A BAYESIAN INFERENCE FOR A MANIFOLD GAUSSIAN PROCESS CLASSIFIER: APPLICATIONS TO IMAGES CLASSIFICATION

Anis Fradi [1,3,4] & Chafik Samir [1] & Anne-Françoise Yao [2]

[1] *CNRS-LIMOS (UMR 6158),UCA, France.*
[2] *CNRS-LMBP (UMR 6620), UCA, France.*
[3] *MAPSFA (LR11ES35),* [4] *Faculty of Sciences of Monastir, University of Monastir, Tunisia.*
anis.fradi@etu.uca.fr & {chafik.samir & Anne.yao}@uca.fr

Résumé. Ce travail est consacré au développement des modèles d'inférence bayésienne lorsque les observations appartiennent à un espace de grande dimension ou de dimension infinie. En général, le problème de la grande dimension a une influence majeure sur les performances des modèles. Pour y remédier, nous considérons un processus gaussien de classification sur une variété (manifold). Cette formulation présente plusieurs avantages : l'apprentissage du processus dans un espace fonctionnel plus général et l'inclusion de certaines contraintes telles que la non-linéarité, la séparabilité, la réduction de dimension, etc. Nous illustrons l'efficacité et la précision de notre modèle proposé pour classer des images comme observations de grande dimension.

Mots-clés. Processus gaussien, inférence en grande dimension, apprentissage automatique, apprentissage sur une variété, classification d'images, données de grande dimension.

Abstract. One of the challenging regression problems consists of learning relevant and meaningful relationships between high dimensional representations across a relatively few observed individuals. Since this problem could have drastic effects on the classification performance, we propose a Bayesian approach with a Gaussian process classifier. It is commonly known that methods based on Gaussian process classifiers are mainly effective in the case of low and medium dimensions. On a mathematical basis, given a finite set of observations $\mathbf{X} = \{x_i\}_{i=1}^{N}$ and a covariance function $c(.,.)$, the most prominent weakness of the standard Gaussian process inference is that it suffers from time complexity due to the inversion and the determinant of the $N \times N$ covariance matrix $\mathbf{C}$. We propose to define and use a manifold Gaussian process classifier's formulation with the advantage of learning a classifier in the feature space under some constraints: Nonlinearity, separability, dimensionality reduction, etc. We illustrate the efficiency and the accuracy of our framework for classifying images of high-dimensional inputs.

Keywords. Gaussian process classifier, statistical machine learning, manifold learning, images classification, inference on high-dimensional data.

1 Introduction

There is currently a significant interest in statistical modeling and machine learning techniques with the challenge of processing massive amounts of complex data (in the form of text documents, images, audio, video, etc.). While significant recent progress has been made in the field of image classification, the problem of high dimensional data remains particularly challenging. This occurs when the number of covariates is relatively large or when the components are highly correlated. The number of equations is consequently less than the number of unknown parameters, which could lead to an infinite number of

solutions.
This article focuses on Gaussian process classifier (GPc) [1], a non-parametric statistical model, which can be used in various scenarios. Each GPc is specified by a mean function, often taken as zero and a covariance function that controls its smoothness. To handle the problem of high dimensional data, we propose a novel framework, called manifold Gaussian process classifier (MGPc), which jointly learns the data and the GPc on a feature (manifold) space. This formulation has the benefit to make it more easy to deal with nonlinearity of data and to create separability in the feature space [2]. In this context, the posterior distribution remains intractable since the likelihood function deviates from mathematically convenient forms. To make the posterior proportionality tractable, the approximation of non-Gaussian posterior distribution with a Gaussian one is explored. The proposed solutions are called manifold Laplace approximation (MLA) and manifold expectation propagation (MEP). For the prediction part, the Bayesian inference has proved its efficiency for optimizing the models' parameters compared to deterministic methods.

2 Background

This section presents tools and notations needed to develop our approach.

2.1 Notations

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and the output space, respectively. We assume that we have N pairs of independent observations: $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \ldots, N$ distributed with the same law as (X, y) where y_i is the i-th observation of the response variable and x_i is its associated covariate vector. In this work we focus on the case where only two classes are discriminated, i.e. each output $y_i \in \mathcal{Y} = \{0, 1\}$ for the associated input $x_i \in \mathcal{X} \subseteq \mathbb{R}^p$, $i = 1, \ldots, N$. For more simplicity, we note the finite set of observations by $\mathbf{X} = \{x_i\}_{i=1}^N$ and $\mathbf{y} = \{y_i\}_{i=1}^N$.

2.2 Weighted ridge logistic

We consider the problem of learning a probabilistic regression logistic model from the available observations $\{\mathbf{X}, \mathbf{y}\}$ by fitting a vector of parameters β in order to better explain the relationship between X and y. The quantity of interest is then the probability of $y = 1$ given $X = x$ denoted by: $\pi_\beta(x) = \mathbb{P}(y = 1|X = x) = \sigma(x^T\beta)$ where σ refers to the sigmoid function. Basically inspired from ridge logistic [3], the idea of weighted ridge logistic consists of considering the weighted sum between the unstructured log-likelihood of logistic regression model: $l(\beta) = \sum_{i=1}^N y_i {x_i}^T\beta - \log(1+\exp({x_i}^T\beta))$ and the l^2 -norm of unknown parameters β, giving the regularized likelihood: $l^\lambda(\beta) = \frac{(1-\lambda)}{2} l(\beta) - \frac{\lambda}{2}||\beta||_2^2$ for $0 \leq \lambda \leq 1$. Therefore, for a optimal choice of λ, the estimator $\beta^{\lambda,*}$ should optimize the log-likelihood compared to the unstructured maximum likelihood estimator (MLE): $\beta^{0,*}$ (obtained for $\lambda = 0$). To get $\beta^{\lambda,*}$, We develop a Taylor expansion of $\nabla l^\lambda(\beta^{k+1})$ at β^k and use the iterative Newton-Raphson (or gradient-descent alternatively) approach iteratively. For the rest, we assume that $\mathcal{Y} = \{-1, +1\}$.

3 The proposed Bayesian approach

In this section, we introduce our proposed MGPc, a nonparametric model effective especially when the size of parameters vector is larger than the number of observations. Another reason for this choice is that Bayesian inference gives more chance for convergence when adding a prior law to likelihood term.

3.1 Problem formulation

We first introduce a new latent function $f : \mathcal{X} \to \mathbb{R}$. The GPc is based on placing a Gaussian process prior over the latent function: $f \sim \mathcal{GP}(0, c)$ where c is a covariance function. We consider a formulation of the probability of class "+1" as: $\pi_f(x) = \mathbb{P}(y = +1|f(x))$ where this term usually refers to the sigmoid $\sigma(f(x))$ or the probit $\Phi(f(x))$. By abuse of notation, we note: $\mathbf{f} = (f_1, \ldots, f_N) = f(\mathbf{X})$. The likelihood function is then the product of individual likelihoods: $\mathbb{P}(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{N} \mathbb{P}(y_i|f_i)$ as well as the posterior distribution is proportional to $\mathbb{P}(\mathbf{f}|\mathbf{X}, \mathbf{y}) \propto \mathbb{P}(\mathbf{f}|\mathbf{X}) \times \mathbb{P}(\mathbf{y}|\mathbf{f})$. Now, we present the MGPc and we make connection to the standard GPc.

3.2 Manifold Gaussian process classifier

Let K be a positive real valued kernel defined on $\mathcal{X} \times \mathcal{X}$ with its corresponding reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$. We assume that we have a partition of $\mathcal{X}$ with centers $\{c_j\}_{j=1}^m$ and an empirical risk function $E : \mathcal{X}^m \to \mathbb{R}$ including data and regularization terms. Then, from the "representer theorem" [4] yields: for $\hat{\Psi} \in \mathcal{H}_K$ satisfying $\hat{\Psi} = \operatorname{argmin}_{\Psi \in \mathcal{H}_K} E(\Psi)$, $\hat{\Psi}$ admits this representation

$$\hat{\Psi}(.) = \sum_{j=1}^{m} \alpha_j K(., c_j) \tag{1}$$

In this work, $\{c_j\}_{j=1}^m$ are supposed to be the centers obtained by the clustering method applied to $\mathbf{X}$ with $m << p$. Consider the reproducing kernel feature map

$$\begin{array}{rccc} \phi_j & : & \mathcal{X} & \to & \mathbb{R} \\ & & x & \mapsto & K(x, c_j) \end{array} \tag{2}$$

As a consequence, $\hat{\Psi}(x) = \sum_{j=1}^{m} \alpha_j \phi_j(x) \in \mathcal{F} \subseteq \mathbb{R}^m$. Therefore, we define a MGPc $M : \mathcal{F} \subseteq \mathbb{R}^m \to \mathbb{R}$ as a GPc $f = M \circ \hat{\Psi} : \mathcal{X} \subseteq \mathbb{R}^p \to \mathbb{R}$ depending on a modified covariance function $\tilde{c}(x, x') = c(\hat{\Psi}(x), \hat{\Psi}(x'))$, which operates on the feature space $\mathcal{F}$. For the following, we recover $\mathbf{Z} = (z_1, \ldots, z_N) = \hat{\Psi}(\mathbf{X})$, $\mathbf{M} = (M_1, \ldots, M_N) = M(\hat{\Psi}(\mathbf{X}))$ and the transformed inputs $\mathbf{Z} = \{z_i\}_{i=1}^N$. The posterior proportionality becomes $\mathbb{P}(\mathbf{M}|\mathbf{Z}, \mathbf{y}) \propto \mathbb{P}(\mathbf{M}|\mathbf{Z}) \times \mathbb{P}(\mathbf{y}|\mathbf{M})$. The question that arises now is how to find an explicit form to the last posterior proportionality distribution?

3.3 Manifold Laplace approximation

By Bayes rule, the log-posterior distribution is proportional to $g(\mathbf{M}) = \log \mathbb{P}(\mathbf{y}|\mathbf{M}) - \frac{1}{2}\mathbf{M}^T \bar{\mathbf{C}}^{-1} \mathbf{M}$ where $\bar{\mathbf{C}} = \tilde{c}(\mathbf{X}, \mathbf{X})$ is the $N \times N$ covariance matrix. Note that $g(.)$ is a concave function leading to a unique global maximum. For MLA, we approximate the maximum a posteriori (MAP) denoted $\hat{\mathbf{M}}$ and

obtained by maximizing the log-posterior distribution, i.e. $\hat{\mathbf{M}} = \arg\max_{\mathbf{M}} \mathbb{P}(\mathbf{M}|\mathbf{Z}, \mathbf{y})$. From a second order Taylor series centered in $\hat{\mathbf{M}}$, we obtain a Gaussian approximation as

$$\hat{\mathbb{P}}(\mathbf{M}|\mathbf{Z}, \mathbf{y}) = \mathcal{N}(\mathbf{M}|\hat{\mathbf{M}}, \mathbf{K}^{-1}) \propto \exp(-\frac{1}{2}(\mathbf{M} - \hat{\mathbf{M}})^T \mathbf{K} (\mathbf{M} - \hat{\mathbf{M}})) \tag{3}$$

where $\mathbf{K} = -\nabla^2 \log \mathbb{P}(\mathbf{M}|\mathbf{Z}, \mathbf{y})_{|\mathbf{M}=\hat{\mathbf{M}}} = \bar{\mathbf{W}} + \bar{\mathbf{C}}^{-1}$ and $\bar{\mathbf{W}}$ is a $N \times N$ diagonal matrix with $\bar{\mathbf{W}}_{ii} = -\frac{\partial^2 \log \sigma(y_i M_i)}{\partial^2 M_i} = \frac{\exp(M_i)}{(1+\exp(M_i))^2}$. We use the Newton-based method to find the MAP estimator $\hat{\mathbf{M}}$ iteratively as: $\mathbf{M}^{t+1} = (\bar{\mathbf{C}}^{-1} + \bar{\mathbf{W}})^{-1}(\bar{\mathbf{W}}\mathbf{M}^t + \nabla \mathbb{P}(\mathbf{y}|\mathbf{M}^t))$. The predictive distribution for the MLA at a test input $z^* = \hat{\Psi}(x^*)$ is

$$\hat{\mathbb{P}}(M(z^*)|\mathbf{Z}, \mathbf{y}, z^*) = \mathcal{N}(M(z^*)|\mu(z^*), \sigma^2(z^*)) \tag{4}$$

$$\mu(z^*) = \bar{\mathbf{C}}_*^T \bar{\mathbf{C}}^{-1} \hat{\mathbf{M}} \tag{5}$$

$$\sigma^2(z^*) = \bar{\mathbf{C}}_{**} - \bar{\mathbf{C}}_*^T (\bar{\mathbf{C}} + \bar{\mathbf{W}}^{-1})^{-1} \bar{\mathbf{C}}_* \tag{6}$$

where $\bar{\mathbf{C}}_{**} = \tilde{c}(x^*, x^*)$ and $\bar{\mathbf{C}}_* = \tilde{c}(\mathbf{X}, x^*)$. Given the mean $\mu(z^*)$ and the variance $\sigma^2(z^*)$, we approximate the predictor for $y^* = +1$ as: $\bar{\pi}(z^*) \approx \int_{\mathbb{R}} \sigma(M^*) \hat{\mathbb{P}}(M^*|\mathbf{Z}, \mathbf{y}, z^*) \mathrm{d}M^*$.

3.4 Manifold expectation propagation

The key idea of MEP is to replace the likelihood terms: $\mathbb{P}(y_i|M_i) = \Phi(y_i M_i)$ by un-normalized Gaussian distributions: $\tilde{L}_i \times \mathcal{N}(M_i|\tilde{\mu}_i, \tilde{\sigma}_i^2)$ since $\mathbb{P}(y_i|M_i)$ is not a distribution on M_i. Based on local approximations, the posterior can be approximated by

$$\hat{\mathbb{P}}(\mathbf{M}|\mathbf{Z}, \mathbf{y}) \propto \mathbb{P}(\mathbf{M}|\mathbf{Z}) \times \prod_{i=1}^{N} \tilde{L}_i \times \mathcal{N}(M_i|\tilde{\mu}_i, \tilde{\sigma}_i^2) = \mathcal{N}(\mathbf{M}|\mu, \Sigma) \tag{7}$$

where $\tilde{\mu} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_N)$, $\tilde{\Sigma} = \mathrm{diag}(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_N^2)$, $\mu = \Sigma \tilde{\Sigma}^{-1} \tilde{\mu}$, and $\Sigma = (\bar{\mathbf{C}}^{-1} + \tilde{\Sigma}^{-1})^{-1}$. The moments of prediction are: $\mu(z^*) = \bar{\mathbf{C}}_*^T (\bar{\mathbf{C}} + \tilde{\Sigma})^{-1} \tilde{\mu}$ and $\sigma^2(z^*) = \bar{\mathbf{C}}_{**} - \bar{\mathbf{C}}_*^T (\bar{\mathbf{C}} + \tilde{\Sigma})^{-1} \bar{\mathbf{C}}_*$. Therefore, the approximate predictor for $y^* = +1$ is

$$\bar{\pi}(z^*) \approx \int_{\mathbb{R}} \Phi(M^*) \hat{\mathbb{P}}(M^*|\mathbf{Z}, \mathbf{y}, z^*) \mathrm{d}M^* = \Phi(\frac{\bar{\mathbf{C}}_*^T (\bar{\mathbf{C}} + \tilde{\Sigma})^{-1} \tilde{\mu}}{\sqrt{1 + \bar{\mathbf{C}}_{**} - \bar{\mathbf{C}}_*^T (\bar{\mathbf{C}} + \tilde{\Sigma})^{-1} \bar{\mathbf{C}}_*}}) \tag{8}$$

4 Application

We evaluate the performance and the efficiency of the proposed methods on a data base of 2042 images of manufacturing defects (" $+1$" defective and " -1" non-defective). We learn the model parameters from a training set and use the rest for evaluation. The subdivision has been performed randomly 15 times and the accuracy rates are given as a mean. We consider the False Positives (FP: non-defective but classified as defective), False Negatives (FN: defective but classified as non-defective) and Classification Errors (CE). We also use the ROC curve to show the sensitivity and the specificity jointly. We compare the MGPc against the regularization approach using two iterative algorithms: gradient and Newton.

Results of logistic regression. The error rates of usual logistic regression are summarized in Table 1. Accordingly, one can observe that gradient field as a representation of an image achieves the lowest error with a significant margin. Consequently, we keep this feature for the rest of tests.
Results of weighted ridge logistic. From Figure 1 (left and middle) the Newton-based method outperforms gradient. This result is confirmed by the ROC curve of Newton-MLE weighted ridge in Figure 1 (right).

Error rates	gradient	Gabor	binarization
FN	20%	51%	53%
FP	27%	47%	43%
CE	25%	48%	45%

Table 1: Classification accuracy using Newton-MLE

Error rates	MLA	MEP
FN	11%	8.5%
FP	19%	10%
CE	17%	9.5%

Table 2: Classification accuracy using MGPc

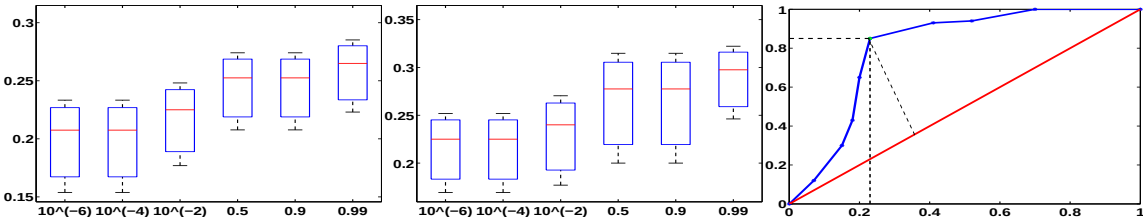

Fig. 1: Errors as a function of regularization parameters (FP=upper values, FN=lower values and CE=values in the middle) obtained by: Newton method (left) and gradient (middle). The ROC curve of Newton is given on the right.

Results of MGPc. Table 2 summarizes results of the MGPc's model. We can observe that both MLA and MEP improve the specificity and sensibility with better results for MEP. The ROC curves of Figure 2 confirm that MEP has the most predictive power and generalization capability where FP is 8.5% and FN is 10%.
For illustration, Figure 3 shows an example of the key steps to classify a new transformed input $z^* = \hat{\Psi}(x^*)$. We remark that, in this example with a test input $(y^* = -1)$, MEP has a better prediction accuracy where $\bar{\pi}(z^*) = 0.33$ for MLA and $\bar{\pi}(z^*) = 0.41$ for MEP.

5 Conclusion

In this work, we have reformulated a real classification problem as a weighted ridge logistic with the advantage of reducing the research space of optimal regularization parameters, eventually when the feature vectors are high dimensional and the number of samples is relatively small. We have also proposed an efficient solution providing details of two Manifold-based inferences to build supervised Gaussian process classifiers. Experiments have been conducted to classify images and have shown that proposed methods were successful.

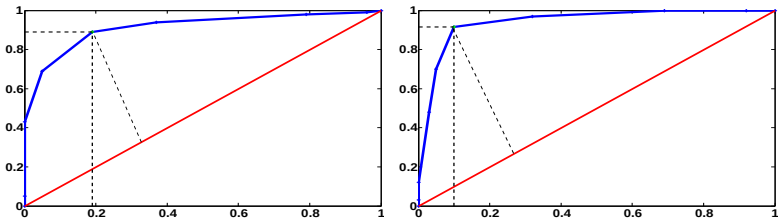

Fig. 2: ROC curves for MLA (left) and MEP (right).

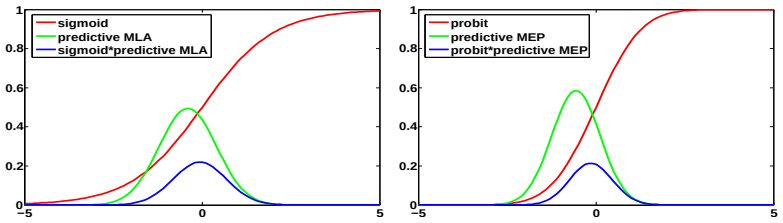

Fig. 3: An illustration of key steps to classify a test input $(y^* = -1)$ using MLA (left) and MEP (right).

References

[1] C. E. Rasmussen, and C. K. I. Williams. (2006), Gaussian Pocesses for Machine Learning, *Adaptive computation and machine learning*.

[2] R. Calandra, J. Peters, C. E. Rasmussen, and M. P. Deisenroth. (2016), Manifold gaussian processes for regression, in *2016 International Joint Conference on Neural Networks, IJCNN 2016, Vancouver, BC, Canada*, July, pp. 24-29.

[3] J. M. Pereira, M. Basto, and A. F. da Silva. (2016), The logistic lasso and ridge regression in predicting corporate failure, *Procedia Economics and Finance*, 39, pp. 634 - 641.

[4] B. Schölkopf, R. Herbrich, and A. J. Smola. (2001), A generalized representer theorem, in Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory, *Springer-Verlag*.