COMPARISON OF THE GAUSSIAN PROCESS CLASSIFICATION APPROACH WITH CLASSICAL PENALISED CLASSIFICATION METHODS IN HIGH DIMENSIONAL OMICS DATA

Miika Ahdesmäki[1,2] *and Korbinian Strimmer*[1]

[1] Institut für Medizinische Informatik, Statistik und Epidemiologie,
Härtelstr. 16-18, DE-04107, Germany,
[2]Department of Signal Processing, Tampere University of Technology,
P.O. Box 553, FI-33101 Tampere, Finland
miika.ahdesmaki@imise.uni-leipzig.de, strimmer@uni-leipzig.de

ABSTRACT

In this study we investigate (linear) Gaussian process (GP) priors for Bayesian classification comparing with other more classical penalised classification methods. For training the GP classifier we employ variational Bayesian estimation. In the comparison we include support vector machines and several diagonal and nondiagonal modifications of linear discriminant analysis (with and without penalisation). Relative performance is assessed using synthetic data and real data from gene expression and proteomic experiments. Based on the data analysis, We discuss the advantages and disadvantages of GP methods for high-dimensional classification.

1. INTRODUCTION

Gaussian processes refer to nonlinear nonparametric Bayesian regression (see [1, 2]), where the values of a function y are modelled directly without parameterising $y(x)$[1]. In the Gaussian process framework the prior distribution for the function values (the targets) is a Gaussian distribution with usually zero mean and a covariance function that relates the entries of the p-dimensional inputs (before observing any targets) by some given criteria. In many cases the covariance function is chosen so that spatial closeness implies higher covariance, i.e. the target values of measurements that are spatially close to each other should be correlated.

The prior distribution for the function values y gives rise to a space of functions and together with the measured targets and a likelihood function, where each target is seen as a Gaussian random variable, posterior predictive probabilities can be obtained.

Gaussian processes can also be seen as smoothers. The prior defines the smoothness of the function space and the role of the likelihood function is to give higher probabilities to functions that are close to the data.

For continuous target values, the likelihood function can also be defined as Gaussian, and thus the posterior can be analytically tractable.

For classification (binomial or multinomial), the likelihood is usually chosen as a sigmoid function (e.g. logit, probit). In classification, latent continuous variables are introduced and priors over them are given. These latent variables are then squashed through the chosen sigmoid to produce proper probabilities. For multinomial regression, variational Bayes can be applied to approximate the posterior as a Gaussian. Better results are obtained with the variational Bayes approximation than with Laplace approximation and the results are comparable to the ones obtained with MCMC [2].

In genomics and proteomics the dimension of the processed data is usually high, in the thousands, whereas the number of measurements is lower, usually in tens or hundreds. With so little data and so high dimensionality, the application of any complex classification rule will most likely yield worse generalisation performance than with a linear rule. This is mostly due to the increased number of parameters in the more complex models and the problem of estimating these parameters. See e.g. [3] for more discussion. Our experience with analysing high dimensional data has also shown that applying e.g. the radial basis covariance function kernel in the GP classifier yields often classifiers whose generalisation performance is random and worse than that of the linear (dot-product) kernel.

The rest of the paper is organised as follows. In the section METHODS we describe the basic Gaussian process classification framework and the variational Bayes approximation in classification (as in [2]). In RESULTS we show the essential results of the comparison study with other classifiers, some of which able to perform feature selection and also model shrinkage. We also make some concluding remarks in section CONCLUSIONS.

2. METHODS

We briefly desribe here the algorithm for the linear two-class variational Bayesian Gaussian process classification. The more general version of the variational Bayes algorithm can be found in [2]. We start by defining the model matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times P}$, the target vector $\boldsymbol{t} \in \{-1, 1\}^{N \times 1}$, the latent GP random variable $\boldsymbol{m} \in$

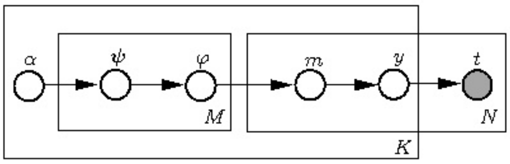

Figure 1. The hierarchical GP model from [2].

$\mathbb{R}^{N\times 1}$, the auxiliary latent variable $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$ (corresponding to the unknown ground truth, $\boldsymbol{y} = \boldsymbol{m} + \epsilon$, $\epsilon \sim \mathcal{N}(0, 1)$), the covariance hyperparameters $\boldsymbol{\varphi} \in \mathbb{R}^{M\times 1}$, where M is the number of hyperparameters in the chosen covariance function (here $M = N$), and the hyperhyperparameters $\boldsymbol{\psi}, \boldsymbol{\alpha}$. The model is illustrated in Fig. 1 where K is the number of classes for the more general case and can now be omitted.

The priors of the latent variables are set $\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\varphi} \sim GP(\boldsymbol{\varphi}) = \mathcal{N}(0, \boldsymbol{C}_{\boldsymbol{\varphi}})$, where $\boldsymbol{C}_{\boldsymbol{\varphi}} \in \mathbb{R}^{N\times N}$ gives the GP covariance function values between the measurements and $\boldsymbol{y}|\boldsymbol{m} \sim \mathcal{N}_{\boldsymbol{y}}(\boldsymbol{m}, \boldsymbol{I})$. For the covariance function hyperparameters an independent exponential distribution prior $\varphi_n \sim \text{Exp}(\psi_n)$ is set and a gamma prior is placed on the mean values of the exponential, i.e. $\psi_n \sim \Gamma(\alpha, \tau)$. The relationship between $\boldsymbol{y}$ and $\boldsymbol{t}$ is given by $t_n = 1$ if $y_n > 0$. The probit likelihood used here is of the form $P(\boldsymbol{t} = 1|\boldsymbol{m}) = \Phi(m)$.

The joint likelihood for the model, $p(\boldsymbol{t}, \boldsymbol{y}, \boldsymbol{m}, \boldsymbol{\varphi}, \boldsymbol{\psi}|\boldsymbol{X}, \alpha, \tau)$, is given in [2] (left out for clarity). Based on the joint likelihood and the variational Bayes approximation, the posterior means (denoted with tildes above the symbols) of the necessary parameters are given by the following iterations:

$$\tilde{\boldsymbol{x}} \leftarrow \boldsymbol{C}_{\tilde{\boldsymbol{\varphi}}} \left(\boldsymbol{I} + \boldsymbol{C}_{\tilde{\boldsymbol{\varphi}}}\right)^{-1} (\tilde{\boldsymbol{m}} + \boldsymbol{p}), \tag{1}$$

$$\tilde{\boldsymbol{\varphi}} \leftarrow \sum_s \tilde{\boldsymbol{\varphi}}^s w(\tilde{\boldsymbol{\varphi}}^s) \tag{2}$$

$$\tilde{\psi}_n \leftarrow \frac{\alpha + 1}{\tau + \tilde{\psi}_n} \tag{3}$$

where elements of $\boldsymbol{p}$ are given by

$$p_n = t_n \mathcal{N}_{\tilde{m}_n}(0, 1)/\Phi(t_n \tilde{m}_n).$$

$\boldsymbol{\varphi}^s$ are drawn using an importance sampler, i.e. by drawing S samples such that $\varphi_n^s \sim \text{Exp}(\tilde{\psi}_n)$ and evaluating

$$w(\tilde{\boldsymbol{\varphi}}^s) = \frac{\mathcal{N}_{\tilde{\boldsymbol{m}}}(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{\varphi}^s})}{\sum_{s'=1}^{S} \mathcal{N}_{\tilde{\boldsymbol{m}}}(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{\varphi}^{s'}}).} \tag{4}$$

For more details see [2].

The posterior probability of a new sample belonging to class 1 is given by evaluating

$$\tilde{m}^{new} = \tilde{\boldsymbol{y}}^\top \left(\boldsymbol{I} + \boldsymbol{C}_{\tilde{\boldsymbol{\varphi}}}\right)^{-1} \boldsymbol{C}_{\tilde{\boldsymbol{\varphi}}}^{new}, \tag{5}$$

$$\tilde{\nu}^{new} = \sqrt{1 + c_{\tilde{\boldsymbol{\varphi}}}^{new} - (\boldsymbol{C}_{\tilde{\boldsymbol{\varphi}}}^{new})^\top \left(\boldsymbol{I} + \boldsymbol{C}_{\tilde{\boldsymbol{\varphi}}}\right)^{-1} \boldsymbol{C}_{\tilde{\boldsymbol{\varphi}}}^{new}}, \tag{6}$$

$$P(t_{new} = 1|\boldsymbol{x}_{new}, \boldsymbol{X}, \boldsymbol{t}) = \Phi(\frac{\tilde{m}^{new}}{\tilde{\nu}^{new}}), \tag{7}$$

where $\boldsymbol{C}_{\tilde{\boldsymbol{\varphi}}}^{new}$ contains the covariance function values between the new point and those in $\boldsymbol{X}$, and $c_{\tilde{\boldsymbol{\varphi}}}^{new}$ is the self-vs-self covariance function value for the test data point.

To evaluate the GP covariances, the input entries in $\boldsymbol{X}$ are first transformed by subtracting the column means (note that this is not the same as in evaluating the ordinary $p \times p$ covariance estimate) and then to allow for a class boundary that does not go through the origin, a column of ones is added to the design. Dot-products are then evaluated between all the row vectors of the transformed $\boldsymbol{X}$, so that entries in $\boldsymbol{C}_{\boldsymbol{\varphi}}$ are given by $\boldsymbol{C}_{ij} = \boldsymbol{x}_i^\top \text{diag}(\boldsymbol{\varphi})\boldsymbol{x}_j$, and the $N \times N$ covariance matrix is then scaled into the corresponding correlation matrix. Note that the covariance / correlation matrix used in the GP framework gives a measure of relatedness between the measurements, not the variables. Thus we are directly modelling the measurements without parameterising the model.

We initiate the algorithm by drawing $\tilde{\boldsymbol{m}}$ from a normal distribution, set $\tilde{\boldsymbol{\varphi}} = \boldsymbol{1}$ and assume vague priors $\alpha = \tau = 10^{-3}$ (notice the effect of τ on $\tilde{\psi}_n$ in the importance sampler when $\tilde{\psi}_n$ vanishes). We then perform the updates in Eqs. 1 to 3 for a maximum of 20 times (a less ad-hoc stopping criterion will be left for future work) and for each round draw 300 samples for the importance sampler to train the classifier.

Having obtained the posterior means for $\boldsymbol{m}$ and $\boldsymbol{\varphi}$ we can then estimate the posterior probabilities for new samples as given in Eq. 7

Most of the time in the algorithm will be spent in estimating the hyperparameters in the importance sampler, i.e. in the feature selection phase, involving an $O(N^3)$ step (Eq. 4).

3. RESULTS

To assess the relative performance of different pattern classification methods, we applied several classifiers in a simulation study on synthetic data and also with real data. When applying to real data, cross-validation (CV) was performed to estimate classification performance.

The competing classifiers chosen for the simulation study were the shrunken centroids diagonal discriminant method dubbed *PAM* [4], shrinkage linear discriminant analysis (SLDA), shrinkage diagonal discriminant analysis (SDDA) [5], support vector machine (SVM) [6] (without tuning and with the default values for the kernel) and the partial least squares classifier by [7] and the standard linear discriminant analysis (LDA) classifier. Since the sample covariance estimate is singular in the $p > n$-setting, the pseudo inverse was used in the standard LDA.

For the analysis of real data we chose the winner of the simulation study, the SLDA, to compare with the GP classifier.

3.1. Synthetic data

For the synthetic data we generated training samples ($N_1 = 25$, $N_2 = 25$) for classes 1 and 2 from two multivari-

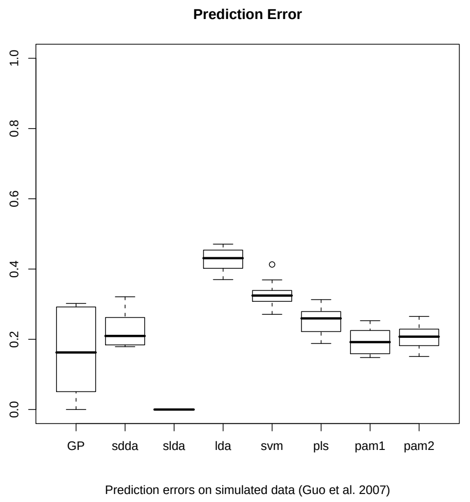

Figure 2. Simulation results: prediction error rates as box-plots for the chosen classifiers.

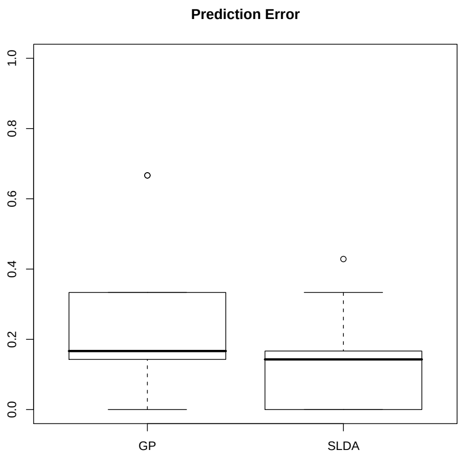

Figure 3. Results on the colon data: prediction error rates as box-plots for the GP and SLDA classifiers.

ate Gaussian distributions ($p = 1000$), specified by μ_1 and μ_2 and a common covariance matrix Σ. For the covariance matrix we chose a block diagonal covariance matrix from [8], with the blocks representing autoregressive models (correlation 0.99^l with l the distance from the diagonal). The block size was set to 25×25 and the first 50 entries of μ_1 were set to 1, otherwise μ_1 and μ_2 were set to zero. The performance was then evaluated on 1000 test samples generated from the corresponding multivariate normals (500 from each). The procedure was repeated 10 times and the resulting error rates are plotted in Fig. 2.

The results show that the shrinkage LDA captures the covariance structure of the problem and classifies the samples without error. The GP and the diagonal discrimant analysis based methods (sdda and pam) show an approximately 20% error rate, with the median of the GP being slighly lower but with a higher variability. This variability of the GP is probably a result of the unoptimal number of rounds for the variational Bayes iterations and the importance sampler.

3.2. Real genomic data

We applied the GP and SLDA classification methods on a colon cancer microarray data set ([9]) to compare the classification performance. 10-fold (stratified) CV, repeated 2 times, was used to estimate the classification error rates. The results are plotted in Fig. 3.

The median values of the classification errors are close to each other for the classifiers, but the distribution is clearly lower for the SLDA. To better capture the properties in the data and to enhance classification performance, several covariance functions could be applied in parallel in the GP approach.

4. CONCLUSIONS

In this paper we considered the two-class classification problem and employed several classification methods in a comparison study. The results show that the linear kernel GP performs similar to the regularised diagonal discriminant analysis based competitors. The GP implements feature selection naturally and in a proper probabilistic manner. Operating in the Bayesian framework means also that the amount of ad-hoc tuning is kept to the minimum and overfitting is avoided naturally. The computational time of the GP approach is a limiting factor if the hyperparameters must be estimated, but only dependent on N, not P. Future work includes e.g. optimising the number of needed iteration rounds for the importance sampler and combining the linear covariance kernel with nonlinear kernels (such as the radial basis function).

5. ACKNOWLEDGMENTS

M.A. is a postdoctoral research fellow of the Alexander von Humboldt Foundation and would like to thank the foundation for its support.

6. REFERENCES

[1] D. J. C. Mackay, *Neural Networks and Machine Learning (edited by C. M. Bishop)*, chapter Introduction to Gaussian Processes, pp. 133–166, NATO ASI Series, 1998.

[2] M. Girolami and S. Rogers, “Variational bayesian multinomial probit regression with gaussian process priors,” *Neural Computation*, vol. 18, pp. 1790–1817, 2006.

[3] D. Hand, “Classifier technology and the illusion of progress,” *Statistical Science*, vol. 21, pp. 1–14, 2006.

[4] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, “Diagnosis of multiple cancer types by shrunken centroids of gene expression,” *Proceedings of the National Academy of Sciences of the USA*, vol. 99, pp. 6567–6572, 2002.

[5] M. Ahdesmäki and K. Strimmer, "Feature selection in "omics" prediction problems using cat scores and false non-discovery rate control," *arXiv*, 2009.

[6] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," Tech. Rep., Department of Computer Science, National Taiwan University, 2009.

[7] A. L. Boulesteix and K. Strimmer, "Partial least squares: a versatile tool for the analysis of high-dimensional genomic data," *Briefings in Bioinformatics*, vol. 7, pp. 32–44, 2007.

[8] Y. Guo, T. Hastie, and R. Tibshirani, "Regularized linear discriminant analysis and its application in microarrays," *Biostatistics*, vol. 8, pp. 86–100, 2007.

[9] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Cell Biology*, vol. 96(12), pp. 6745–6750, 1999.