ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 75-80

Flexible and Robust Bayesian Classification by Finite Mixture Models

Cédric Archambeau, Fédéric Vrins and Michel Verleysen*

Université catholique de Louvain (UCL) - Machine Learning Group
Place du Levant 3, 1348 Louvain-la-Neuve, Belgium
{archambeau,vrins,verleysen}@dice.ucl.ac.be

Abstract. The regularized Mahalanobis distance is proposed in the framework of finite mixture models to avoid commonly faced numerical difficulties encountered with EM. Its principle is applied to Gaussian and Student-t mixtures, resulting in reliable density estimates, the model complexity being kept low. Besides, the regularized models are robust to various noise types. Finally, it is shown that the quality of the associated Bayesian classification is near optimal on Ripley's synthetic data set.

1 Introduction

Probability density function (PDF) estimation is essential in pattern recognition, exploratory data analysis and other related fields [5, 3]. Among others, it provides a solid basis to learning tasks such as clustering and classification.

A popular technique for estimating an unknown PDF nonparametrically, i.e. not assuming any a priori shape, is by the Parzen window estimator [9]. It consists in placing a Gaussian kernel on each data point of the learning set. The PDF is approximated by summing all the kernels, which are multiplied by a normalizing factor. While the computational complexity of Parzen's estimator is relatively small, its model complexity is large, increasing unnecessarily for large data sets as it is proportional to the number of learning samples. As a result, when using Parzen's estimator, oscillations may occur in the distribution tails and in less dense regions of the input space. This leads to inaccuracies in the PDF estimates and their associated Bayesian classification.

By contrast, finite mixture models (FM) are used in semi-parametric PDF estimation problems or clustering tasks [8]. The unknown PDF is approximated by a weighted sum of mixture distributions. In general, finite Gaussian mixture models (FGM) are used. However, Student-t mixture models (FTM) are efficient alternatives that can deal with a limited number of outliers [10]. Both model types

*M.V. is a Senior Research Associate of the Belgian FNRS. This work was partially supported by the European Commission (IST-2000-25145) and the Belgian FRSM (# 3.4590.02).

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 75-80

provide a powerful tool when the number of components in the mixture is small, can be guessed easily and the data clusters are Gaussian shape. Each density in the mixture is then fitted to a cluster of data samples in the learning set.

Yet, FM may be used in a more general framework in order to estimate PDFs of arbitrary shape [3], the number of components being learnt as a hyperparameter. Compared to Parzen's estimator, FM are computationally expensive, but their flexibility is increased by the introduction of weighting factors. This leads to a lower model complexity, which in turns shows better generalization capabilities and avoids oscillations in the PDF estimates. In practice however, when the number of components increases, numerical difficulties arise, as mixture densities may possibly collapse. This problem was extensively discussed in [1] and accredited to the concept of isolation. In [2], it was proposed to use the regularized Mahalanobis distance [6] in order to improve the density estimates of FGM. In this paper, we extend its use to FTM and show that the quality of the resulting Bayesian classification is high for both regularized FM.

This paper is organized as follows. In Section 2, FGM and FTM are recalled. In Section 3, the use of the regularized Mahalanobis distance is proposed and motivated in the context of FM. Finally, in Section 4, the classification performance of regularized FGM (RFGM) and regularized FTM (RFTM) on Ripley's synthetic data set [11] is presented and discussed. More specifically, the robustness of RFGM and RFTM is assessed by corrupting the learning set by additive Gaussian noise or uniform random noise (atypical observations).

2 Finite Mixture Models

Finite mixture models [8] can approximate any continuous PDF, provided the model has a sufficient number of components and provided the parameters of the model are chosen correctly [3]. Let us consider a d-dimensional continuous random vector $X \in \Re^d$. Its true PDF can be approximated by a linear combination of M component densities $p(\mathbf{x}|j)$:

$$p(\mathbf{x}) = \sum_{j=1}^{M} P(j)p(\mathbf{x}|j), \tag{1}$$

where the mixing proportions $P(j)$ are non-negative and must sum to one. Consider an i.i.d. realization $\chi = \{\mathbf{x}_n\}_{n=1}^{N}$ of X. We may define the corresponding log-likelihood function:

$$L(\theta) = \log \prod_{n=1}^{N} p(\mathbf{x}_n), \tag{2}$$

where θ summarizes the model parameters. By means of the expectation-maxi mization (EM) algorithm [4], the maximum likelihood estimate of θ can be approximated iteratively, avoiding the intricacy of non-linear optimization schemes.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 75-80

Maximizing the log-likelihood function is then equivalent to finding the most probable PDF estimate provided the data set χ. For further details on EM and its extensions, we refer to [7].

A popular choice for the component density $p(\mathbf{x}|j)$ is the multivariate Gaussian distribution. Each component j is then characterized by its center $\mathbf{c}_j$ and its covariance matrix Σ_j.

When the data set contains atypical observations such as outliers, the estimates of the centers $\mathbf{c}_j$ and the covariance matrices Σ_j of the Gaussian components can be severely affected. Therefore, providing protection against outliers is essential in many practical problems. Robustness can be introduced by embedding the Gaussian distribution of each mixture component in a wider class of elliptically symmetric distributions, called the t-Student distribution. As a result, a heavy-tailed alternative to the Gaussian family is provided:

$$p(\mathbf{x}|j) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{(\nu\pi)^{\frac{d}{2}}|\Sigma_j|^{1/2}\Gamma\left(\frac{\nu}{2}\right)}\left[1+\frac{1}{\nu}\left(\mathbf{x}-\mathbf{c}_j\right)^T\Sigma_j^{-1}\left(\mathbf{x}-\mathbf{c}_j\right)\right]^{-\frac{\nu+d}{2}}. \quad (3)$$

In this equation, $\Gamma(\cdot)$ denotes the gamma function. Parameter ν is called the degree of freedom. It may be viewed as a robustness tuning parameter. If ν tends to infinity, the t distribution tends to the Gaussian distribution.

Applying EM to (2) leads to the following iterative scheme for FTM [10]:

E-step:

$$P^{(t)}(j|\mathbf{x}_n) = \frac{p^{(t)}(\mathbf{x}_n|j)P^{(t)}(j)}{p^{(t)}(\mathbf{x}_n)}, \quad (4)$$

$$Q^{(t)}(j|\mathbf{x}_n) = \frac{\nu+d}{\nu+\frac{1}{\nu}\left(\mathbf{x}_n-\mathbf{c}_j^{(t)}\right)^T{\Sigma_j^{(t)}}^{-1}\left(\mathbf{x}_n-\mathbf{c}_j^{(t)}\right)}. \quad (5)$$

M-step:

$$\mathbf{c}_j^{(t+1)} = \frac{\sum_{n=1}^N P^{(t)}(j|\mathbf{x}_n)Q^{(t)}(j|\mathbf{x}_n)\mathbf{x}_n}{\sum_{n=1}^N P^{(t)}(j|\mathbf{x}_n)Q^{(t)}(j|\mathbf{x}_n)}, \quad (6)$$

$$\Sigma_j^{(t+1)} = \frac{\sum_{n=1}^N P^{(t)}(j|\mathbf{x}_n)Q^{(t)}(j|\mathbf{x}_n)\left(\mathbf{x}_n-\mathbf{c}_j^{(t+1)}\right)\left(\mathbf{x}_n-\mathbf{c}_j^{(t+1)}\right)^T}{\sum_{n=1}^N P^{(t)}(j|\mathbf{x}_n)}, \quad (7)$$

$$P^{(t+1)}(j) = \frac{1}{N}\sum_{n=1}^N P^{(t)}(j|\mathbf{x}_n). \quad (8)$$

Theoretically, the degree of freedom ν can be estimated by EM [8]. However, in practice, the convergence of EM is slow for unknown ν and the one-dimensional search of its solution, computed at each iteration step, is very time consuming. Therefore, ν is considered as a regular hyperparameter and is learnt in a classical way by exhaustive search.

The corresponding E- and M-step for FGM are easily obtained by setting $Q^{(t)}(j|\mathbf{x}_n)$ equal to 1 for each component j and each data sample $\mathbf{x}_n$. Indeed, by computing the limit, we get $\lim_{\nu\to+\infty} Q^{(t)}(j|\mathbf{x}_n) = 1$.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 75-80

3 Regularized Mahalanobis Distance

When one wants to approximate an unknown PDF of arbitrary shape by increasing the number of components arbitrarily, numerical difficulties might occur. This problem was traced in [1] and linked to the concept of isolation. Actually, when maximizing the log-likelihood function $L(\theta)$, the width of a component in the mixture may tend to zero. Yet, if sufficient data samples are available and the singularities of the likelihood function can be avoided, we may approximate the true PDF arbitrarily well. In order to recover from singular covariance matrices, the regularized Mahalanobis distance is proposed.

The Mahalanobis distance D_M is defined as follows:

$$D_M(\mathbf{x}_n, \mathbf{c}_j) = (\mathbf{x}_n - \mathbf{c}_j)^T \Sigma_j^{-1} (\mathbf{x}_n - \mathbf{c}_j) . \tag{9}$$

From (3), one can easily see that both the multivariate Gaussian distribution and the multivariate Student-t distribution use D_M to determine their shape. When the number of data samples contributing to the computation of the covariance matrix of a component is small with respect to the dimension d of the data samples, it may be singular. Moreover, as discussed in [6], D_M tends to produce hyperellipsoidal components, leading to unusually large and elongated densities. By contrast, when we consider the Euclidean distance D_E, large data clusters tend to split unnecessarily, as the component densities are hyperspherical. If the covariance matrix Σ_j is replaced by the $d \times d$ identity matrix I in (9), one finds the definition of the Euclidean distance.

Based on the hyperspherical character of D_E and the hyperellipsiodal character of D_M, we can construct the *regularized Mahalanobis distance* [6], which is a convex combination of both distances:

$$D_{ME}(\mathbf{x}_n, \mathbf{c}_j) = (\mathbf{x}_n - \mathbf{c}_j)^T \left[(1-\lambda) (\Sigma_j + \epsilon I)^{-1} + \lambda I \right] (\mathbf{x}_n - \mathbf{c}_j) , \tag{10}$$

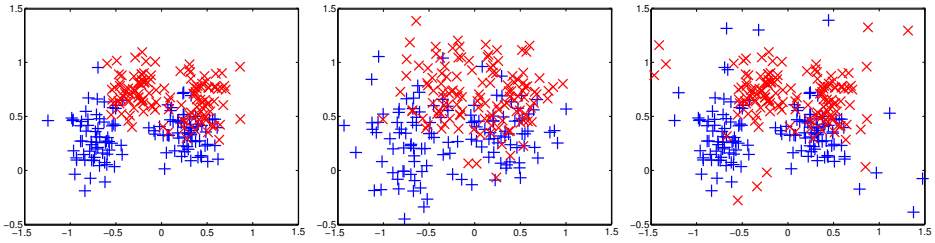
where ϵ and λ are learning parameters. Parameter λ is in the interval $[0, 1]$ and should be learnt properly. It controls the tradeoff between D_M and D_E. When the covariance matrices cannot be estimated reliably, a large value of λ should be used. By contrast, a careful estimation of ϵ is not required. Indeed, its role is to stabilize the learning process by converting a singular matrix to a non-singular one. As a result, using different values of ϵ do not make much difference as long as they are significantly smaller than the average variance of the data samples [2].

Consider again the E- and M-step for computing the model parameters of FGM and FTM. Introducing the regularized Mahalanobis distance consists in adapting, at each iteration step t, the covariance matrix of each component density according to (10). Therefore, the following adaptation rule is inserted in the M-step:

$$\Sigma'^{(t+1)}_j = \left[(1-\lambda) \left(\Sigma_j^{(t+1)} + \epsilon I \right)^{-1} + \lambda I \right]^{-1} , \tag{11}$$

where $\Sigma_j^{(t+1)}$ is still computed according to (7).

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 75-80



(a) (b) (c)

Figure 1: Ripley's learning set (a), corrupted by (b) additive Gaussian noise ($\sigma = 0.2$) or (c) uniform random noise (10% atypical observations). The class labels are denoted by '+' and '×' respectively.

4 Results and Discussion

FM that exploits the regularized Mahalanobis distance can be used for supervised classification. First, one learns the unknown PDF $p(\mathbf{x}|c = C)$ of each class C. Next, each data sample can be classified according to *Bayes' rule*:

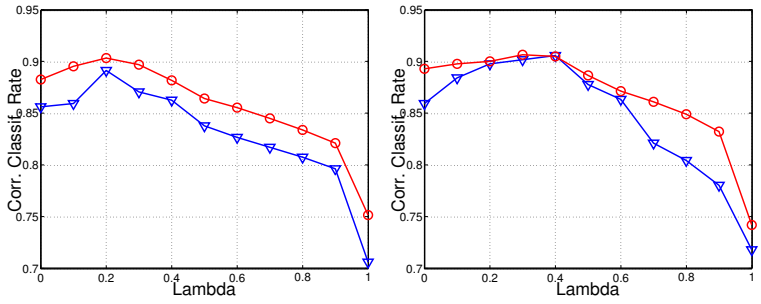
$$\hat{C}(\mathbf{x}_n) = \arg\max_c p(\mathbf{x}_n|c = C)P(c = C), \tag{12}$$

where $P(c = C)$ is the class prior of class C.

In Figure 1, Ripley's synthetic learning set is shown [11]. For both regularized FM (RFGM and RFTM), learning has been performed using learning sets corrupted by two types of noise. Both are common in practice. First, a strong additive Gaussian noise was considered ($\sigma = 0.2$). Subsequently, the robustness of the models to uniform random noise in the input space was investigated (up to 10% atypical observations). Their class label were randomly assigned.

Figure 2 shows the average correct classification rate on Ripley's test set [11], when learning is biased by the two types of noise. In both cases, RFTM outperforms RFGM. The results should be compared to the Bayes' optimal error rate of 8%, obtained when the class densities are known and the learning set is noiseless. In presence of the Gaussian noise source, we obtain an error rate of 10.8% and 9.6% for RFGM and RFTM ($\nu = 7$) respectively. For both techniques it was found that the optimal parameter $\lambda_{opt} = 0.2$. The error rate in presence of uniform random noise is 9.4% and 9.3% for RFGM and RFTM ($\nu = 5$) respectively. While in this case RFTM only slightly outperforms RFGM, one can see that the RFTM estimate is more flexible as it uses less constrained covariance matrices ($\lambda_{RFGM,opt} > \lambda_{RFTM,opt}$).

The number of components M, and the optimal parameters λ and ν were selected by exhaustive search. The models were computed using a learning set χ_L and their performance was evaluated on a test set χ_T. It was assumed that the true class PDFs were unknown. Both RFGM and RFTM used 5 mixture components during learning. It should be emphasized that the regular FGM and FTM could not be computed as they were collapsing.



(a) (b)

Figure 2: Correct classification rate of Ripley's test set in function of λ ($\epsilon = 10^{-6}$). The learning set is corrupted by (a) additive Gaussian noise and (b) uniform random noise. RFGM and RFTM are denoted by '$\bigtriangledown$' and '$\bigcirc$' respectively.

5 Conclusion

The regularized Mahalanobis distance was introduced in the context of finite mixture models, making them applicable in a wider range of problems, as for example for nonparametric PDF estimation. Regularized Gaussian mixtures and regularized Student-t mixtures were proposed. They avoid the numerical difficulties faced with finite mixture models when estimating the model parameters by the EM algorithm. It was shown that the resulting PDF estimates are flexible and robust to different types of noise. Finally, it was shown that the associated Bayesian classification is near optimal on Ripley's data set.

References

[1] C. Archambeau, J.A. Lee, and M. Verleysen. On the convergence problems of the *EM* algorithm for finite Gaussian mixtures. In *ESANN'03*, pages 99–106, Bruges, Belgium.

[2] C. Archambeau and M. Verleysen. Fully nonparametric probability density estimation with finite Gaussian mixture models. In *ICAPR'03*, pages 81–84, Calcutta, India.

[3] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford university press, 1995.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the *EM* algorithm (with discussion). *J. Roy. Stat. Soc., B*, 39:1–38, 1977.

[5] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.

[6] J. Mao and A. K. Jain. A self-orgainizing network for hyperellipsoidal clustering (HEC). *IEEE trans. Neural Networks*, 7(1):16–29, 1996.

[7] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, NY, 1997.

[8] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York, NY, 2000.

[9] E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Stat.*, 33:1065–1076, 1962.

[10] D. Peel and G. J. McLachlan. Robust mixture modelling using the t distribution. *Stat. Comp.*, 10(4):339–348, 2000.

[11] B. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.