Max-Margin Learning of Gaussian Mixtures with Sequential Minimal Optimization

Trinh Minh Tri Do
LIP6, Universite Pierre et
Marie Curie, Paris, France
Trinh-Minh-Tri.Do@lip6.fr

Thierry Artieres
LIP6, Universite Pierre et
Marie Curie, Paris, France
Thierry.Artieres@lip6.fr

Abstract

This works deals with discriminant training of Gaussian Mixture Models through margin maximization. We go one step further previous work, we propose a new formulation of the learning problem that allows the use of efficient optimization algorithm popularized for Support Vector Machines, yielding improved convergence properties and recognition accuracy on handwritten digits recognition.

1 Introduction

A Gaussian Mixture Model (GMM) is a generative probabilistic model that implements a class conditional density function of the form:

$$p(x|y) = \sum_{k=1}^{K} p(k) \times p(x|N_{y,k}) \tag{1}$$

where y denotes the class of the sample x, $p(k)$ denotes the prior probabilities of the k^{th} components, and $N_{y,k}$ stands for the k^{th} component of the mixture, which is a Gaussian distribution with mean $\mu_{y,k}$ and a covariance matrice $\Sigma_{y,k}$:

$$p(x|N_{yk}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_{yk}|}} exp \left[-\frac{1}{2}(x-\mu_{yk})\Sigma_{yk}^{-1}(x-\mu_{yk}) \right] \tag{2}$$

GMMs are used in many areas such as speech or image processing and recognition. GMMs owe part of their popularity first to central theorem that gives Gaussian distribution a unique status, and second to their generic feature. Indeed, one can implement any pdf that respect some regularity properties with a Gaussian mixture. Also, GMMs have been shown to be rather efficient, robust, and easy to use. Finally, GMMs inherit the popularity of Hidden Markov Models (HMMs) that have been intensively used for years for sequence classification and segmentation in speech, off-line and on-line handwriting recognition.

GMMs are traditionally learned with a non discriminant criterion. One GMM model is learned for each class, y, with positive-only samples (e.g. via Maximum Likelihood Estimation (MLE) [6, 15]) to implement probability density functions (pdf) $p(x|y)$. Then one builds a Bayes decision rules through $\text{argmax}_y\, p(x|y) * p(y)$ (often assuming uniform priors). This generative approach is usually less efficient (e.g. wrt. error-rate) than discriminant methods. However generative approaches have been extensively used for dealing with sequences and more generally with structured data because of the difficulty to learn discriminant models for such data. Hence, many speech and handwriting recognizers are based on non discriminant learning schemes.

Some works investigated the discriminant learning of models for sequence classification, less have focused on sequence segmentation. First studies consisted in introducing discriminant criteria (for GMM or HMMs) and in using gradient descent like algorithms for optimizing the chosen criterion. One can mention criteria such as Class Conditional Likelihood [14], Maximum Mutual Information (MMI) [2, 16, 5, 22] or Minimum Classification Error (MCE) [8], see [11] for a synthetic review of these methods. Also [21] proposed to learn a classifier that minimizes the probability of error of generative models. Other techniques have been proposed for building discriminant systems based on generative models, like the use of Fisher scores [7], or the use of kernels between models [13].

Lastly, in the last few years a number of studies have focused on mixing Markovian models, exploiting Gaussian distributions, and Support Vector Machines (SVM) based discriminant algotihms [23, 9, 12, 19, 20]. For instance the technique in [19, 20] consists in learning GMM with a maximum margin criterion. These latter works are very promising but suffer from some limitations. A first limit comes form the nature of the parameters learned (e.g. only Gaussian means are learned in [12]). A second limit comes from the convergence rate which may be very slow, hence does not yield a good solution.

Here we build on the work of [19, 20] and aim at learning efficiently generative models through maximization of the margin. We focus on GMMs and propose a learning algorithm that differs from the one in [19, 20] from several aspects. One major component of our work is a new formulation of the learning problem as an optimization problem which allows the use of SMO-like algorithms (Sequential Minimal Optimization) [17, 4, 1]. Thanks to this formulation we propose a more robust algorithm (wrt. initialization) with faster convergence and better experimental recognition results.

We present first the original algorithm proposed in [19, 20] in Section 2. In Section 3 we detail our approach by reformulating the problem in its dual form and by showing how to instantiate the SMO algorithm in this particular case. Finally, we provide experimental results on off-line handwritten digit recognition in Section 4.

2 Discriminant Learning of GMMs

We consider here the discrimination of d-dimensional samples (feature vectors) $x = [x^1, x^2, .., x^d]$. We want to build a GMM-based classifier from a labeled training data set $(x_1, y1), .., (x_N, y_N)$. We describe briefly the standard non discriminant training algorithm, then we present the approach proposed in [19, 20].

2.1 Maximum Likelihood learning

Assuming a known distribution law (e.g. Gaussian), the non discriminant approach consists in estimating class-conditional density probabilities $p(x|y)$ independently for each class y using a MLE. Classification of a test sample is performed by computing class posterior probabilities (or joint probability $P(x, y)$) and choosing the class y that maximizes this probability.

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, p(x|y) \times p(y) \tag{3}$$

where $p(y)$ stands for the prior probability of class y. Class conditional probability densities $p(x|y)$ are mixtures of Gaussian distribution (Cf. Eq. (1)). MLE learning a GMM for class y consists in searching parameters maximizing the likelihood of the training samples of that class, this is performed with an EM-like algorithm. In practice some numerical problems may arise, especially for high dimensional data, since obtained covariance matrices are not always well conditioned. A well-known solution is to use a regularization term. For instance one can smooth, at every iteration, the covariance matrices according to:

$$\Sigma_{y,k} = \Sigma_{y,k} + \lambda Id \tag{4}$$

where λ is usually chosen as a function of the diagonal terms in $\Sigma_{y,k}$ of of its condition number. We will refer to this method as EM in the following. Note that we will consider the following alternative decision rule where a test sample is affected to the class whose component maximizes the component posterior probability.

$$\hat{y} = \underset{y}{\operatorname{argmax}} \left[\max_k p(x|N_{y,k}) \times p(k|y) \times p(y) \right] \tag{5}$$

2.2 GMM learning with margin maximization

Let consider first the case where $K = 1$, i.e. class-conditional densities are single Gaussian distributions. [19, 20] proposed to express the decision rule as a discriminant function based on Mahalanobis distance. Consider the matrix Φ_y that is built from the parameters of the Gaussian distribution for class y, namely its mean μ_y, and its inverse covariance matrix $\psi_y = \Sigma_y^{-1}$:

$$\Phi_y = \begin{bmatrix} \psi_y & -\psi_y \mu_y' \\ -\mu_y' \psi_y & \mu_y \psi_y \mu_y' + \beta_y \end{bmatrix} \tag{6}$$

where β_y is a real value corresponding to the $log(p(y))$. Note that ψ_y being an (inverse) covariance matrix, it is positive semi-definite (PSD hereafter). Noting $z = [x, 1] = [x^1, x^2, .., x^d, 1]$ an extended sample, the discriminant function takes the form:

$$\hat{y} = argmin_y z \Phi_y z' \tag{7}$$

Then, a sample z_i with label y_i is well classified if $z_i \Phi_{y_i} z_i' < z_i \Phi_y z_i' \forall y \neq y_i$. Hence, for a training dataset $(x_1, y_1), .., (x_N, y_N)$ one can learn Φ_y by solving the following optimization problem:

$$\begin{array}{lll} \min_{\Phi,\xi} & \frac{1}{2}\sum_y \|\psi_y\|^2 + C\sum_i \xi_i & \\ s.c. & z_i \Phi_{y_i} z_i' \leq z_i \Phi_y z_i' - 1 + \xi_i & \forall i \forall y \neq y_i \\ & \xi_i \geq 0 & \forall i \\ & \psi_y \succ 0 & \forall y \end{array} \tag{8}$$

where $\psi_y \succ 0$ means that matrix ψ_y is PSD. In the above formulation, usual slack variables are used to deal with the unseparable case. Note that the regularization term in the above criterion concerns a subpart of matrix Φ_y only since it does appear to be relevant to regularize means of the distributions.

The above formalization is interesting since the objective function is quadratic with constraints that are either linear or convex. This opens possibilities to the use of efficient optimization techniques such as the ones developed in the field of Support Vector Machines (SVM). To extend the formulation to the case of mixtures of K Gaussian distributions per class, we introduce additional notations. Let y_i be the class of training sample x_i and k_i the index of the component that produced the sample, hence $r_i = (k_i, y_i)$ is a unique identifier of the Gaussian that is $responsible$ for x_i. For a sample x we note $r = (k, y)$ the identifier of the Gaussian which emitted it. Also, we will note $R(y)$ the set of all component distributions in the class conditional density for class y. Note that in the particular case where all r_i variables are known the learning problem becomes:

$$\begin{array}{lll} \min_{\Phi,\xi} & \frac{1}{2}\sum_r \|\psi_r\|^2 + C\sum_i \xi_i & \\ s.c & z_i\Phi_{r_i}z_i' \leq z_i\Phi_r z_i' - 1 + \xi_i & \forall i \forall r \notin R(y_i) \\ & \xi_i \geq 0 & \forall i \\ & \psi_r \succ 0 & \forall r \end{array} \tag{9}$$

[19, 20] propose to remove slack variables in Equation (9) by introducing the $hinge$ function, where $hinge(z) = max(0, z)$. This changes the problem into:

$$\begin{array}{ll} \min_{\Phi} & \frac{1}{2}\sum_r \|\psi_r\|^2 \\ & +C\sum_{i=1:N}\sum_{r\notin R(y_i)} hinge(1 + z_i\Phi_{r_i}z_i' - z_i\Phi_r z_i') \\ s.c & \psi_r \succ 0\,\forall r \end{array} \tag{10}$$

The objective function is convex and so are the constraints, which can be solved with e.g. a projected gradient descent technique [3, 18]. Every parameter update step is followed by a projection step. If the constraints are not satisfied parameters are projected in the subspace of the parameter space where constraints are satisfied. In the particular case of PSD of covariance matrices the projection step consists in finding the closest matrice to ψ_r that is PSD, which may be not so simple. [19, 20] proposed to set all negative eigen values of ψ_r to zero. This is a simple but rough method that makes the convergence relatively slow, hence much sensitive to initialization as suggested by the authors. In practice, they use as inialization the solution of MLE training.

3 Max-margin learning of GMMs in the dual

We just showed that PSD (non linear) constraints on ψ_r prevent the use of the dual form of the optimization problem and lead to slow and inefficient optimization algorithms. We propose here an improved alternative formulation allowing more efficient optimization algorithms.

3.1 Dual formulation

Noting that $M \succ 0 \iff \forall x, xMx \geq 0$ we propose to replace the PSD constraint $M \succ 0$ by a set of constraints of the form $xMx \geq 0$. For instance, considering these constraints for all training samples Eq. (9) becomes:

$$\begin{array}{lcl} \min_{\Phi,\xi} & \frac{1}{2}\sum_r \|\psi_r\|^2 + C\sum_i \xi_i & \\ \text{s.c} & z_i\Phi_{r_i}z_i' \leq z_i\Phi_r z_i' - 1 + \xi_i & \forall i \forall r \notin R(y_i) \\ & \xi_i \geq 0 & \forall i \\ & (x_i - \mu_t)\psi_r(x_i - \mu_t)' \geq 0 & \forall i, \forall r \end{array} \tag{11}$$

where μ_t denotes the mean of all training samples (it is not considered as a variable in the following). Of course, satisfying all the constraints $(x - \mu_t)\psi_r(x - \mu_t) \geq 0$ for the whole training set does not warranty that ψ_r is PSD. The idea behind our proposition is that we expect that if the constraint $(x - \mu_t)\psi_r(x - \mu_t) \geq 0$ is satisfied for all training samples then the matrice ψ_r should be PSD. Although we did not demonstrate such a result we did not encounter a counter-example in our experiments. To improve the clarity of the presentation we introduce temporary variables θ_i as in [1], Eq. (11) becomes:

$$\begin{array}{lcl} \min_{\Phi,\xi,\theta} & \frac{1}{2}\sum_r \|\psi_r\|^2 + C\sum_i \xi_i & \\ \text{s.c} & z_i\Phi_{r_i}z_i' \leq \theta_i - 1 + \xi_i & \forall i \\ & \theta_i \leq z_i\Phi_r z_i' & \forall i \forall r \notin R(y_i) \\ & \xi_i \geq 0 & \forall i \\ & (x_i - \mu_t)\psi_r(x_i - \mu_t)' \geq 0 & \forall i, \forall r \end{array} \tag{12}$$

Following standard derivation, this optimization problem can be solved by writing the Lagrangian then noticing that the solution is given by a saddle point of the Lagrangian, that must be minimized wrt. parameters Φ,ξ,θ and maximized wrt. Lagrange multipliers. Omitting details, one can get the dual form:

$$\begin{array}{lll} \max_{\alpha,\gamma} & -\frac{1}{2}\sum_r \|\psi_r\|^2 + \sum_i \alpha_i^{r_i} & \\ s.c & \alpha_i^r > 0, \gamma_i^r > 0, \alpha_i^{r_i} < C & \forall i \forall r \\ & \sum_r y_i^r \alpha_i^r = 0 & \forall i \\ & \sum_i y_i^r \alpha_i^r = 0 & \forall r \\ & \sum_i y_i^r \alpha_i^r x_i = 0 & \end{array} \tag{13}$$

3.2 Optimization with SMO

To optimize efficiently (13) we looked at decomposing it in smaller problems, this is the SMO strategy.

3.2.1 Principle

The principle of SMO is to (iteratively) select a training sample and then to optimize the objective function w.r.t. the variables associated to the selected sample [17, 4, 1]. The optimization step is performed through the iteration of minimal optimization steps, each concerns a pair only of variables that are linked through a constraint. These elementary optimization steps should be solved analytically. This algorithm relies on heuristics for the choice of the sample, and for the choice of the two variables to be optimized in a single elementary step. In our implementation, we first roughly evaluate the expected gain for every training sample (by examining what happens for a particular pair of variables). Also, to avoid a costful procedure to evaluate the most interesting pair of variables we rely on KKT conditions to determine it efficiently (see [1]).

3.2.2 SMO for GMM learning

Applying SMO in our case is not straightforward because of the constraints $\sum_i y_i^r \alpha_i^r x_i = 0$ in Eq. (13). The problem lies in that these constraints actually link variables that are associated to all the training samples. Then, there (may) exist pairs of variables that cannot be optimized in a SMO step, i.e. one cannot change their values while still satisfying the constraints. Note that these constraints concern the last rows and columns of matrices Φ_r, i.e. quantities involved are $\mu_r \psi_r$ and $\mu_r \psi_r \mu_r' + \beta_r$. A solution is to consider these quantities as variables that we note Ξ_r. This makes sens since, provided ψ_r is invertible (i.e. strictly positive since it is already PSD), quantities $\nu_r = \mu_r \psi_r$ and $\nu_r = \mu_r \psi_r$ may be viewed as variables that are independent from ψ_r. This is not always true in practice but this leads to good convergence behavior.

Based on this discussion, we propose to overcome the difficulty of handling the PSD constraint by distinguishing between two sets of variables to be learned, the matrices ψ_r on the one hand and the remaining parameters Ξ_r on the other hand, and to optimize iteratively and alternatively these two sets of parameters. The convexity of the objective function and of the constraints (wrt. Φ_r) warranties the convergence toward the global optimum solution. Besides, the optimization wrt. Ξ_r is linear, hence simple. We come back now to the optimization wrt. matrices ψ_r and detail the SMO algorithm. First we express the primal optimization problem of Equation(11), while considering optimization wrt. ψ_r only (Ξ_r are assumed constant).

$$\begin{array}{lcc}
\min_{\psi,\xi,\theta,\delta} & \frac{1}{2}\sum_r \|\psi_r\|^2 + C\sum_i \xi_i & \\
s.c & x_i \psi_{r_i} x_i' - 2x_i \nu_{r_i} + \delta_{r_i} \leq \theta_i - 1 + \xi_i & \forall i \\
 & \theta_i \leq x_i \psi_r x_i' - 2x_i \nu_r + \delta_r \forall i \forall r & \notin R(y_i) \\
 & \xi_i \geq 0 & \forall i \\
 & (x_i - \mu_t)\psi_r (x_i - \mu_t)' \geq 0 & \forall i \forall r
\end{array} \tag{14}$$

The problem in Eq. (14) is still a quadratic program which we can get the dual form as previously, with only one equation corresponding to $\frac{\delta L}{\delta \psi_r} = 0$, leading to:

$$\psi_r = \sum_i \gamma_i^r (x_i - \mu_t)'(x_i - \mu_t) - \sum_i y_i^r a_i^r x_i' x_i \tag{15}$$

We then get the dual:

$$\begin{array}{lll}
\max_{\alpha,\gamma} & -\frac{1}{2}\sum_r \|\psi_r\|^2 + \sum_i \alpha_i^{r_i} & \\
 & +\sum_{i,r} \alpha_i^r y_i^r (-2x_i \nu_r + \delta_r) & \\
s.c & \alpha_i^r > 0, \gamma_i^r > 0, \alpha_i^{r_i} < C & \forall i \forall r \\
 & \sum_r y_i^r \alpha_i^r = 0 & \forall i \\
 & \sum_i y_i^r \alpha_i^r = 0 & \forall r
\end{array} \tag{16}$$

where $\psi_r = \sum_i \gamma_i^r (x_i - \mu)'(x_i - \mu) - \sum_i y_i^r a_i^r x_i' x_i$

3.2.3 Elementary step

The above problem (16) is ready for decomposition. We present now the elementary step for a training sample x_i, whose associated variables are α_i^r and γ_i^r, with associated constraints $\alpha_i^r > 0$, $\alpha_i^{r_i} \leq C$,$\gamma_i^r > 0$, and $\alpha_i^{r_i} = \sum_{r \notin R(y_i)} \alpha_i^r$. We discuss now the optimization of α_i^r then the optimization of γ_i^r.

Optimization of α_i^r consists in first selecting a pair of variables r_a et r_b (corresponding to two Gaussian components), then searching the new values for $\alpha_i^{r_a}$, $\alpha_i^{r_b}$ that maximize the dual, while still satisfying the constraint $\alpha_i^{r_i} = \sum_{r \notin R(y_i)} \alpha_i^r$. We distinguish two cases. In the first case, one of the Gaussian is the one associated to the training sample x_i, let assume $r_a = r_i$. Then, if we increase $\alpha_i^{r_a}$ by v, then we should increase similarly $\alpha_i^{r_b}$ which does not belong to $R(x_i)$. In that case optimization consists in determining optimal value v that maximizes the dual. This may be done analytically since the latter is a quadratic function of v, all other variables being fixed. In the second case, none of the two variables correspond to a Gaussian component in $R(y_i)$. Then, since $\sum_{r \notin R(y_i)} \alpha_i^r$ cannot change one has to remove value v to $\alpha_i^{r_a}$, if it is added to $\alpha_i^{r_b}$. Here again the dual is a quadratic function of v and the optimal value may be found analytically. Note that in any case if the optimal value $v*$ leads to the violation of a constraint such as ($\alpha_i^r \geq 0$,$\alpha_i^{r_i} \leq C$) then the closest value to $v*$ that satisfy the constraint is chosen. For instance if $\alpha_i^r + v* \leq C$ then we choose $v = v*$ else we choose $v = C - \alpha_i^{r_i}$.

Finally, optimization of variables γ_i^r does not come with any problem since no constraint link these variables. γ_i^r occur in ψ_r only (Cf. Eq. (16)) so that optimal changes of these variables may also be determined analytically.

4 Experiments

We report experimental results gained on two off-line handwritten digit recognition benchmark databases, USPS dataset [10] and MNIST[1] dataset. USPS dataset consists in 7291 training samples and 2007 test samples, each digit is a 16x16 image. MNIST dataset consists in 60000 traing samples and 10000 test samples, each digit is a 28x28 image. We used a standard preprocessing for the two datasets ([11]), consisting in a Principal Components Analysis (PCA) where we keep the 50 principal components.

Experiments aim at comparing results of discriminant learning with the two optimization methods, projected gradient with the hinge function as in [19, 20] and SMO algorithm in our case. In all the experiments, the discriminant training is systematically initialized with the result of non discriminant training (MLE estimation with an EM

[1] http://yann.lecun.com/exdb/mnist/index.html

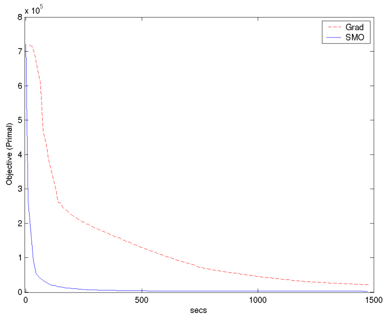

Figure 1. Comparison of convergence rates for the projected gradient based method in [19] (Grad) and our algorithm (SMO).

algorithm). A side effect of this initialization is the labeling of all training samples with the emitting Gaussian component (r_i).

We begin with the analysis of convergence for the two discriminant methods. Both algorithms are initialized with the MLE solution and are used with the same hyper parameter C. Figure 1 shows the evolution of objective function (value of the primal) as a function of time for the gradient method (Grad) and for our algorithm (SMO), for the USPS dataset. Absolute value of time doesn't matter here, what is actually interesting lies in the difference between the two methods. One may easily see on this figure that our technique based on SMO converges much faster than the projected gradient method, which takes much more time to converge to a good solution. While in theory the optimization problem in Equation (10) is convex and can be solved with a projected gradient method, it turns out that things are not so simple in practice. Convergence may be very slow so that reaching a good solution is not warrantied at all. Experimentally we observed that the solution of the Gradient method is much dependent on the initialization since numeric problems often prevent convergence to the optimal solution. From this point of view out method appears to be much more robust wrt. initialization.

Next figures (Figures (2) and (3) illustrate the sensitivity of both approaches by comparing the performances of the regularized non discriminant solution of Eq. (4) (EM) and of the two discriminant approaches (initialized with the EM solution), as a function of the regularization parameter λ (Cf Eq.(4)). Figure (2) shows the case of mixture models with two Gaussian component models per class (i.e. $K = 2$ in Eq.(1)) while Fig. (3) shows what happens with mixture models with four Gaussian components per class ($K = 4$). Whatever the EM initialization is (i.e. whatever λ), and in both cases ($K = 2$ or $K = 4$), our approach performs better that the reference method in [19]. Also our method appears to be less sensitive to initialization, which confirms convergence analysis results. This is an important point since a good initialization with EM usually comes with numeric problems and requires a careful regularized solution. With our method such a careful (i.e. manual) initialization is not mandatory so that one can use an automatically determined λ (i.e. non optimal EM solution) while still converging to the optimal maximum margin solution.

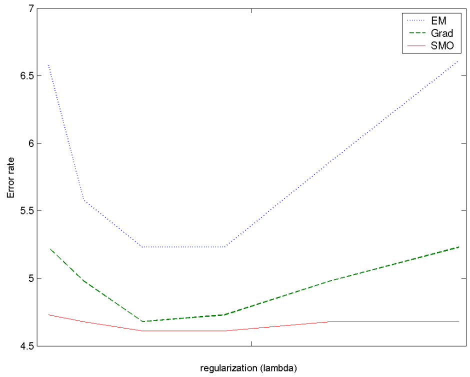

Figure 2. Comparison of the classification performances (on USPS dataset) for the regularized non discriminant solution (EM), algorithm [19] (Grad) and our algorithm (SMO) as a function of λ. Class models are mixtures of two Gaussian distributions.

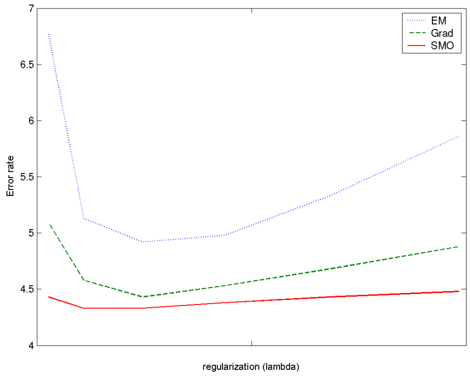

Figure 3. Comparison of the classification performances (on USPS dataset) for the regularized non discriminant solution (EM), algorithm in [19] (Grad) and our algorithm (SMO) as a function of λ. Class models are mixtures of four Gaussian distributions.

Finally we provide comparative results for the three methods on two datasets (USPS and MNIST) when K varies from 1 to 8 and for two typical cases of strong reg-

Table 1. Error rates for USPS digits, for regularized EM, [19] (Grad) and our algorithm (SMO), when λ is large (a) and small (b), for various K.

K	EM	Grad	SMO
1	7.22	5.23	4.88
2	6.61	5.23	4.68
4	5.86	4.88	4.48
6	5.46	4.73	4.43

(a)

K	EM	Grad	SMO
1	5.83	5.13	4.88
2	5.30	4.68	4.61
4	4.92	4.43	4.33
6	4.90	4.43	4.33

(b)

Table 2. Error rates for MNIST digits, for regularized EM, [19] (Grad) and our algorithm (SMO), when λ is large (a) and small (b), for various K.

K	EM	Grad	SMO
1	5.72	2.31	2.03
2	5.01	2.24	1.91
4	3.72	2.02	1.79
8	3.00	1.91	1.69

(a)

K	EM	Grad	SMO
1	3.93	2.10	2.03
2	3.48	2.05	1.90
4	2.65	1.99	1.79
8	2.07	1.78	1.69

(b)

ularization (large value of λ) and of light regularization (small value of λ). A first comment is that in any case our method performs similarly or better than the reference method, which confirms earlier results. Here again we see that our method is less sensitive to initialization whatever the number of components and whatever the dataset. Finally one note that the difference between our method and the gradient based one is less significant when regularization is light.

5 Conclusion

We proposed a new algorithm for learning GMMs through margin maximization. We proposed a new formulation of the learning problem that led us derive a new algorithm based on the SMO algorithm. Experiments show that our algorithm exhibits increased robustness and better convergence properties (wrt. convergence speed and quality) which translate into reduced error-rates on handwritten digit recognition for two benchmark datasets.

References

[1] F. Aiolli and A. Sperduti, "Multi-prototype Support Vector Machine", *IJCAI*, 2003, pp 541.

[2] L. Bahl, P. Brown, P. de Souza and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition", *ICASSP*, 1986.

[3] D. Bertsekas, *Nonlinear programming*, Athena Scientific, 2nd edition, 1999.

[4] K. Crammer and Y. Singer, "On the Learnability and Design of Output Codes for Multiclass Problems", *Machine Learning*, 47:201, 2002.

[5] J. Dahmen, R. Schluter and H. Ney, "Discriminative Training of Gaussian Mixtures for Image Object Recognition", *DAGM-Symposium*, 1999, pp 205–212.

[6] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society (Series B), 39:1-38*, 1977.

[7] T. Jaakkola, M. Diekhans and D. Haussler, "Using the Fisher kernel method to detect remote protein homologies", *International Conference on Intelligent Systems for Molecular Biology*, 1999.

[8] B.-H. Juang and S. Katagiri, "Discriminative Learning for Minimum Error Classification", *IEEE Trans. Acoustics, Speech, and Signal Processing*, 1992.

[9] S. E. Kruger, M. Schaffoner, M. Katz, E. Andelic and A. Wendemuth, "Mixture of Support Vector Machines for HMM based Speech Recognition", *Proceedings of the 18th International Conference on Pattern Recognition*, 2006.

[10] Y. LeCun, B. Boser, J. S. Denker, S. A. Solla, R. E. Howard and L. D. Jackel, "Back-propagation applied to handwritten zip code recognition", *Neural Computation*, 1:541–551, 1989.

[11] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-Based Learning Applied to Document Recognition", *Proceedings of the IEEE*, 86(11):2278–2324, Nov. 1998.

[12] X. Li, H. Jiang and C. Liu, "Large margin HMMs for speech recognition", *Proc. of ICASSP 2005*, 2005.

[13] P. J. Moreno, P. P. Ho and N. Vasconcelos, "A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications", S. Thrun, L. Saul and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, 2004, Cambridge, MA. MIT Press.

[14] A. Nadas, "A decision-theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood", *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1983, pp 31(4):814?817.

[15] R. Neal and G. Hinton, "A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants", M. I. Jordan, editor, *Learning in Graphical Models*, 1998. Kluwer.

[16] Y. Normandin, "Hidden Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Problem", *PhD dissertation, Dept. of Electrical Eng., McGill Univ., Montreal, Canada*, 1991.

[17] J. C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines", Technical report, Microsoft Research, 19981998 John Platt.

[18] N. Ratliff, J. A. Bagnell and M. Zinkevich, "Subgradient methods for maximum margin structured learning", *AISTATS 2007*, 2007.

[19] F. Sha and L. K. Saul, "Large margin Gaussian mixture modeling for phonetic classification and recognition", *Proc. of ICASSP 2006*, 2006.

[20] F. Sha and L. K. Saul, "Large margin hidden Markov models for automatic speech recognition", *Advances in Neural Information Processing Systems 19*, 2007.

[21] S. Tong and D. Koller, "Restricted Bayes Optimal Classifiers", *AAAI/IAAI*, 2000, pp 658.

[22] V. Valtchev, J. Odell, P. Woodland and S. Young., "MMIE training of large vocabulary recognition systems", *Speech Communication*, 1997, pp 22:303?314.

[23] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 2 edition, 1999.