A Bayesian Kernel Logistic Discriminant Model: An Improvement to the Kernel Fisher's Discriminant

R. Ksantini[1], D. Ziou[2], B. Colin[2] and F. dubeau[2]
(1) University of Windsor, Computer Science School
ksantini@uwindsor.ca
(2) University of Sherbrooke, Computer Science and Mathematic Department
{djemel.ziou, bernard.colin, francois.dubeau}@usherbrooke.ca

Abstract

The Kernel Fisher's Discriminant (KFD) has proven to be competitive to several state-of-the-art classifiers. However, it is assuming equal covariance structure for all transformed classes, which is not true in many applications. In this paper, we propose a novel Bayesian Kernel Logistic Discriminant model (BKLD) which goes one step further by representing each transformed class by its own covariance matrix. This can perform better than the KFD. An extensive comparison of the BKLD to the KFD and to other state-of-the-art non-linear classifiers is performed.

Introduction

The KFD was proposed by (Mika et al. 1999) and its main idea is to perform the traditional Fisher's linear discriminant in the feature space obtained by the kernel trick. However, it suffers from the small sample size problem since the kernel-induced feature space is typically of very high dimensionality. Furthermore, it is incapable of dealing with heteroscedastic data (classes with different covariance matrices) that are commonly found in real-world applications. In this paper, we propose a Bayesian Kernel Logistic Discriminant model (BKLD) which is capable of dealing with heteroscedastic data by representing each transformed class by its own covariance matrix. This can perform better than the KFD. The posterior distribution of the BKLD model is elegantly approximated by a tractable Gaussian form using variational transformation and Jensen's' inequality, which allow a straightforward computation of the weights. In order to avoid small sample size problem and to speed up the computation of the model weights, we introduce a sparsity-promoting Gaussian prior over them. In the next section, we detail the derivation of the BKLD. In the third section, we compare the BKLD to the KFD and other non-linear classifiers on a collection of data sets. Finally, we present our conclusions.

Copyright © 2008, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The Bayesian Kernel Logistic Discriminant Model

Let $\mathcal{X}_1 = \{X_i\}_{i=1}^{N_1}$ and $\mathcal{X}_2 = \{X_i\}_{i=N_1+1}^{N}$ be two different classes constituting an input space of N samples or vectors. Applying the kernel trick, we use a function Φ to map the classes $\mathcal{X}_1$ and $\mathcal{X}_2$ to two feature classes $\mathcal{F}_1 = \{\Phi(X_i)\}_{i=1}^{N_1}$ and $\mathcal{F}_2 = \{\Phi(X_i)\}_{i=N_1+1}^{N}$, respectively, wherein $\Phi(X_i) = (1, \mathcal{K}(X_i, X_1), \mathcal{K}(X_i, X_2), ..., \mathcal{K}(X_i, X_N)) \; \forall \; i \in \{1, 2, ..., N\}$, where $\mathcal{K}$ is a kernel function. Let us denote by $\underline{\Phi}_1$ and $\underline{\Phi}_2$ two random vectors whose realizations represent the vectors of $\mathcal{F}_1$ and the vectors of $\mathcal{F}_2$, respectively. We suppose that $\underline{\Phi}_1 \sim g_1(\underline{\Phi}_1)$ and $\underline{\Phi}_2 \sim g_2(\underline{\Phi}_2)$, where g_1 and g_2 are two Gaussian distributions whose means and covariance matrices are empirically computed from $\mathcal{F}_1$ and $\mathcal{F}_2$. With $\underline{\Phi}_1$ we associate a tag $t_1 = 0$, and with $\underline{\Phi}_2$ we associate a tag $t_2 = 1$. The unknown parameters (weights) are considered as random variables and are denoted by the random vector $\mathbf{w} = (w_0, w_1, ..., w_N)$. We define a 'likelihood' function as:

$$P(t_1 = 0, t_2 = 1|\mathbf{w}) = \sum_{\underline{\Phi}_1 \in \mathcal{F}_1, \underline{\Phi}_2 \in \mathcal{F}_2} \left[\prod_{i=1}^{2} P(t_i = i - 1|\underline{\Phi}_i, \mathbf{w}) g_i(\underline{\Phi}_i) \right], \quad (1)$$

where, given $F(x) = \frac{e^x}{1+e^x}$, $P(t_i = i-1|\underline{\Phi}_i, \mathbf{w}) = F((2i-3)\mathbf{w}^T\underline{\Phi}_i) \; \forall \; i \in \{1, 2\}$ represent logistic modelings of t_1 and t_2 given the realizations of $\underline{\Phi}_1$ and $\underline{\Phi}_2$, respectively. We adopt a Bayesian perspective, and 'constrain' the parameters by defining an zero-mean Gaussian prior distribution over $\mathbf{w}$:

$$\pi(\mathbf{w}|\boldsymbol{\beta}) = \prod_{i=0}^{N} \mathcal{N}(w_i|0, \beta_i^{-1}), \quad (2)$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_N)$ a vector of $N + 1$ prior parameters. Having defined the prior, Bayesian inference proceeds by computing, from the Bayes' rule, the posterior over the unknown weights:

$$P(\mathbf{w}|t_1 = 0, t_2 = 1) = \frac{\sum_{\underline{\Phi}_1 \in \mathcal{F}_1, \underline{\Phi}_2 \in \mathcal{F}_2} \left[\prod_{i=1}^{2} P(t_i = i - 1|\underline{\Phi}_i, \mathbf{w}) g_i(\underline{\Phi}_i) \right] \pi(\mathbf{w}|\boldsymbol{\beta})}{P(t_1 = 0, t_2 = 1)}. \quad (3)$$

The computation of the posterior distribution is intractable. However, we can approximate it by a variational posterior approximation with a Gaussian form. To obtain this approximation, we perform two successive approximations to the likelihood function, in order to bound it by an exponential form which is a conjugate of the Gaussian prior. The first approximation is based on a variational transformation of the sigmoid function $F(x)$ in $H_i = (2i-3)\mathbf{w}^T\underline{\Phi}_i \; \forall \; i \in \{1,2\}$ (Jaakkola and Jordan 2000). So the likelihood function can be approximated as follows:

$$\sum_{\underline{\Phi}_1 \in \mathcal{F}_1, \underline{\Phi}_2 \in \mathcal{F}_2} \left[\prod_{i=1}^{2} P(t_i = i-1 | \underline{\Phi}_i, \mathbf{w}) g_i(\underline{\Phi}_i) \right] \geq \sum_{\underline{\Phi}_1 \in \mathcal{F}_1, \underline{\Phi}_2 \in \mathcal{F}_2} \left[\prod_{i=1}^{2} P(t_i = i-1 | \underline{\Phi}_i, \mathbf{w}, \epsilon_i) g_i(\underline{\Phi}_i) \right], \quad (4)$$

where $P(t_i = i-1|\underline{\Phi}_i, \mathbf{w}) = F(\epsilon_i) e^{\frac{(H_i - \epsilon_i)}{2} - \varphi(\epsilon_i)\left(H_i^2 - \epsilon_i^2\right)}$, $\epsilon_i > 0$ is the variational parameter and $\varphi(\epsilon_i) = \frac{tanh(\frac{\epsilon_i}{2})}{4\epsilon_i}$. The second approximation is based on Jensen's' inequality, which uses the convexity of the function e^x. Using Jensen's' inequality, we obtain

$$\begin{aligned} & \sum_{\underline{\Phi}_1 \in \mathcal{F}_1, \underline{\Phi}_2 \in \mathcal{F}_2} \left[\prod_{i=1}^{2} P(t_i = i-1 | \underline{\Phi}_i, \mathbf{w}, \epsilon_i) g_i(\underline{\Phi}_i) \right] \quad (5) \\ \geq & \left[\prod_{i=1}^{2} F(\epsilon_i) \right] e^{\left[\sum_{i=1}^{2} \left[\frac{E_{g_i}[H_i] - \epsilon_i}{2} \right] - \sum_{i=1}^{2} \left[\varphi(\epsilon_i)\left(E_{g_i}[H_i^2] - \epsilon_i^2\right) \right] \right]}, \\ = & \underline{\mathrm{P}}\left(t_1 = 0, t_2 = 1 | \mathbf{w}, \left\{\epsilon_i\right\}_{i=1}^{2}\right), \end{aligned}$$

where E_{g_1} and E_{g_2} are the expectations with respect to g_1 and g_2, respectively. Finally, given that $\pi(\mathbf{w}|\boldsymbol{\beta})$ is a Gaussian which is a conjugate of the exponential variational form $\underline{\mathrm{P}}\left(t_1 = 0, t_2 = 1 | \mathbf{w}, \left\{\epsilon_i\right\}_{i=1}^{2}\right)$, the variational posterior approximation denoted by $P\left(\mathbf{w}|t_1 = 0, t_2 = 1, \left\{\epsilon_i\right\}_{i=1}^{2}, \boldsymbol{\beta}\right)$ is a Gaussian with mean μ_{post} and covariance matrix Σ_{post}. Thus, omitting the algebra, Σ_{post} and μ_{post} are given by the following Bayesian update equations:

$$(\Sigma_{post})^{-1} = A^{-1} + 2\sum_{i=1}^{2} \left[\varphi(\epsilon_i) E_{g_i}[\underline{\Phi}_i \underline{\Phi}_i^T]\right], \quad (6)$$

$$\mu_{post} = \Sigma_{post}\left[\sum_{i=1}^{2} \left[(i - \frac{3}{2}) E_{g_i}[\underline{\Phi}_i]\right]\right], \quad (7)$$

with $A = \mathrm{diag}(\beta_0^{-1}, \beta_1^{-1}, ..., \beta_N^{-1})$. We have to find the values of $\left\{\epsilon_i\right\}_{i=1}^{2}$ and $\left\{\beta_i\right\}_{i=0}^{N}$ that yield a tight lower bound in (5). In the EM formalism, this can be achieved by iteratively maximizing the following expectation

$$\int log\left(\underline{\mathrm{P}}\left(t_1 = 0, t_2 = 1 | \mathbf{w}, \left\{\epsilon_i\right\}_{i=1}^{2}\right) \pi(\mathbf{w}|\boldsymbol{\beta})\right) P\left(\mathbf{w}|t_1 = 0, t_2 = 1, \left(\left\{\epsilon_i\right\}_{i=1}^{2}\right)^{old}, \boldsymbol{\beta}^{old}\right) d\mathbf{w}.$$

Taking the partial derivatives of the above expectation with respect to $\left\{\epsilon_i\right\}_{i=0}^{1}$ and $\left\{\beta_i\right\}_{i=0}^{N}$ and equalizing to zero leads to

$$\epsilon_i^2 = E_{g_i}[\underline{\Phi}_i^T \Sigma_{post} \underline{\Phi}_i] + \mu_{post}^T \left[E_{g_i}[\underline{\Phi}_i \underline{\Phi}_i^T]\right] \mu_{post}, \quad (8)$$

$$\beta_j = \frac{1}{\Sigma_{post,jj} + \mu_{post,j}^2}, \quad (9)$$

$\forall \; i \in \{1,2\}$ and $j \in \{0,1,...,N\}$. The weight computation algorithm has two phases. The first phase is the initialization; the second is iterative and allows the computation of Σ_{post} and μ_{post} through the Bayesian update equations (6) and (7), respectively, while using equations (8) and (9) to find the variational parameters and prior parameters at each iteration.

Experimental Results

We compared the BKLD to the KFD, the single RBF classifier, the regularized AdaBoost (AB_R) and the SVM (with Gaussian RBF kernel $\mathcal{K}(X, X_i) = e^{-||X - X_i||^2/\sigma}$), where σ is the positive 'width' parameter. For the BKLD we used Gaussian RBF too. We used 8 artificial and real word data sets.[1] On each of these data sets we trained and tested all classifiers. The optimization of the necessary parameters for each classifier were performed using a 5-fold cross validation procedure (see (Ratsch, Onoda, and Muller 2000) for details). The results in table 1 show the test classification errors. The experiments show that the BKLD is superior to the KFD and other classifiers on all data sets.

	RBF	AB_R	SVM	KFD	BKLD
B. Cancer	27.6	26.5	26.0	*25.8*	**25.1**
Diabetes	24.3	23.8	23.5	*23.2*	**22.7**
Heart	17.6	16.5	*16.0*	16.1	**15.7**
Ringnorm	1.7	1.6	1.7	*1.5*	**0.8**
F. Solar	34.4	34.2	*32.4*	33.2	**30.9**
Thyroid	4.5	4.6	4.8	*4.2*	**4.0**
Twonorm	2.9	2.7	3.0	*2.6*	**1.8**
Waveform	10.7	*9.8*	9.9	9.9	**8.7**

Table 1: *Comparison among the five classifiers: Estimation of the test classification errors in % on 8 data sets (best method in bold face, second best emphasized).*

Conclusion

We have proposed a novel BKLD model which is capable of dealing with heteroscedastic data. However, it assumes that the classes are normally distributed. Therefore, future work will be dedicated to make the BKLD adaptive to multimodel data.

References

Jaakkola, T. S., and Jordan, M. I. 2000. Bayesian parameter estimation via variational methods. *Statistics and Computing* 10(1): 25-37.

Mika, S. *et al.* 1999. Fisher Discriminant Analysis with Kernels. *In Proceedings of IEEE Neural Networks for Signal Processing Workshop*, 41-48.

Ratsch, G., Onoda, T., and Muller, K. -R. 2000. Soft Margins for Adaboost. *Machine Learning* 42(3): 287-320.

[1] The data sets can be obtained via **http://www.first.gmd.de/~raetsch/**