MANIFOLD-BASED INFERENCE FOR A SUPERVISED GAUSSIAN PROCESS CLASSIFIER

Anis Fradi [1], *Chafik Samir* [1], *and Anne-Françoise Yao* [2]

[1] Laboratoire d'Informatique de Modélisation et d'Optimisation des Systèmes, CNRS (UMR 6158)
[2] Laboratoire de Mathématiques Blaise Pascal, CNRS (UMR 6620)
[1,2] University of Clermont Auvergne, France

ABSTRACT

One of the challenging classification 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 alternative in the case of logistic regression. The proposed method has the additional benefit to learn both the adaptive embedding, as a Gaussian process, and the dimensionality reduction, jointly within the same Bayesian framework. We illustrate the efficiency and the accuracy of our framework for classifying images of manufacturing defects.

Index Terms— Machine Learning, Gaussian Process, Manifold Embedding, Regression, Image Classification.

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.). In particular, those methods become interesting in many applications such as image classification, object recognition or detection [1]. While significant recent progress has been made in the field of image classification, the problem of high dimensional data remains particularly challenging [2]. This occurs when the number of covariates is relatively large or when the components are highly correlated [3]. The number of equations is consequently less than the number of unknown parameters, which could lead to an infinite number of solutions.

To avoid the effects of high dimensional data on the classification performance, intensive research has been conducted [4,6], for which the goal is to build effective predictive models despite the aforementioned problems. In general, the previous methods in this context can be divided into two main categories. The first category known as shrinkage or regularization methods. This technique is typically based on constraining or regularizing the coefficient estimates, or equivalently, on shrinking the coefficient estimates towards zero. The shrinking of the coefficient estimates has a significant impact of reducing their variance [5].

The ridge regression and the Lasso are widely employed methods in this context. Ridge regression [7] is an example of shrinkage method applied to maximum log-likelihood (MLE) (or equivalently ordinary least squares (OLS)) estimator. The Lasso [8, 9] is another well known shrinkage method which replaces the l^2 norm on ridge by l^1 made use of Bayesian networks. In particular, some approaches are based on the approximation of non-Gaussian joint posterior with a Gaussian one. For instance, [10, 11] uses a Laplace approximation whereas [11,12] uses an expectation propagation (EP) assuming a density filtering and an extended version of the Kalman filter. Similarly, [13] employs an adaptive cavity approximation to capture significant correlations.

In the same context, to handle the problem of high dimensional data, we propose a new method, called manifold Gaussian processes (MGP) classifier, which jointly combines learning the data mapping into a feature space and a Gaussian processes classifier [14] on the embedding manifold (Hilbert space). This formulation has the benefit to make it more easy to deal with nonlinearity of data and to create separability in the embedding space [15]. For the prediction part, the Bayesian inference has proved its efficiency for optimizing the model parameters.

The rest of the paper is organized as follows. We first introduce the background and the MLE-ridge weighted regression for estimating parameters in Section 2. We give details of the proposed framework: MGP classifier in Section 3 and manifold Gaussian processes (MEP) in Section 4. The experimental results are presented and discussed in Section 5. We finally conclude in Section 6.

2. NOTATIONS AND BACKGROUND

This preliminary section introduces notations that will be used throughout the paper and provides a brief set of basic principles and background. Though mildly technical, it is useful as we focus on a particular sub-set of results that pertain directly in building our model. In this work, we suppose that we observe N independent individuals (X_1, Y_1),..,(X_N, Y_N) distributed with the same law as (X, Y) and we consider the problem of learning a probabilistic regression model from available observations, to better explain the relationship be-

tween the response variable Y and X. We are particularly interested on studying the case of binary classification, where only two classes are discriminated, i.e $Y \in \{0, 1\}$. We remind that logistic regressions have been widely used as $\pi_\beta(x) = \mathbb{P}(Y = 1|X = x) = \sigma(x^T\beta)$ where σ usually refers to the sigmoid function. We denote by $\pi_\beta(x_i)$ the probability of $Y_i = 1$ for a given $X_i = x_i$ for all $i \in \{1, .., N\}$. Basically inspired from ridge logistic [16], the idea of weighted ridge logistic consists of considering the weighted sum between the 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 square sum of β, which gives

$$l^\lambda(\beta) = \frac{(1-\lambda)}{2} l(\beta) - \frac{\lambda}{2}||\beta||_2^2 \tag{1}$$

where the regularization parameter satisfies $0 < \lambda < 1$. We denote by $\beta^{\lambda,*}$ the optimal solution. Therefore for a better choice of λ, the estimator $\beta^{\lambda,*}$ should maximize the log-likelihood compared to the unstructured MLE: $\beta^{0,*}$, ie. MSE($\beta^{\lambda,*}$) < MSE($\beta^{0,*}$) [17]. The gradient vector of $l^\lambda(\beta)$ is

$$\nabla l^\lambda(\beta) = \frac{(1-\lambda)}{2} \nabla l(\beta) - \lambda\beta \tag{2}$$

where $\nabla l(\beta) = \mathbb{X}^T(\mathbb{Y} - \pi_\beta(\mathbb{X}))$, $\mathbb{X} = (x_1, \ldots, x_N)$, and $\mathbb{Y} = (y_1, \ldots, y_N)$. Then the estimator is a solution of $\nabla l^\lambda(\beta) = 0$ and the negative Hessian of $l^\lambda(\beta)$ is given by

$$H^\lambda(\beta) = \frac{(1-\lambda)}{2} \mathbb{X}^T H(\beta)\mathbb{X} + \lambda I \tag{3}$$

where $H(\beta)$ is an $N \times N$ diagonal matrix with $H_{ii}(\beta) = \pi_\beta(x_i)(1 - \pi_\beta(x_i))$. We develop a Taylor expansion of $\nabla l^\lambda(\beta^{k+1})$ at β^k and use iterative Newton or gradient descent approaches iteratively until convergence, which gives

$$\beta^{k+1} \approx \frac{(1-\lambda)}{2} (H^\lambda(\beta^k))^{-1}(\mathbb{X}^T H(\beta^k)\mathbb{X}\beta^k + \nabla l(\beta^k)) \tag{4}$$

For the rest of this paper, we assume that $Y \in \{-1, 1\}$.

3. MANIFOLD GAUSSIAN PROCESS: A BAYESIAN INFERENCE WITH LAPLACE APPROXIMATION

The manifold embedding consists of searching for an optimal mapping from the data space to a new manifold under some constraints (e.g. reduce non-linearity, increase separability.) [18]. If we denote this mapping Ψ and $(\mathcal{H}, <.,.>_\mathcal{H})$ the new feature space, the learning will be made on $(\Psi(X), Y)$ where it is desired that the data will be more linearly separable. Furthermore, we assume that $\mathcal{H}$ is finite dimensional and that a basis of $\mathcal{H}$ could be determined in unsupervised manner. In the remainder of this paper, we consider that $\mathcal{H} = span\{\phi_1, .., \phi_m\}$ based on the spectral theorem [19].

At this stage, we give details of the manifold Gaussian processes (MGP) classifier, where the mapping Ψ and the Gaussian processes (GP) classifier are learned jointly from data. We use Laplace based method to approximate the Bayesian inference. To achieve this goal, we introduce a new latent variable f and we consider a new formulation of the logistic $\pi_f(x) = \sigma(f(x))$. The GP classification is based on placing a GP prior over the latent variable $f \sim \mathcal{GP}(0, c)$ where c is a covariance function [20]. By abuse of notation, we note $\mathbf{f} = (f_1, .., f_N)^T = (f(x_1), .., f(x_N))^T$. We remind that the Laplace approximation employs a Gaussian approximation $\hat{\mathbb{P}}(f|\mathbb{X}, \mathbb{Y})$ to the true posterior $\mathbb{P}(f|\mathbb{X}, \mathbb{Y})$ from the second order Taylor expansion to $\log \mathbb{P}(\mathbf{f}|\mathbb{X}, \mathbb{Y})$ around the MAP estimator: $\hat{\mathbf{f}} = \arg\max_\mathbf{f} \log \mathbb{P}(\mathbf{f}|\mathbb{X}, \mathbb{Y})$.

We first present the MGP classifier model and then we make connection to the standard GP classifier. Note that by doing so, this framework guides the learning of Ψ toward representations that are useful for the overall function $f = G \circ \Psi$. This later is the key insight where the mapping Ψ and the Gaussian processes classifier G are learned jointly following the same supervised objective. Let assume that Ψ is a deterministic, parametrized function that maps the input space $\mathbb{R}^p$ into $\mathcal{H}$, which serves as the domain for the GP classification $G : \mathcal{H} \to \mathbb{R}$. Therefore if the input vector $x \in \mathbb{R}^p$, the MGP is equivalent to a GP for $f : \mathbb{R}^p \to \mathbb{R}$ with a covariance function C such that $C(x, x') = c(\Psi(x), \Psi(x'))$ [20].

We recover $\mathbb{Z} = (z_1, .., z_N)^T = (\Psi(x_1), .., \Psi(x_N))^T$ and $\mathbf{M} = (G(z_1), .., G(z_N))^T$. For the learning part, we approximate the MAP donated $\hat{\mathbf{M}}$ by maximizing the posterior density $\mathbb{P}(\mathbf{M}|\mathbb{Z}, \mathbb{Y})$, i.e. $\hat{\mathbf{M}} = \arg\max_\mathbf{M} \mathbb{P}(\mathbf{M}|\mathbb{Z}, \mathbb{Y})$. By Bayes rule, its logarithm is proportional to

$$g(\mathbf{M}) = \log \mathbb{P}(\mathbb{Y}|\mathbf{M}) - \frac{1}{2}\mathbf{M}^T \bar{C}^{-1}\mathbf{M} \tag{5}$$

where $\bar{C}$ is the covariance matrix with $\bar{C}_{ij} = C(x_i, x_j)$. Note that $g(.)$ is concave leading to a unique optimum. From this proportionality we obtain a Gaussian approximation

$$\begin{aligned} \hat{\mathbb{P}}(\mathbf{M}|\mathbb{Z}, \mathbb{Y}) &= \mathcal{N}(\mathbf{M}|\hat{\mathbf{M}}, H^{-1}) \\ &\propto \exp(-\frac{1}{2}(\mathbf{M} - \hat{\mathbf{M}})^T H(\mathbf{M} - \hat{\mathbf{M}})) \end{aligned} \tag{6}$$

where $H = -\nabla^2 \log \mathbb{P}(\mathbf{M}|\mathbb{Z}, \mathbb{Y})_{|\mathbf{M}=\hat{\mathbf{M}}} = \bar{W} + \bar{C}^{-1}$ and $\bar{W}$ is a $N \times N$ diagonal matrix with $\bar{W}_{ii} = -\frac{\partial^2 \log \mathbb{P}(y_i|M_i)}{\partial^2 M_i}$ $= \frac{\exp(M_i)}{(1+\exp(M_i))^2}$. We use the Newton-based method iteratively to find the MAP estimator

$$\mathbf{M}^{k+1} = (\bar{C}^{-1} + \bar{W})^{-1}(\bar{W}\mathbf{M}^k + \nabla \mathbb{P}(\mathbb{Y}|\mathbf{M}^k)) \tag{7}$$

The predictive distribution for the MGP of a test input $M^* = G(\Psi(x^*)) = G(z^*)$ is

$$\hat{\mathbb{P}}(M^*|\mathbb{Z}, \mathbb{Y}, z^*) = \mathcal{N}(\mu(z^*), \sigma^2(z^*)) \tag{8}$$

$$\mu(z^*) = \bar{C}_*^T \bar{C}^{-1}\hat{\mathbf{M}} \tag{9}$$

$$\sigma^2(z^*) = \bar{C}_{**} - \bar{C}_*^T(\bar{C} + \bar{W}^{-1})^{-1}\bar{C}_* \tag{10}$$

where $\bar{C}_{**} = C(x^*, x^*)$ and $\bar{C}_* = (C(x_1, x^*), .., C(x_N, x^*))^T$. Given the mean $\mu(z^*)$ and the variance $\sigma^2(z^*)$ of M^*, we approximate the predictor for $y^* = 1$ as

$$\bar{\pi}_* \approx \mathbb{E}_{\hat{\mathbb{P}}}(\pi_*|\mathbb{Z}, \mathbb{Y}, z^*) = \int_{\mathbb{R}} \sigma(M^*)\hat{\mathbb{P}}(M^*|\mathbb{Z}, \mathbb{Y}, z^*)\mathrm{d}M^* \quad (11)$$

4. MANIFOLD GAUSSIAN PROCESS: A BAYESIAN INFERENCE WITH EXPECTATION PROPAGATION

The EP approach is usually used to approximate marginal moments and can be generalized for Gaussian processes [21]. The key idea is to replace the likelihood terms by unnormalized Gaussians, we refer to [11] for more details. We call this method manifold expectation propagation (MEP) model and we remind that the formulation is similar to MGP classifier, derived in Section 3, except that the Laplace approximation is replaced by the expectation propagation. We then use the same notations and write the posterior distribution over $\mathbf{M}$ as the product of the prior and the likelihood function

$$\mathbb{P}(\mathbf{M}|\mathbb{Z}, \mathbb{Y}) = \frac{1}{L}\mathbb{P}(\mathbf{M}|\mathbb{Z}) \times \prod_{i=1}^{N} \mathbb{P}(y_i|M_i) \quad (12)$$

where the normalization term is

$$L = \mathbb{P}(\mathbb{Y}|\mathbb{Z}) = \int_{\mathbb{R}^N} \mathbb{P}(\mathbf{M}|\mathbb{Z}) \times \prod_{i=1}^{N} \mathbb{P}(y_i|M_i)\mathrm{d}\mathbf{M} \quad (13)$$

In the following, we consider that $\mathbb{P}(y_i|M_i) = \Phi(y_i \times M_i)$, where $\Phi(.)$ denotes the probit likelihood for binary classification. It is clear that the posterior is analytically intractable. To build the MEP framework, we can approximate the likelihood, locally, with an un-normalized Gaussian

$$\begin{aligned} \mathbb{P}(y_i|M_i) &\approx t_i(M_i|\tilde{L}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) \\ &= \tilde{L}_i \times \mathcal{N}(M_i|\tilde{\mu}_i, \tilde{\sigma}_i^2) \end{aligned} \quad (14)$$

However, the likelihood approximation should not be normalized since the exact likelihood do not have this property. The product of the local likelihoods is $\prod_{i=1}^{N} \tilde{L}_i \times \mathcal{N}(\mathbf{M}|\tilde{\mu}, \tilde{\Sigma})$ where $\tilde{\mu} = (\tilde{\mu}_1, .., \tilde{\mu}_N)^T$ and $\tilde{\Sigma} = diag(\tilde{\sigma}_1^2, .., \tilde{\sigma}_N^2)$. Based on local approximations, the posterior can be approximated by

$$\begin{aligned} \hat{\mathbb{P}}(\mathbf{M}|\mathbb{Z}, \mathbb{Y}) &= \frac{\mathbb{P}(\mathbf{M}|\mathbb{Z})}{L} \times \prod_{i=1}^{N} t_i(M_i|\tilde{L}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) \\ &= \mathcal{N}(\mathbf{M}|\mu, \Sigma) \end{aligned} \quad (15)$$

with $\mu = \Sigma\tilde{\Sigma}^{-1}\tilde{\mu}$, $\Sigma = (\bar{C}^{-1} + \tilde{\Sigma}^{-1})^{-1}$.
To summarize, the iterative steps of our algorithm are:

1. Choose one $t_i(M_i|\tilde{L}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2)$ to update
2. Compute the cavity distribution of M_i

$$\hat{\mathbb{P}}_{-i}(M_i) \propto \frac{\hat{\mathbb{P}}(M_i|\mathbb{Z}, \mathbb{Y})}{t_i(M_i|\tilde{L}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2)} = \mathcal{N}(M_i|\mu_{-i}, \sigma_{-i}^2) \quad (16)$$

where $\hat{\mathbb{P}}(M_i|\mathbb{Z}, \mathbb{Y}) = \mathcal{N}(M_i|\mu_i, \sigma_i^2 = \Sigma_{ii})$, $\sigma_{-i}^2 = (\sigma_i^{-2} - \tilde{\sigma}_i^{-2})^{-1}$, and $\mu_{-i} = \sigma_{-i}^2(\sigma_i^{-2}\mu_i - \tilde{\sigma}_i^{-2}\tilde{\mu}_i)$

3. Define $\mathbb{P}_i(M_i)$, the pseudo-exact posterior marginal distribution of M_i, as

$$\mathbb{P}_i(M_i) = \mathbb{P}(y_i|M_i) \times \hat{\mathbb{P}}_{-i}(M_i) \quad (17)$$

4. Compute $\hat{\mathbb{P}}(M_i) = \hat{L}_i \times \mathcal{N}(M_i|\hat{\mu}_i, \hat{\sigma}_i^2)$ by minimizing the Kullback-Leibler divergence $(K.L)$,

$$(\hat{L}_i, \hat{\mu}_i, \hat{\sigma}_i^2) = \underset{(\hat{Z}_i, \hat{\mu}_i, \hat{\sigma}_i^2)}{\operatorname{argmin}} K.L(\mathbb{P}_i(M_i)||\hat{\mathbb{P}}(M_i)) \quad (18)$$

then the desired posterior marginal moments are

$$\begin{aligned} &\hat{L}_i = \Phi(l_i), \quad \hat{\sigma}_i^2 = \sigma_{-i}^2 - \frac{\sigma_{-i}^4 \times \mathcal{N}(l_i|0,1)}{(1+\sigma_{-i}^2) \times \Phi(l_i)}(l_i + \frac{\mathcal{N}(l_i|0,1)}{\Phi(l_i)}) \\ &\hat{\mu}_i = \mu_{-i} + \frac{y_i \times \sigma_{-i}^2 \times \mathcal{N}(l_i|0,1)}{\Phi(l_i) \times \sqrt{1+\sigma_{-i}^2}}, \quad l_i = \frac{y_i \times \mu_{-i}}{\sqrt{1+\sigma_{-i}^2}} \end{aligned} \quad (19)$$

5. Update $(\tilde{L}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2)$ with $t_i(M_i|\tilde{L}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) = \frac{\hat{\mathbb{P}}(M_i)}{\hat{\mathbb{P}}_{-i}(M_i)}$,

$$\begin{aligned} &\tilde{\mu}_i = \tilde{\sigma}_i^2(\hat{\sigma}_i^{-2}\hat{\mu}_i - \sigma_{-i}^{-2}\mu_{-i}), \quad \tilde{\sigma}_i^2 = (\hat{\sigma}_i^{-2} - \sigma_{-i}^{-2})^{-1}, \\ &\tilde{L}_i = \hat{L}_i\sqrt{2\pi(\sigma_{-i}^2 + \tilde{\sigma}_i^2)}\exp(\tfrac{1}{2}\tfrac{(\mu_{-i} - \tilde{\mu}_i)^2}{\sigma_{-i}^2 + \tilde{\sigma}_i^2}) \end{aligned} \quad (20)$$

As in the previous section, the prediction by MEP are

$$\mu(z^*) = \bar{C}_*^T\bar{C}^{-1}\mu = \bar{C}_*^T(\bar{C} + \tilde{\Sigma})^{-1}\tilde{\mu} \quad (21)$$

$$\sigma^2(z^*) = \bar{C}_{**} - \bar{C}_*^T(\bar{C} + \tilde{\Sigma})^{-1}\bar{C}_* \quad (22)$$

Therefore, the approximate predictor for $y^* = 1$ is

$$\bar{\pi}_* = \Phi(\frac{\bar{C}_*^T(\bar{C} + \tilde{\Sigma})^{-1}\tilde{\mu}}{\sqrt{1 + \bar{C}_{**} - \bar{C}_*^T(\bar{C} + \tilde{\Sigma})^{-1}\bar{C}_*}}) \quad (23)$$

5. EXPERIMENTAL RESULTS

We evaluate the performance and the efficiency of the proposed methods on a database of 2042 images of manufacturing defects. The database contains 530 images of defective metallic boxes and 1512 images of non-defective ones. Figure 1 shows two examples of original images and several extracted features (vertical gradient, binary gradient, and Gabor filter) used to represent them for evaluation. First, we test the efficiency of our framework to classify defective and non-defective images. We learn the model parameters from 75% of the dataset as training and use the rest for test. To evaluate the classification quality, we consider the False Negatives (FN: non-defective but classified as defective) and False Positives (FP: defective but classified as non-defective). The subdivision has been performed randomly at least 15 times and the recognition rates are given as a mean. We compare our

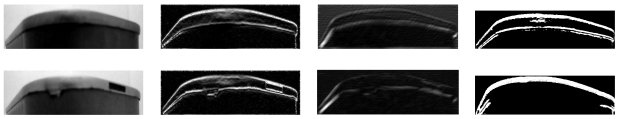

Fig. 1. Non-defective (top) and defective (bottom) boxes with different representations.

approach with two iterative methods: gradient and Newton as detailed in Section 2 using the same experimental protocol.

To compute the classification error, we first compute the estimator and its parameters from the training $(\mathbb{X}, \mathbb{Y})$ then, given a new observation x^* for test, we apply the regression model to determine its class y^*. We also use the ROC curve which generalizes the choice of the threshold by controlling the sensitivity and the specificity.

The error rates of logistic regression are summarized in Table 1. Accordingly, one can observe that vertical gradient achieves the lowest error with a significant margin.

Table 1. Classification performance using Newton-MLE.

Error rates \ features	gradient	Gabor	binarization
FP	20%	51%	53%
FN	27%	47%	43%

Results for method described in section 2. Figure 2 shows the False Positives, the False Negatives and the classification errors (CE) for different values of $\lambda \in]0, 1[$. Note that the Newton-based optimization outperforms the gradient descent, see Figure 2 (left and middle). This result is confirmed by the ROC curve of Newton-MLE weighted ridge in Figure 2 (right).

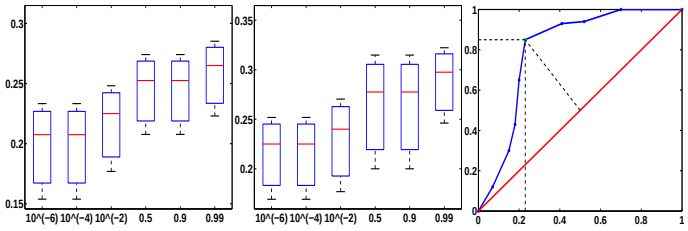

Fig. 2. Errors as a function of regularization parameters (FN=upper values, FP=lower values and CE=values in the middle) obtained by: Newton method (left) and gradient (middle). The ROC curve of Newton is given on the right.

Illustration of predictions using MGP and MEP. Figure 3 shows an example of the key steps to classify a new input z^*. For this particular example, we display details for MGP: the sigmoid function $\sigma(z)$ (red line), the posterior predictive law $\mathcal{N}(z, \mu(z^*) = -0.4, \sigma^2(z^*) = 0.8)$ (green line), the product of sigmoid and predictive law $\sigma(z) \times \mathcal{N}(z, \mu(z^*), \sigma^2(z^*))$ (blue line), and the area between the x-axis and the blue line $\bar{\pi}_* = 0.41$. We then use the same example for MEP: the probit function $\sigma(z)$ (red line), the posterior predictive law $\mathcal{N}(z, \mu(z^*) = -0.56, \sigma^2(z^*) = 0.68)$ (green line), the product of probit and the predictive law $\sigma(z) \times \mathcal{N}(z, \mu(z^*), \sigma^2(z^*))$ (blue line), and the area between the x-axis and the blue line $\bar{\pi}_* = 0.33$. We remark that, in this example with a test input $(y^* = 0)$, MEP has a better predictor with $\bar{\pi}_* = 0.33$ than MGP with $\bar{\pi}_* = 0.41$.

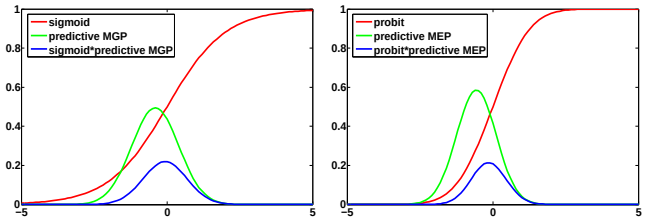

Fig. 3. An illustration of key steps to classify a test input $(y^* = 0)$ using MGP (left) and MEP (right).

Results using MGP and MEP. Table 2 summarizes results of the proposed methods. We can observe that both MGP and MEP improve the specificity and sensibility with better results for MEP. The ROC curves of Figure 4 confirm that MEP has the most predictive power and generalization capability where the risk of FP is approximately 8.5% and the risk of FN is 10%.

Table 2. Classification performance using MGP and MEP.

Error rates \ methods	MGP	MEP
FP	11%	8.5%
FN	19%	10%

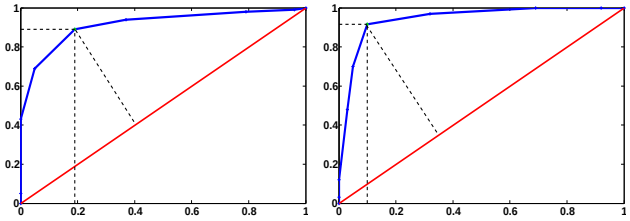

Fig. 4. ROC curves for MGP (left) and MEP (right).

6. CONCLUSION

In this paper, we have formulated the classification problem as a regression and have proposed an efficient solution when the feature vectors are high dimensional whereas the number of samples is relatively small. The proposed framework provides details of two Manifold-based inferences to build supervised Gaussian process classifiers. Experiments have been conducted to classify images of manufacturing defects and have shown that proposed methods achieve high accuracy.

7. REFERENCES

[1] Y.-J. Chen, J.-C. Tsai, and Y.-C. Hsu, "A real-time surface inspection system for precision steel balls based on machine vision," *Measurement Science and Technology*, vol. 21, no. 7, pp. 76–82, 2016. 1

[2] J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*. Springer Publishing Company, Incorporated, 2007. 1

[3] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006. 1

[4] P. R. Anukrishna and V. Paul, "A review on feature selection for high dimensional data," in *2017 International Conference on Inventive Systems and Control (ICISC)*, Jan 2017, pp. 1–4. 1

[5] I. Fodor, "A survey of dimension reduction techniques," Tech. Rep., 2002. 1

[6] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, Jan. 2014. 1

[7] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, 1970. 1

[8] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society (Series B)*, vol. 58, pp. 267–288, 1996. 1

[9] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, Dec. 2006. 1

[10] C. K. I. Williams and D. Barber, "Bayesian classification with gaussian processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1342–1351, 1998. 1

[11] C. E. Rasmussen and C. K. I. Williams, *Gaussian Pocesses for Machine Learning*, ser. Adaptive computation and machine learning. MIT Press, 2006. 1, 3

[12] T. P. Minka, "Expectation propagation for approximate bayesian inference," in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 2001, pp. 362–369. 1

[13] M. Opper and O. Winther, "Gaussian processes for classification: Mean-field algorithms," *Neural Computation*, vol. 12, no. 11, pp. 2655–2684, Nov. 2000. 1

[14] R. Calandra, J. Peters, C. E. Rasmussen, and M. P. Deisenroth, "Manifold gaussian processes for regression," in *2016 International Joint Conference on Neural Networks, IJCNN 2016, Vancouver, BC, Canada, July 24-29*, 2016, pp. 3338–3345. 1

[15] B. Schölkopf, R. Herbrich, and A. J. Smola, "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, 2001, pp. 416–426. 1

[16] J. M. Pereira, M. Basto, and A. F. da Silva, "The logistic lasso and ridge regression in predicting corporate failure," *Procedia Economics and Finance*, vol. 39, pp. 634 – 641, 2016. 2

[17] A. Albert and J. A. Anderson, "On the existence of maximum likelihood estimates in logistic regression models," *Biometrika*, vol. 71, no. 1, pp. 1–10, 1984. 2

[18] T. Gneiting, W. Kleiber, and M. Schlather, "Matérn cross-covariance functions for multivariate random fields," *Journal of the American Statistical Association*, vol. 105, pp. 1167–1177, 2010. 2

[19] J. Weidmann, *Linear Operators in Hilbert Spaces*. Graduate Texts in Mathematics, 1980, vol. 68. 2

[20] M. L. Stein, *Interpolation of Spatial Data*, ser. Springer Series in Statistics. Springer-Verlag New York, 1999. 2

[21] T. P. Minka, *A Family of Algorithms for Approximate Bayesian Inference*. Cambridge, MA, USA: Massachusetts Institute of Technology, 2001. 3