitted to the average training data treats each annotator equally and does not learn individual noise levels. Each

model used a Gaussian kernel (cf. Eq. (1)), a zero mean function, and Gaussian likelihood function.

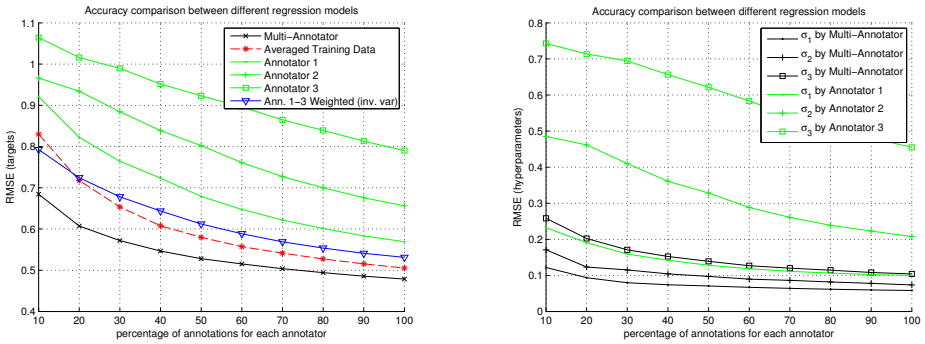

Fig. 1. The RMSE of the GP multi-annotator model, the GP fitted to the average response, and GPs fitted to each individual annotator on the 'housing' dataset (506 instances, 13 features). Left: RMSE for predicted targets. Right: RMSE for predicted noise-level hyperparameters.

In Fig 1, left panel, we plot the RMSE results with respect to the target labels. In Fig 1, right panel, we plot the RMSE results with respect to the noise-levels of the annotators used for data generation. Clearly, the GP multi-annotator model outperforms the other models, on both target prediction and hyperparameter estimation, but the precise amount depends on the dataset, the number of annotations, and the number of annotators and their noise levels.

The GP multi-annotator model improves upon earlier reported results [2, 9]; (1) It provides a non-parametric framework, (2) it is not necessary to annotate all samples by all annotators, and (3) the tuning of hyperparameters is fully automatic. In addition, since all data is properly combined using a Bayesian framework, ad hoc methods such as adding additional constraints to control annotator influence [11] or pruning low-quality annotators [12] are unnecessary.

5 Conclusions and Future Work

In this paper we presented a GP framework for regression with multiple noisy annotators. GPs provide a flexible, non-parametric framework, which naturally deals with missing annotations, and allows automatic tuning of hyperparameters using the evidence. The individual annotator noise levels can therefore be learned from data allowing for a better weighting of annotations leading to superior performance compared to a model fitted to an average response or a weighting of individually trained models.

For future work the GP multi-annotator model can be extended to be robust against outliers and relax the assumption of homogeneous Gaussian noise for

each annotator. Furthermore, the model can be extended to other supervised learning tasks like binary classification. GPs can be extended to binary classification by using the GP as a latent function whose sign determines the class label. Exact evaluations are, however, intractable, because of the non-Gaussian likelihood function and approximations are needed.

Acknowledgement This work has been carried out as part of the OCTOPUS project with Océ Technologies B.V. under the responsibility of the Embedded Systems Institute. This project is partially supported by the Netherlands Ministry of Economic Affairs under the Bsik program.

References

1. Raykar, V.C., Yu, S., Zhao, L.H., Jerebko, A., Florin, C.: Supervised learning from multiple experts: whom to trust when everyone lies a bit. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 889–896 (2009)
2. Raykar, V.C., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. JMLR, 11, 1297–1322 (2010)
3. Smyth, P., Fayyad, U., Burl, M., Perona, P., Baldi, P.: Inferring ground truth from subjective labelling of venus images. In: Advances in neural information processing systems 7, pp. 1085–1092 (1995)
4. Sheng, V.S., Provost, F., Ipeirotis, P.G.: Get another label? Improving data quality and data mining using multiple, noisy labelers. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 614–622 (2008)
5. Snow, R., O'Connor, B., Jrafsky, D., Ng, A.: Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 254–263 (2008)
6. Sorokin, A., Forsyth, D.: Utility data annotation with Amazon Mechanical Turk. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–8 (2008)
7. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE). An algorithm for the validation of image segmentation. IEEE Trans. Med. Imag. 23(7), 903–921 (2004)
8. Cholleti, S.R., Goldman, S.A., Blum, A., Politte, D.G., Don, S.: Veritas: combining expert opinions without labeled data. In: 20th IEEE International Conference on Tools with Artificial Inteligence (2008)
9. Ristovski, K., Das, D., Ouzienko, V., Guo, Y., Obradovic, Z.: Regression Learning with Multiple Noisy Oracles. In: 19th European Conference on Artificial Intelligence, pp. 445–450 (2010)
10. Rasmussen, C.E., Williams, C.K.I.: Gaussian processes for machine learning. MIT Press, Cambridge, MA (2006)
11. Dekel, O., Shamir, O.: Good Learners for Evil Teachers In: Proceedings of the 26th International Conference on Machine Learning, pp. 233–240 (2009)
12. Dekel, O., Shamir, O.: Vox populi: Collecting high-quality labels from a crowd. In: Proceedings of the 22nd Annual Conference on Learning Theory, pp. 377–386 (2009)