Manifold Constrained Variational Mixtures

Cédric Archambeau* and Michel Verleysen**

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

Abstract. In many data mining applications, the data manifold is of lower dimension than the dimension of the input space. In this paper, it is proposed to take advantage of this additional information in the frame of variational mixtures. The responsibilities computed in the VBE step are constrained according to a discrepancy measure between the Euclidean and the geodesic distance. The methodology is applied to variational Gaussian mixtures as a particular case and outperforms the standard approach, as well as Parzen windows, on both artificial and real data.

1 Introduction

Finite mixture models [1] are commonly used for clustering purposes and modeling unknown densities. Part of their success is due to the fact that their parameters can be computed in an elegant way by the expectation-maximization algorithm (EM) [2]. Unfortunately, it is well known that mixture models suffer from an inherent drawback. EM maximizes iteratively the data log-likelihood, which is an ill-posed problem that can lead to severe overfitting; maximizing the likelihood may result in setting infinite probability mass on a single data point.

Among others, the variational Bayesian framework was introduced in order to avoid this problem [3]. In variational Bayes (VB) a factorized approximation of the joint posterior of the latent variables and the model parameters is used in order to compute a variational lower bound on the marginal data likelihood. In addition, VB allows determining the optimal number of components in the mixture by comparing the value of this variational lower bound. In [4] a variant was proposed to perform automatic model selection.

Recently, manifold Parzen [5] was introduced in order to improve nonparametric density estimation when the data is lying on a manifold of lower dimensionality than the one of the input space. In this paper, a related technique for variational mixtures is proposed by constraining the responsibilities according to the mismatch between the Euclidean and the geodesic distance. The key idea is to favor the directions along the manifold when estimating the unknown density, rather than wasting valuable density mass in directions perpendicular to the manifold orientation. The approach is applied to VB Gaussian mixtures as a particular case. Manifold constrained variational Gaussian mixtures (VB-MFGM) are compared experimentally to standard VB-FGM and standard Parzen.

* C.A. is supported by the European Commission project IST-2000-25145.

** M.V. is a Senior Research Associate of Belgian National Fund for Scientific Research.

W. Duch et al. (Eds.): ICANN 2005, LNCS 3697, pp. 279–284, 2005.

2 Variational Bayes for Mixtures Models

Let $X = \{\mathbf{x}_n\}_{n=1}^N$ be an i.i.d. sample, $Z = \{\mathbf{z}_n\}_{n=1}^N$ the latent variables associated to X and $\Theta = \{\boldsymbol{\theta}_m\}_{m=1}^M$ the model parameters, M being the number of mixture components. Finite mixture models are latent variable models in the sense that we do not know by which component a data sample was generated. We may thus associate to each data sample $\mathbf{x}_n$ a binary latent vector $\mathbf{z}_n$, with latent variables $z_{nm} \in \{0, 1\}$ that indicate which component has generated $\mathbf{x}_n$ ($z_{nm} = 1$ if $\mathbf{x}_n$ was generated by component m and 0 otherwise).

In Bayesian learning, both the latent variables Z and the model parameters Θ are treated as random variables. The quantity of interest is the marginal data likelihood, also called incomplete likelihood (i.e. of the observed variables only). For a fixed model structure $\mathcal{H}_M$, it is obtained by integrating out the nuisance parameters Z and Θ:

$$p(X|\mathcal{H}_M) = \int_\Theta \int_Z p(X, Z, \Theta|\mathcal{H}_M) dZ d\Theta \ . \tag{1}$$

This quantity is usually untractable. However, for any arbitrary density $q(Z, \Theta)$ a lower bound on $p(X|\mathcal{H}_M)$ can be found using Jensen's inequality:

$$\log p(X|\mathcal{H}_M) \geq \int_\Theta \int_Z q(Z, \Theta) \log \frac{p(X, Z, \Theta|\mathcal{H}_M)}{q(Z, \Theta)} dZ d\Theta \tag{2}$$

$$= \log p\left(X|\mathcal{H}_M\right) - \mathrm{KL}\left[q(Z, \Theta)||p\left(Z, \Theta|X, \mathcal{H}_M\right)\right] \ , \tag{3}$$

where KL[$\cdot$] is the Kullback-Leibler (KL) divergence. It is easily seen from (3) that the equality holds when $q(Z, \Theta)$ is equal to the joint posterior $p(Z, \Theta|X, \mathcal{H}_M)$.

In VB, the variational posterior approximates the joint posterior by assuming the latent variables and the parameters are independent:

$$p(Z, \Theta|X, \mathcal{H}_M) \approx q(Z, \Theta) = q(Z)q(\Theta) \ . \tag{4}$$

By assuming this factorization, the lower bound (2) on the marginal likelihood is tractable and the gap between both can be minimized by minimizing the KL divergence between the true and the variational posterior. Setting the derivatives of KL with respect to $q(Z)$ and $q(\Theta)$ to zero results in an EM-like scheme [6]:

$$\textbf{VBE step} : \quad q(Z) \propto \exp\left(\mathrm{E}_\Theta\{\log p(X, Z|\Theta, \mathcal{H}_M)\}\right) \ . \tag{5}$$

$$\textbf{VBM step} : \quad q(\Theta) \propto p(\Theta|\mathcal{H}_M) \exp\left(\mathrm{E}_Z\{\log p(X, Z|\Theta, \mathcal{H}_M)\}\right) \ . \tag{6}$$

In these equations $\mathrm{E}_\Theta\{\cdot\}$ and $\mathrm{E}_Z\{\cdot\}$ denote respectively the expectation with respect to Θ and Z, and $p(X, Z|\Theta, \mathcal{H}_M)$ is the complete likelihood (i.e. of the observed and unobserved variables). Note also that the prior $p(\Theta|\mathcal{H}_m)$ on the parameters appears in (6). If we choose $p(\Theta|\mathcal{H}_m)$ conjugate[1] to the exponential family, learning in the VB framework consists then simply in updating the parameters of the prior to the parameters of the posterior.

[1] The prior $p(\Theta)$ is said to be conjugate to $r(\Theta)$ if the posterior $q(\Theta)$ is of the same form as $p(\Theta)$, that is $q(\Theta) \propto p(\Theta)r(\Theta)$. In (6), $r(\Theta)$ is of the exponential family.

Since sample X is i.i.d., the posterior $q(Z)$ factorizes to $\prod_n q(\mathbf{z}_n)$. Furthermore, in the case of mixture models $q(\mathbf{z}_n)$ factorizes as well, such that $q(Z) = \prod_n \prod_m q(z_{nm})$. The resulting VBE step for mixture modes is:

$$q(z_{nm}) \propto \exp\left(\mathrm{E}_{\boldsymbol{\theta}_m}\{\log p(\mathbf{x}_n, \mathbf{z}_n | \boldsymbol{\theta}_m, \mathcal{H}_M)\}\right) \ , \tag{7}$$

where $\mathrm{E}_{\boldsymbol{\theta}_m}\{\cdot\}$ is the expectation with respect to $\boldsymbol{\theta}_m$. As in EM, the quantities computed in the VBE step are the responsibilities, each of them being proportional to the posterior probability of having a component m when $\mathbf{x}_n$ is observed.

3 Manifold Constrained Mixtures Models

Nonlinear data projection techniques [7,8] aim at finding the lower dimensional data manifold (if any) embedded in the input space and at unfolding it. A central concept is the geodesic distance, which is measured along the manifold and not through the embedding space, akin the Euclidean distance. The geodesic distance thus takes the intrinsic geometrical structure of the data into account.

Data Manifold. Consider two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ on the multidimensional manifold $\mathcal{M}$ of lower dimensionality than the one embedding space. The geodesic distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as the minimal arc length in $\mathcal{M}$ connecting both data samples. In practice, such a minimization is untractable. However, geodesic distances can easily be approximated by graph distances [9]. The problem of minimizing the arc length between two data samples lying on $\mathcal{M}$ reduces to the problem of minimizing the length of path (i.e. broken line) between these samples, while passing through a number of other data points of $\mathcal{M}$. In order to follow the manifold, only the smallest jumps between successive samples are permitted. This can be achieved by using either the K-rule, or the ϵ-rule. The former allows jumping to the K nearest neighbors. The latter allows jumping to samples lying inside a ball of radius ϵ centered on them. Below, we only consider the K-rule as the choice for ϵ is more difficult in practice.

The data and the set of allowed jumps constitute a weighted graph, the vertices being the data, the edges the allowed jumps and the edge labels the Euclidean distance between the corresponding vertices. In order to be a distance, the path length (i.e. the sum of successive jumps) must satisfy the properties of non-negativity, symmetry and triangular inequality. The first and the third are satisfied by construction. Symmetry is ensured when the graph is undirected. For the K-rule, this is gained by adding edges as follows: if $\mathbf{x}_j$ belongs to the K nearest neighbors of $\mathbf{x}_i$, but $\mathbf{x}_i$ is not a neighbor of $\mathbf{x}_j$ then the corresponding edge is added. Besides, extra edges are added to the graph in order to avoid disconnected parts. For this purpose a minimum spanning tree [10] is used.

The only remaining problem for constructing the distance matrix of the weighted undirected graph is to compute the shortest path between all data samples. This is done by repeatedly applying Dijkstra's algorithm [11], which computes the shortest path between a source vertex and all other vertices in a weighted graph provided the labels are non-negative (which is here the case).

Manifold Constrained VBE step. Let us denote the Euclidean and graph distances (i.e. approximate geodesic distances) between sample $\mathbf{x}_n$ and the component center $\boldsymbol{\mu}_m$ by $\delta^e(\mathbf{x}_n, \boldsymbol{\mu}_m)$ and $\delta^g(\mathbf{x}_n, \boldsymbol{\mu}_m)$ respectively. The exponential distribution $\mathcal{E}(y|\eta, \zeta) = \zeta^{-1} \exp\{-(y-\eta)/\zeta\}$ is suitable to measure the discrepancy between both distances by setting η to $\delta^e(\mathbf{x}_n, \boldsymbol{\mu}_m)^2$ and y to $\delta^g(\mathbf{x}_n, \boldsymbol{\mu}_m)^2$, since $\delta^e(\mathbf{x}_n, \boldsymbol{\mu}_m) \leq \delta^g(\mathbf{x}_n, \boldsymbol{\mu}_m)$. The manifold constrained responsibilities are obtained by penalizing the complete likelihood by the resulting discrepancy:

$$\begin{aligned} q'(z_{nm}) &\propto \exp\left(\mathrm{E}_{\boldsymbol{\theta}_m}\{\log p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta}_m, \mathcal{H}_M)\mathcal{E}(\delta^g(\mathbf{x}_n, \boldsymbol{\mu}_m)^2|\delta^e(\mathbf{x}_n, \boldsymbol{\mu}_m)^2, \zeta = 1)\}\right) \\ &\approx q(z_{nm})\ \exp\left(\delta^e(\mathbf{x}_n, \boldsymbol{\alpha}_m)^2 - \delta^g(\mathbf{x}_n, \boldsymbol{\alpha}_m)^2\right) \ , \end{aligned} \tag{8}$$

where it is assumed that the variance of $\boldsymbol{\mu}_m$ is small and $\boldsymbol{\alpha}_m = \mathrm{E}_{\boldsymbol{\theta}_m}\{\boldsymbol{\mu}_m\}$. Choosing ζ equal to 1 leaves the responsibility unchanged if both distances are identical. However, when the mismatch increases, $q'(z_{nm})$ decreases, which means that it is less likely that $\mathbf{x}_n$ was generated by m because the corresponding geodesic distance is large compared to the Euclidean distance. This results in a weaker responsibility, reducing the influence of $\mathbf{x}_n$ when updating the variational posterior of the parameters of m in the VBM step.

4 Manifold Constrained Variational Gaussian Mixtures

In this section, the manifold constrained variational Bayes machinery is applied to the Gaussian mixture case. A finite Gaussian mixture (FGM) [1] is a linear combination of M multivariate Gaussian distributions with means $\{\boldsymbol{\mu}_m\}_{m=1}^M$ and covariance matrices $\{\boldsymbol{\Sigma}_m\}_{m=1}^M$: $\hat{p}(\mathbf{x}) = \sum_{m=1}^M \pi_m \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$, with $\mathbf{x} \in \mathbb{R}^d$. The mixing proportions $\{\pi_m\}_{m=1}^M$ are non-negative and must sum to one. Their conjugate prior is a Dirichlet $p(\pi_1, ..., \pi_M) = \mathcal{D}(\pi_1, ..., \pi_M|\kappa_0)$ and the conjugate prior on the means and the covariance matrices is a product of Normal-Wisharts $p(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = \mathcal{N}(\boldsymbol{\mu}_m|\boldsymbol{\alpha}_0, \boldsymbol{\Sigma}_m/\beta_0)\mathcal{W}\left(\boldsymbol{\Sigma}_m^{-1}|\gamma_0, \boldsymbol{\Lambda}_0\right)$. The variational posterior factorizes similarly as the prior and is of the same functional form. The posterior on the mixture proportions $q(\pi_1, ..., \pi_M)$ are jointly Dirichlet $\mathcal{D}(\pi_1, ..., \pi_M|\kappa_1, ..., \kappa_m)$ and the posterior on the means and the covariance matrices $q(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ are Normal-Wishart $\mathcal{N}(\boldsymbol{\mu}_m|\boldsymbol{\alpha}_m, \boldsymbol{\Sigma}_m/\beta_m)\mathcal{W}\left(\boldsymbol{\Sigma}_m^{-1}|\gamma_m, \Lambda_m\right)$.

Training Procedure. The parameters of manifold constrained variational Gaussian mixtures (VB-MFGM) can be learnt as follows:

1. Construct the training manifold by the K-rule and compute the associated distance matrix $\delta^g(\mathbf{x}_i, \mathbf{x}_j)$ by Dijkstra's shortest path algorithm.
2. Repeat until convergence:
   Update the distance matrix of the expected component means.
   Find for each $\boldsymbol{\alpha}_m$ the K nearest training samples $\{\mathbf{x}_k\}_{k=1}^K$ and compute its graph distances to all training data: $\delta^g(\mathbf{x}_n, \boldsymbol{\alpha}_m) = \min_k\{\delta^g(\mathbf{x}_n, \mathbf{x}_k) + \delta^e(\mathbf{x}_k, \boldsymbol{\alpha}_m)\}$.
   VBE step. Compute the manifold constrained responsibilities using (8):

$$\begin{aligned} q'(z_{nm}) \propto \tilde{\pi}_m \tilde{\Lambda}_m^{1/2} \exp\left\{-\frac{\gamma_m}{2}\left(\mathbf{x}_n - \boldsymbol{\alpha}_m\right)^{\mathrm{T}} \boldsymbol{\Lambda}_m \left(\mathbf{x}_n - \boldsymbol{\alpha}_m\right) - \frac{d}{2\beta_m}\right\} \\ \times \exp\left\{\delta^e(\mathbf{x}_n, \boldsymbol{\alpha}_m)^2 - \delta^g(\mathbf{x}_n, \boldsymbol{\alpha}_m)^2\right\} \ , \end{aligned} \tag{9}$$

where $\log \tilde{\pi}_m \equiv \psi(\kappa_m) - \psi\left(\sum_m \kappa_m\right)$ and $\log \tilde{\Lambda}_m \equiv \sum_{i=1}^{d} \psi\left(\frac{\gamma_m+1-i}{2}\right) + d \log 2 - \log |\boldsymbol{\Lambda}_m|$, with $\psi(\cdot)$ the digamma function.

VBM step. Update the variational posteriors by first computing the following quantities:

$$\bar{\boldsymbol{\mu}}_m = \frac{\sum_n q'(z_{nm}) \mathbf{x}_n}{\sum_n q'(z_{nm})} \ , \bar{\boldsymbol{\Sigma}}_m = \frac{\sum_n q'(z_{nm}) C(\mathbf{x}_n, \bar{\boldsymbol{\mu}}_m)}{\sum_n q'(z_{nm})} \ , \bar{\pi}_m = \frac{\sum_n q'(z_{nm})}{N} \ ,$$

where $C(\mathbf{x}_n, \bar{\boldsymbol{\mu}}_m) = (\mathbf{x}_n - \bar{\boldsymbol{\mu}}_m)(\mathbf{x}_n - \bar{\boldsymbol{\mu}}_m)^{\mathrm{T}}$. Next, update the parameters of the posteriors:

$$\boldsymbol{\alpha}_m = \frac{N \bar{\pi}_m \bar{\boldsymbol{\mu}}_m + \beta_0 \boldsymbol{\alpha}_0}{\beta_m} \ , \quad \beta_m = N \bar{\pi}_m + \beta_0 \ , \quad \gamma_m = N \bar{\pi}_m + \gamma_0 \ , \tag{10}$$

$$\boldsymbol{\Lambda}_m^{-1} = N \bar{\pi}_m \bar{\boldsymbol{\Sigma}}_m + \frac{N \bar{\pi}_m \beta_0 C(\bar{\boldsymbol{\mu}}_m, \boldsymbol{\alpha}_0)}{\beta_m} + \boldsymbol{\Lambda}_0^{-1} \ , \quad \kappa_m = N \bar{\pi}_m + \kappa_0 \ . \tag{11}$$

The computational overhead at each iteration step is limited with respect to standard VB-FGM, as the number of components in the mixture is usually small and updating $\delta^g(\mathbf{x}_n, \boldsymbol{\alpha}_m)$ does not require to recompute $\delta^e(\mathbf{x}_i, \mathbf{x}_j)$.

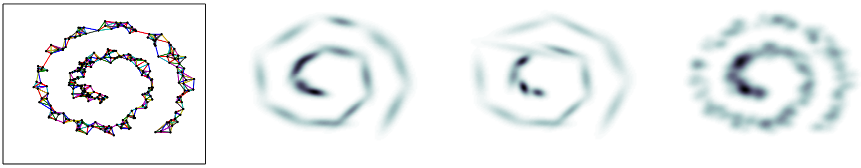
5 Experimental Results and Conclusion

In order to asses the quality of the density estimators the average negative log-likelihood of the test set $\{\mathbf{x}_q\}_{q=1}^{N_t}$ is used: ANLL $= -\frac{1}{N_t} \sum_{q=1}^{N_t} \log \hat{p}(\mathbf{x}_q)$. In the following, VB-MFGM is compared to standard VB-FGM and standard nonparametric density estimation (Parzen) [12] on artificial and real data.

The first example is presented for illustrative purposes. The data samples are generated from a two dimensional noisy spiral: $\mathbf{x}_1 = 0.04t \sin t + \epsilon_1$ and $\mathbf{x}_2 = 0.04t \cos t + \epsilon_2$, where t follows a uniform $\mathcal{U}(3, 15)$ and $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0, .03)$ is zero-mean Gaussian noise in each direction. The training, validation and test sets have respectively 300, 300 and 10000 samples. The optimal parameters are $M = 15$ and $K = 5$. The estimators are shown in Figure 1. On the one hand, VB-MFGM avoids manifold related local minima in which standard VB-FGM may get trapped into by forcing the expected component centers to move through the training manifold and the covariance matrices to be oriented along it. On the other hand, VB-MFGM clearly produces smoother estimators than Parzen.

In order to asses the performance of VB-MFGM on a real data set, the density of the Abalone[2] data is estimated after normalization. Note that the information regarding the sex is not used. The available data is divided in 2500 training, 500 validation, and 1177 test points. The optimal parameters are $M = 7$ and $K = 20$. The optimal width of the Gaussian kernel in Parzen is 0.17. The ANLL of test set for Parzen windows, VB-FGM and VB-MFGM are respectively 2.49, 0.84 and <u>0.37</u>. Remark that the improvement of VB-MFGM compared to VB-FGM is statistically significant (the standard error of the ANLL is 0.025).

[2] The Abalone data is available from the UCI Machine Learning repository: htttp://www.ics.uci.edu/ mlearn/MLRepository.html.



(a) Learn. Manif. (b) VB-MFGM (-.50). (c) VB-FGM (-.45). (d) Parzen (-.48)

Fig. 1. Training manifold of a noisy spiral, as well as the VB-MFGM, the standard VB-FGM and the Parzen window estimator. For each one, the ANLL of the test set is between parentheses (and the best is underlined).

Conclusion. The knowledge that the data is lying on a lower dimensional manifold than the dimension of the embedding space can be exploited in the frame of variational mixtures. By penalizing the complete data likelihood, the responsibilities (VBE step) are biased according to a discrepancy between the Euclidean and the geodesic distance. Following this methodology, manifold constrained variational Gaussian mixtures (VB-MFGM) were constructed. It was demonstrated experimentally that the resulting estimators are superior to standard variational approaches and nonparametric density estimation. In the future, we plan to investigate alternative mismatch measures.

References

1. McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York, NY. (2000)
2. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Stat. Soc., B **39** (1977) 1–38.
3. Attias, H.: A variational bayesian framework for graphical models. In Solla, S., Leen, T., Mller, K.R., eds.: NIPS 12. MIT Press. (1999)
4. Corduneanu, A., Bishop, C.M.: Variational bayesian model selection for mixture distributions. In Jaakkola, T., Richardson, T., eds.: AISTATS 8, Morgan Kaufmann (2001) 27–34.
5. Vincent, P., Bengio, Y.: Manifold Parzen windows. In S. Becker, S.T., Obermayer, K., eds.: NIPS 15. MIT Press (2003) 825–832.
6. Beal, M.J.: Variational Algorithms for Approximate Bayesian Inference. PhD thesis, University College London (UK). (2003)
7. Lee, J.A., Lendasse, A., Verleysen, M.: Nonlinear projection with curvilinear distances: Isomap versus Curvilinear Distance Analysis. Neucom **57** (2003) 49–76.
8. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290** (2000) 2319–2323.
9. Bernstein, M., de Silva, V., Langford, J., Tenenbaum, J.: Graph approximations to geodesics on embedded manifolds. Techn. report Stanford University, CA. (2000)
10. West, D.B.: Introduction to Graph Theory. Prentice Hall, Upper Saddle River, NJ. (1996)
11. Dijkstra, E.W.: A note on two problems in connection with graphs. Num. Math. **1** (1959) 269–271.
12. Parzen, E.: On estimation of a probability density function and mode. Ann. Math. Stat. **33** (1962) 1065–1076.