Lin Xiao and Li Deng

A Geometric Perspective of Large-Margin Training of Gaussian Models

Large-margin techniques have been studied intensively by the machine learning community to balance the empirical error rate on the training set and the generalization ability on the test set. However, they have been mostly developed together with generic discriminative models such as support vector machines (SVMs) and are often difficult to apply in parameter estimation problems for generative models such as Gaussians and hidden Markov models. The difficulties lie in both the formulation of the training criteria and the development of efficient optimization algorithms.

In this article, we consider the basic problem of large margin training of Gaussian models. We take the geometric perspective of separating patterns using concentric ellipsoids, a concept that has not generally been familiar to signal processing researchers but which we will elaborate on here. We describe the approach of finding the maximum-ratio separating ellipsoids (MRSEs) [1] and derive an extension with soft margins. We show how to formulate the soft-margin MRSE problem as a convex optimization problem, more specifically a semidefinite program (SDP). In addition, we derive its duality theory and optimality conditions and apply this method to a vowel recognition example, which is a classical problem in signal processing.

RELEVANCE

A considerable number of recent works on automatic speech recognition [2]–[4] incorporate margins into the discriminative training of Gaussian hidden Markov models to improve their generalization capability. All of these works use convex optimization techniques, more specifically SDP, to formulate tractable training criteria and develop efficient solution methods. In particular, the method developed in [3] measures the separation margins in terms of Mahalanobis distances, which is closely related to the approach of MRSE. Similar formulations for large-margin training using ellipsoids have also been considered in, for example, [5]–[7].

The soft-margin MRSE problem captures the geometric essence of many large-margin ellipsoidal classifiers mentioned above and is a clear analogy of the classical linear SVM. In fact, it is equivalent to SVM using quadratic kernels with an additional positive semidefinite constraint that ensures the separating boundaries are ellipsoids, which is most appropriate for Gaussian models.

PREREQUISITES

We assume familiarity with basic signal processing tools such as Gaussian models and maximum-likelihood estimation. General familiarity with convex optimization is also required, in particular SDP and duality theory.

BACKGROUND

Consider the problem of classifying an unknown object into one of several presumed classes. We assume the observed data $x \in \mathbf{R}^d$ and its label $k \in \{1, \ldots, m\}$ are both random variables. Given the prior probability $p(k)$ and the conditional probability $p(x|k)$, the optimal classifier that minimizes classification error rate is based on the maximum a posterior (MAP) rule

$$\begin{aligned}\hat{k} &= \arg\max_k p(k|x) \\ &= \arg\max_k p(x,k) \\ &= \arg\max_k p(x|k)p(k). \end{aligned} \tag{1}$$

However, in practice, the probability functions $p(k)$ and $p(x|k)$ are usually unknown and need to be estimated from available training data. To make the estimation problem computationally tractable, we often adopt a parametric modeling approach where we assume the probability functions, in particular, the conditional probability $p(x|k)$, belong to a parametrized function family.

We focus on the most widely used family for probability distributions—Gaussian models. In this case, the conditional probability density functions are parametrized by the mean $\mu_k \in \mathbf{R}^d$ and the covariance matrix $\Sigma_k \in \mathbf{R}^{d\times d}$, for each class $k \in \{1, \ldots, m\}$. In the MAP decision rule (1), we replace the conditional probability $p(x|k)$ with

$$p(x|\,\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \times \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right).$$

Given a set of n-labeled training data $\{(x_i, c_i)\}_{i=1}^n$, and assuming that they are independent and identically distributed, the maximum-likelihood (ML) estimation of the model parameters is

$$\mu_k^{\mathrm{ML}} = \frac{1}{n_k}\sum_{i:\,c_i=k} x_i,$$
$$\Sigma_k^{\mathrm{ML}} = \frac{1}{n_k}\sum_{i:\,c_i=k}(x_i-\mu_k^{\mathrm{ML}})(x_i-\mu_k^{\mathrm{ML}})^T,$$

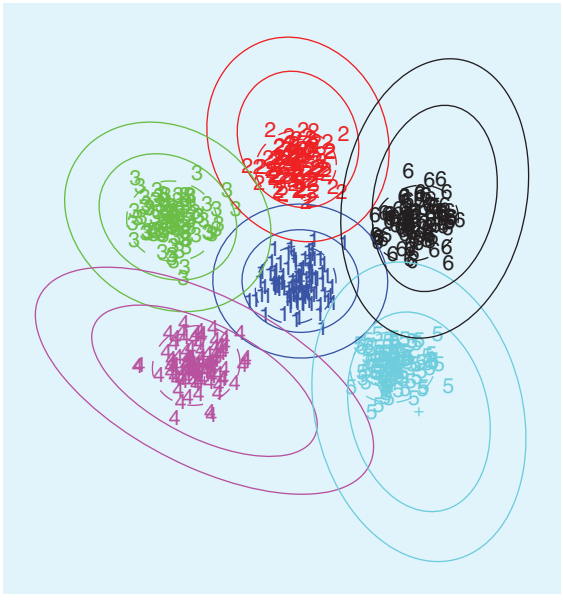
where $n_k = |\{i : c_i = k\}|$. The prior probabilities can be estimated as $p(k) = n_k/n$. In practical applications, ML estimation may not work well because of several reasons, including: model mismatch between data and assumptions and between training and testing data sets; and inadequate amount of training data.

In recent years, large-margin-based discriminative training has become a

Digital Object Identifier 10.1109/MSP.2010.938085

1053-5888/10/$26.00©2010IEEE

promising alternative for parameter estimation of generative models. Intuitively, a classifier with a larger margin (distance between the well-classified examples and the decision boundary) can better tolerate classification errors, which may arise due to insufficient training data, deficiency of the classifier, or mismatches between the training and test sets. It can be shown [8] that the test-set error rate is bounded by the sum of the empirical error rate on the training set and a generalization score associated with the margin. Minimizing this combined bound can lead to better performance on the test set than minimizing the empirical training-set error rate only.



[FIG1] MRSEs for separating six patterns.

For Gaussian models, the log-likelihood function is defined as

$$\begin{aligned}\mathcal{L}(x_i|\mu_k,\Sigma_k) &= \log(p(x_i|\mu_k,\Sigma_k)p(k))\\ &= -\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\\ &\quad -\frac{1}{2}\log|\Sigma_k| + \pi_k,\end{aligned}$$

where π_k is a constant depending on the prior probability $p(k)$. If we define the margin as the difference between the log-likelihoods of different model parameters, then the large-margin discriminative training problem can be stated as

$$\begin{aligned}&\text{maximize} \quad s\\ &\text{subject to}\\ &\mathcal{L}(x_i|\mu_{c_i},\Sigma_{c_i}) - \mathcal{L}(x_i|\mu_k,\Sigma_k) \geq s,\\ &\quad \forall\, k \neq c_i,\ i = 1,\ldots,n\\ &\quad \Sigma_k \succcurlyeq 0,\ k = 1,\ldots,m. \qquad (2)\end{aligned}$$

Here the optimization variables are the model parameters (μ_k, Σ_k) for $k = 1, \ldots, m$, and the scalar s, which is a lower bound on all the margins. The constraints $\Sigma_k \succcurlyeq 0$ state that all the covariance matrices must be positive semidefinite.

The large-margin training problem (2) is not convex, and in general is computationally very hard to solve. Several convex relaxations have been developed for this problem and its extensions for Gaussian hidden Markov models. In [2], the covariance matrices Σ_k are fixed and an SDP relaxation is developed for optimizing the means μ_k. In [4], an SDP relaxation is developed for simultaneous optimization of the mean and variances, assuming the covariance matrices are diagonal. In [3], the term $-(1/2)\log|\Sigma_k|$ in the log-likelihood function is ignored and an SDP relaxation is formulated by measuring the margins in terms of Mahalanobis distances. The MRSE problem discussed in this article has close connection with the last approach, and we take a new look at the problem providing a geometric perspective.

PROBLEM STATEMENT

The idea of separating multiclass patterns using ellipsoids is quite natural. For each class k, we would like to find an ellipsoid that encloses all labeled points x_i with $c_i = k$, while leaving all other points outside. In this context, the concept of separating margin is well served by the ratio between two concentric ellipsoids with the same shape, with all correctly labeled points enclosed in the inner ellipsoid and all other points excluded outside of the outer ellipsoid; see Figure 1 for an illustration.

An ellipsoid in $\mathbf{R}^d$ can be parametrized by its center μ and a symmetric, positive semidefinite matrix P that determines its shape and size

$$\begin{aligned}\mathcal{E}_d(\mu,P) = \{x \in \mathbf{R}^d \mid (x-\mu)^T\\ \times P(x-\mu) \leq 1\}.\end{aligned}$$

The ellipsoid is degenerate if P is singular. A scaled concentric ellipsoid with the same shape can be obtained by dividing the matrix P with a scalar $\rho > 0$. The scaled ellipsoid $\mathcal{E}_d(\mu, P/\rho)$ has an equivalent representation

$$\begin{aligned}\mathcal{E}_d(\mu,P/\rho) = \{x \in \mathbf{R}^d \mid (x-\mu)^T\\ \times P(x-\mu) \leq \rho\}.\end{aligned}$$

Note that $\sqrt{\rho}$ is the ratio between the lengths of the corresponding semimajor axes of the two concentric ellipsoids.

For a Gaussian random variable with mean μ and covariance matrix Σ, the ellipsoids $\mathcal{E}_d(\mu, \Sigma^{-1}/\rho)$ give confidence level sets with different probabilities, depending on the scaling ρ. For example, $\rho = d$ gives about 50% probability that a random sample is inside the ellipsoid $\mathcal{E}_d(\mu, \Sigma^{-1}/d)$, and $\rho = d + 2\sqrt{d}$ gives about 90% probability.

Suppose we are given the labeled training set $\{(x_i, c_i)\}_{i=1}^n$. For each class $k \in \{1, \ldots, m\}$, the associated MRSEs problem [1] is to find the ellipsoids $\mathcal{E}_d(\mu_k, P_k)$ and $\mathcal{E}_d(\mu_k, P_k/\rho_k)$ such that ρ_k is maximized while satisfying the constraints

$$\begin{aligned}&x_i \in \mathcal{E}_d(\mu_k, P_k),\ \ \forall i{:}c_i = k,\\ &x_i \notin \mathcal{E}_d(\mu_k, P_k/\rho_k),\ \ \forall i{:}c_i \neq k.\end{aligned}$$

In other words, the MRSE problem is

$$\begin{aligned}&\text{maximize} \quad \rho_k\\ &\text{subject to} \quad (x_i-\mu_k)^T P_k (x_i-\mu_k) \leq 1,\\ &\qquad\qquad \forall i : c_i = k\\ &\qquad\qquad (x_i-\mu_k)^T P_k (x_i-\mu_k) \geq \rho_k,\\ &\qquad\qquad \forall i : c_i \neq k\\ &\qquad\qquad P_k \succcurlyeq 0, \qquad (3)\end{aligned}$$

where the optimization variables are μ_k, P_k, and ρ_k. The MRSE problem (3) is always feasible. The patterns of class k are separable from all others using ellipsoids if and only if the optimal $\rho_k > 1$.

To gracefully handle the case when the patterns are not strictly separable using ellipsoids, we use the idea of soft margins as in SVM (see, e.g., [9]). In

other words, we introduce a slack variable ξ_i for each of the pattern inclusion or exclusion constraints, and add a weighted penalty term to the objective function as seen in (4) at the bottom of the page.

Here γ is a positive weighting parameter. We call the problem (4) the soft-margin MRSE problem. Note that we can always have the optimal $\rho_k > 1$ by setting a small enough γ. But on the other hand, the optimal ρ_k may be unbounded above if γ is too small.

Both the MRSE and soft-margin MRSE problems are nonconvex. However, they can be transformed into convex optimization problems, more specifically SDPs, using a homogeneous embedding technique [1], which we describe next.

SOLUTION

HOMOGENEOUS EMBEDDING

Any ellipsoid in $\mathbf{R}^d$ can be viewed as the intersection of a homogeneous ellipsoid (one centered at the origin) in $\mathbf{R}^{d+1}$ and the hyperplane

$$\mathcal{H} = \{z \in \mathbf{R}^{d+1} \,|\, z = (x, 1),\ x \in \mathbf{R}^d\}.$$

A homogeneous ellipsoid in $\mathbf{R}^{d+1}$ can be expressed as

$$\mathcal{E}_{d+1}(0, \Phi) = \{z \in \mathbf{R}^{d+1} \,|\, z^T \Phi z \le 1\},$$

where Φ is a symmetric positive semidefinite matrix. To find the intersection of $\mathcal{E}_{d+1}(0, \Phi)$ with the hyperplane $\mathcal{H}$, we partition the matrix Φ as

$$\Phi = \begin{bmatrix} P & q \\ q^T & r \end{bmatrix}, \tag{5}$$

where $P \in \mathbf{R}^{d\times d}$, $q \in \mathbf{R}^d$, and $r \in \mathbf{R}$. If $z = (x, 1)$, then we have

$$z^T \Phi z \le 1 \ \Leftrightarrow\ x^T P x + 2q^T x + r \le 1.$$

Now let

$$\mu = -P^{-1}q, \quad \delta = r - q^T P^{-1} q, \tag{6}$$

then

$$z^T \Phi z \le 1 \ \Leftrightarrow\ (x-\mu)^T P (x-\mu) + \delta \le 1.$$

Since P is positive semidefinite, we always have $\delta \le 1$. In addition, the constraint $\Psi \succcurlyeq 0$ implies that the Schur complement $r - q^T P^{-1} q \ge 0$, i.e., $\delta \ge 0$. Therefore, $0 \le \delta \le 1$. Whenever P is positive definite, we have $0 \le \delta < 1$, and

$$\mathcal{E}_{d+1}(0, \Phi) \cap H = \mathcal{E}_d(\mu, P/(1-\delta)).$$

In this case, we call $\mathcal{E}_{d+1}(0, \Phi)$ a homogeneous embedding of $\mathcal{E}_d(\mu, P/(1-\delta))$.

Given a nondegenerate ellipsoid $\mathcal{E}_d(\mu, P)$ (i.e., $P > 0$), its homogeneous embedding in $\mathbf{R}^{d+1}$ is nonunique, and can be parametrized as $\mathcal{E}_{d+1}(0, \Phi_\delta)$, for $0 \le \delta < 1$, where

$$\Phi_\delta = \begin{bmatrix} (1-\delta)P & -(1-\delta)P\mu \\ -(1-\delta)\mu^T P & (1-\delta)^T \mu^T P \mu + \delta \end{bmatrix}.$$

See Figure 2 for an illustration. The special case $\delta = 0$ is called a canonical embedding. In this case, the embedding $\mathcal{E}_{d+1}(0, \Phi_0)$ is a degenerate ellipsoid because the matrix Φ_0 is singular.

SDP FORMULATIONS

Now we consider the problem of separating patterns in $\mathbf{R}^d$ using homogeneous ellipsoids in $\mathbf{R}^{d+1}$. Since the MRSE approach essentially consists of solving m separate one-versus-others classification problems, we focus on a formulation for binary classification. For this purpose, we divide the n training data into two sets $\{x_i\}_{i=1}^{n_1}$ and $\{y_j\}_{j=1}^{n_2}$, where the x_i's are points belonging to a specific class, and y_j's are all the other points. We have $n_1 + n_2 = n$.

First, we need to embed the training data in the hyperplane $\mathcal{H}$ by letting

$$a_i = \begin{bmatrix} x_i \\ 1 \end{bmatrix},\ i = 1, \ldots, n_1,$$
$$b_j = \begin{bmatrix} y_j \\ 1 \end{bmatrix},\ j = 1, \ldots, n_2.$$

The MRSE problem using homogeneous embedding can be stated as

$$\begin{aligned} \text{maximize} \quad & \rho \\ \text{subject to} \quad & a_i^T \Phi a_i \le 1,\ i = 1, \ldots, n_1 \\ & b_j^T \Phi b_j \ge \rho,\ j = 1, \ldots, n_2 \\ & \Phi \succcurlyeq 0,\ \rho \ge 1, \end{aligned} \tag{7}$$

where the optimization variables are Φ and ρ. This is a convex optimization problem, more specifically an SDP. As a result, it can be solved globally and efficiently using interior-point methods; see, e.g., [10].

Once the problem (7) is solved, we can recover the parametrization in $\mathbf{R}^d$ using the transformations in (5) and (6). The two concentric separating ellipsoids are

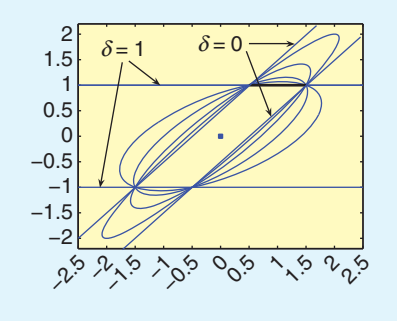
$$\mathcal{E}_d(\mu, P/(1-\delta)), \quad \mathcal{E}_d(\mu, P/(\rho(1-\delta))).$$

Moreover, the following properties of the MRSE problem are shown in [1]:

- If the patterns are separable, then the optimal solution to (7) is always a canonical embedding, i.e., $\delta = 0$.
- If the patterns are nonseparable, then the optimal solution to (7) always has $\rho = 1$, and Φ is degenerate such that $\delta = 1$.

Figure 3 gives a geometric illustration of the first property. The second one can be understood from the illustration presented in Figure 2.

$$\begin{aligned} \text{maximize} \quad & \rho_k - \gamma \textstyle\sum_i \xi_i \\ \text{subject to} \quad & (x_i - \mu_k)^T P_k (x_i - \mu_k) \le 1 + \xi_i, \quad \forall i : c_i = k \\ & (x_i - \mu_k)^T P_k (x_i - \mu_k) \ge \rho_k - \xi_i, \quad \forall i : c_i \ne k \\ & P_k \succcurlyeq 0,\ \xi_i \ge 0\ \ \forall i = 1, \ldots, n. \end{aligned} \tag{4}$$



[FIG2] Homogeneous embedding: the one-dimensional ellipsoid $4(x-1)^2 \le 1$ (i.e., the interval $1/2 \le x \le 3/2$) can be viewed as the intersection of a two dimensional ellipsoid (nonunique) with the hyperplane $\{(x, 1) \,|\, x \in R\}$.

Similarly, we can formulate the soft-margin MRSE problem with homogeneous embedding as an SDP

$$\begin{aligned}\text{maximize}\quad & \rho-\gamma\Big(\sum_i\xi_i+\sum_j\eta_j\Big)\\ \text{subject to}\quad & a_i^T\Phi a_i\le 1+\xi_i,\ i=1,\ldots,n_1\\ & b_j^T\Phi b_j\ge \rho-\eta_j,\ j=1,\ldots,n_2\\ & \Phi\succcurlyeq 0,\ \rho\ge 1\\ & \xi_i\ge 0,\ i=1,\ldots,n_1\\ & \eta_j\ge 0,\ j=1,\ldots,n_2,\end{aligned}\tag{8}$$

where the optimization variables are Φ, ρ, and the slack variables ξ_i and η_j. Here γ is a weighting parameter. Once the SDP (8) is solved, we can also use the transformations in (5) and (6) to find the separating ellipsoids in $\mathbf{R}^d$. The properties of the resulting solution will be discussed next, via duality theory and optimality conditions.

The same homogeneous embedding technique is used in [3] to formulate a large-margin training problem using Mahalanobis' distances as an SDP. We will explain its connection with the soft-margin MRSE problem later.

DUALITY

The dual of a convex optimization problem always gives more insights of the problem structure, and the MRSE problems are no exceptions. Here we only consider the duality theory for the soft-margin MRSE problem (8). Following standard derivation for Lagrange duality (see, e.g., [10, Ch. 5]), we found the dual of (8) is

$$\begin{aligned}\text{minimize}\quad & \textstyle\sum_i\lambda_i-\sum_j\nu_j+1\\ \text{subject to}\quad & \textstyle\sum_i\lambda_i a_i a_i^T-\sum_j\nu_j b_j b_j^T\succcurlyeq 0\\ & 0\le\lambda_i\le\gamma,\quad i=1,\ldots,n_1\\ & 0\le\nu_j\le\gamma,\quad j=1,\ldots,n_2\\ & \textstyle\sum_j\nu_j\ge 1,\end{aligned}\tag{9}$$

where the optimization variables are λ_i and ν_j, which are the Lagrange multipliers for the first two sets of constraints in (8), respectively. The dual problem (9) is also an SDP.

Weak duality always holds, i.e., the maximum objective value of the primal problem (8) is always no larger than the minimum value of the dual problem (9). If both the primal and dual problems are feasible, weak duality gives a simple bound on the maximum objective value of the primal problem, denoted by $\rho^\star$

$$p^\star\le\sum_i\lambda_i-\Big(\sum_j\nu_j-1\Big)\le\sum_i\lambda_i\le n_1\gamma.$$

If the primal problem is unbounded above, then the dual problem must be infeasible. This is the case if the parameter γ is set too small, for example, if $\gamma<1/n_2$. To see this, let Φ, ρ, $\{\xi_i\}_{i=1}^{n_1}$, and $\{\eta_j\}_{j=1}^{n_2}$ be a set of feasible solutions to the primal problem (8), and let

$$\rho'=\rho+\alpha,\ \eta_j'=\eta_j+\alpha,\ j=1,\ldots,n_2.$$

Then Φ, ρ', $\{\xi_i\}_{i=1}^{n_1}$ and $\{\eta_j'\}_{j=1}^{n_2}$ are still feasible for arbitrary $\alpha>0$, and the objective value is

$$\begin{aligned}&\rho'-\gamma\Big(\sum_i\xi_i+\sum_j\eta_j'\Big)\\ &\quad=\rho-\gamma\Big(\sum_i\xi_i+\sum_j\eta_j\Big)\\ &\qquad+(1-n_2\gamma)\alpha.\end{aligned}$$

Therefore, if $\gamma<1/n_2$, then the objective value increases unbounded above as we increase α to infinity. To check that the dual problem is infeasible when $\gamma<1/n_2$, we note that the constraints $0\le\nu_j\le\gamma$ imply that

$$\sum_j\nu_j\le n_2\gamma<n_2\frac{1}{n_2}=1.$$

But this contradicts with the last constraint in (9), which leads to dual infeasibility.

The most interesting case in practice is when both the primal and dual problems are strictly feasible. In this case, strong duality holds, i.e., the optimal values of the primal and dual problems are finite and the same. Moreover, the optimality conditions in this case reveal many interesting properties of the soft-margin MRSE problem. This is discussed next.

OPTIMALITY CONDITIONS

Let $\Phi^\star$, $\rho^\star$, $\{\xi_i^\star\}_{i=1}^{n_1}$ and $\{\eta_j^\star\}_{j=1}^{n_2}$ denote the primal optimal solutions, and $\{\lambda_i^\star\}_{i=1}^{n_1}$ and $\{\nu_j^\star\}_{j=1}^{n_2}$ denote the dual optimal solutions. In addition to be primal and dual feasible, they satisfy the following complementary slackness conditions:

- for the separation ratio,

$$\begin{aligned}&\rho^\star>1\ \Rightarrow\ \textstyle\sum_j\nu_j^\star=1,\\ &\textstyle\sum_j\nu_j^\star>1\ \Rightarrow\ \rho^\star=1;\end{aligned}$$

- for $i=1,\ldots,n_1$,

$$\begin{aligned}&a_i^T\Phi^\star a_i<1\ \Rightarrow\ \lambda_i^\star=0,\\ &\text{or}\ \ \lambda_i^\star>0\ \Rightarrow\ a_i^T\Phi^\star a_i=1+\xi_i^\star,\\ &\text{and}\\ &\xi_i^\star>0\ \Rightarrow\ \lambda_i^\star=\gamma,\\ &\text{or}\ \ \lambda_i^\star<\gamma\ \Rightarrow\ \xi_i^\star=0;\end{aligned}$$

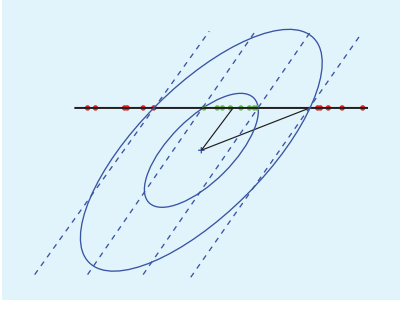
- for $j=1,\ldots,n_2$,

$$\begin{aligned}&b_j^T\Phi^\star b_j>\rho^\star\ \Rightarrow\ \nu_j^\star=0,\\ &\text{or}\ \ \nu_j^\star>0\ \Rightarrow\ b_j^T\Phi^\star b_j=\rho^\star-\eta_j^\star,\\ &\text{and}\\ &\eta_j^\star>0\ \Rightarrow\ \nu_j^\star=\gamma,\\ &\text{or}\ \ \nu_j^\star<\gamma\ \Rightarrow\ \eta_j^\star=0;\end{aligned}$$

- and finally, the matrix complementary slackness condition

$$\text{Tr}\,\Phi^\star\Big(\sum_i\lambda_i^\star a_i a_i^T-\sum_j\nu_j^\star b_j b_j^T\Big)=0.\tag{10}$$

The scalar complementary slackness conditions above parallel those for the SVM; see, e.g., [9]. In particular, the a_i's with $0<\lambda_i^\star<\gamma$ lie on the boundary of the inner ellipsoid, and the b_j's with $0<\nu_j^\star<\gamma$ lie on the boundary



[FIG3] The two homogeneous ellipsoids drawn in solid lines separate the two classes, and so do the two degenerate ellipsoids (canonical embeddings) drawn in dashed lines. They have the same intersections with the hyperplane $\mathcal{H}$, but the two canonical embeddings have the maximum separation ratio.

of the outer ellipsoid. The a_i's with $\xi_i^\star > 0$ (and therefore $\lambda_i^\star = \gamma$) are in-class points that lie outside of the inner ellipsoid, and the b_j's with $\eta_j^\star > 0$ (and therefore $\nu_j^\star = \gamma$) are out-of-class points that lie inside of the outer ellipsoid. All the points with either $\lambda_i^\star > 0$ or $\nu_j^\star > 0$ have the same interpretation of "support vectors" as in the SVM. All other points can be removed without affecting the optimal solution.

The matrix complementary condition (10) is equivalent to the balancing equation

$$\sum_i \lambda_i^\star (a_i^T \Phi^\star a_i) = \sum_j \nu_j^\star (b_j^T \Phi^\star b_j).$$

Since both $\Phi^\star$ and $\sum_i \lambda_i^\star a_i a_i^T - \sum_j \nu_j^\star b_j b_j^T$ are symmetric and positive semidefinite, (10) actually implies a stronger condition

$$\Phi^\star \left(\sum_i \lambda_i^\star a_i a_i^T - \sum_j \nu_j^\star b_j b_j^T \right) = 0. \quad (11)$$

To ease notation, let

$$\sum_i \lambda_i^\star a_i a_i^T - \sum_j \nu_j^\star b_j b_j^T = \begin{bmatrix} \Psi^\star & \theta^\star \\ \theta^{\star T} & \alpha^\star \end{bmatrix},$$

where

$$\Psi^\star = \sum_i \lambda_i^\star x_i x_i^T - \sum_j \nu_j^\star y_j y_j^T,$$
$$\theta^\star = \sum_i \lambda_i^\star x_i - \sum_j \nu_j^\star y_j,$$
$$\alpha^\star = \sum_i \lambda_i^\star - \sum_j \nu_j^\star.$$

Together with the partition of $\Phi^\star$ given in (5), the matrix (11) means

$$P^\star \Psi^\star + q^\star \theta^{\star T} = 0,$$
$$P^\star \theta^\star + \alpha^\star q^\star = 0,$$
$$q^{\star T} \Psi^\star + r^\star \theta^{\star T} = 0,$$
$$q^{\star T} \theta^\star + \alpha^\star r^\star = 0.$$

Many interesting properties of the optimal solutions follow from the above four equations.

First, note that we always have $\alpha^\star \geq 0$ and $P^\star \succcurlyeq 0$ (positive semidefinite). If $\alpha^\star > 0$ and $P^\star > 0$ (positive definite), then

- the center of the ellipsoids can be expressed as

$$\mu^\star = -P^{\star -1} q^\star = \frac{1}{\alpha^\star} \theta^\star = \frac{\sum_i \lambda_i^\star a_i - \sum_j \nu_j^\star b_j}{\sum_i \lambda_i^\star - \sum_j \nu_j^\star}$$

- $\Psi^\star$ is a rank-one matrix

$$\Psi^\star = \alpha^\star \mu^\star \mu^{\star T} = \frac{1}{\alpha^\star} \theta^\star \theta^{\star T}$$

and we always have canonical embedding, because

$$r^\star = -q^{\star T} \mu^\star = q^{\star T} P^{\star -1} q^\star$$
$$\Rightarrow \delta^\star = r^\star - q^{\star T} P^{\star -1} q^\star = 0.$$

If the patterns are nonseparable and the parameter γ is set too big, we could have $\alpha^\star = 0$. In this special case, the optimal values of the primal and dual problems are equal to one, since

$$\rho^\star - \gamma \left(\sum_i \xi_i^\star + \sum_j \eta_j^\star \right) = \sum_i \lambda_i^\star - \sum_j \nu_j^\star + 1 = 1.$$

This also implies $\rho^\star = 1$, $\xi_i^\star = 0$ for all i, and $\eta_j^\star = 0$ for all j, which are the same as the solution to (7) with $\delta^\star = 1$.

ALTERNATIVE FORMULATIONS

When solving the soft-margin MRSE problem, in very rare cases, we could have $\alpha^\star > 0$ and $P^\star$ being singular. This could happen, e.g., if some of the boundary points are aligned in an affine subspace. To fix this problem, we can add a regularization term that accounts for the volume of the ellipsoids to the objective function. Since $\log(\det(P^{-1}))$ directly measures the volume of the ellipsoid (which is infinite if P is singular), we can replace the objective of the problem (8) by

$$\text{maximize } \rho - \gamma_1 \left(\sum_i \xi_i + \sum_j \eta_j \right) - \gamma_2 \log(\det(P^{-1})),$$

where γ_1 and γ_2 are two positive weighting parameters. Note that P is the $d \times d$ leading diagonal block of Φ. The resulting problem is also a convex optimization problem, and can be solved globally and efficiently using interior-point method [11].

Instead of using the weighting parameter γ, we could also parametrize the soft-margin MRSE problem (8) using the ratio ρ. Given any fixed $\rho \geq 1$, for each class $k \in \{1, \ldots, m\}$, we can solve the problem

$$\begin{aligned} \text{minimize} \quad & \textstyle\sum_{i=1}^n \eta_{ik} \\ \text{subject to} \quad & z_i^T \Phi_k z_i \leq 1 + \eta_{ik}, \ \forall i : c_i = k \\ & z_i^T \Phi_k z_i \geq \rho - \eta_{ik}, \ \forall i : c_i \neq k \\ & \Phi_k \succcurlyeq 0, \\ & \eta_{ik} \geq 0, \ i = 1, \ldots, n, \end{aligned} \quad (12)$$

where $z_i = (x_i, 1)$, and the optimization variables are Φ_k and η_{ik}. This problem is equivalent to (8) for certain value of γ. Note that we have switched to notations that are explicit for multiclass problems, even though the m problems, each for a distinct class, can be solved separately.

Such notations are convenient for formulating a variant that allows simultaneous discriminative learning of all classes, similar to the large-margin problem (2). For this purpose, let's fix $\rho = 2$ and consider the m versions of (12), one for each class, altogether. The first two sets of constraints in these SDPs imply

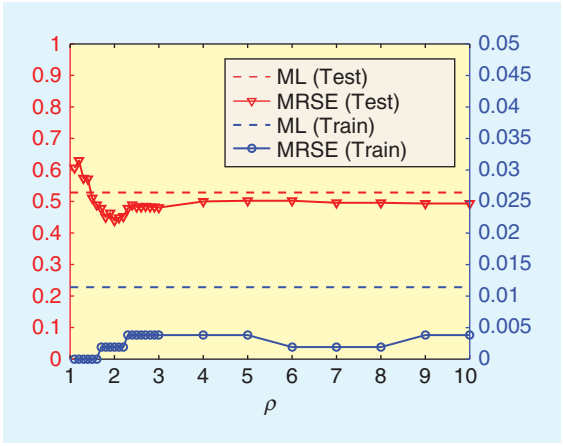
$$z_i^T (\Phi_k - \Phi_{c_i}) z_i \geq 1 - (\eta_{ic_i} + \eta_{ik}), \quad \forall k \neq c_i, \ i = 1, \ldots, n. \quad (13)$$

For convenience, let $\xi_{ik} = \eta_{ic_i} + \eta_{ik}$, and consider the problem of minimizing $\sum_{i,k} \xi_{ik}$, subject to all the constraints in (13). However, we notice that the variables Φ_k, $k = 1, \ldots, m$, can be scaled simultaneously to yield arbitrary large margins in (13). Therefore, we need to add a regularization term to the objective to limit the sizes of the Φ_k's. For example, regularizing the traces of the Φ_k's leads to the formulation

$$\begin{aligned} \text{minimize} \quad & \textstyle\sum_{i,k} \xi_{ik} + \gamma \sum_k \operatorname{Tr} \Phi_k \\ \text{subject to} \quad & z_i^T (\Phi_k - \Phi_{c_i}) z_i \geq 1 - \xi_{ik}, \\ & \xi_{ik} \geq 0, \forall k \neq c_i, i = 1, \ldots, n \\ & \Phi_k \succcurlyeq 0, \ k = 1, \ldots, m, \end{aligned} \quad (14)$$

where γ is a weighting parameter, and the optimization variables are Φ_k and

ξ_{ik}. This is almost the same problem considered in [3], the only difference is that they used the regularization $\mathrm{Tr}\, P_k$ instead of $\mathrm{Tr}\, \Phi_k$, where P_k is the $d \times d$ leading diagonal block of Φ_k. By examining the equations (5) and (6), we see that using $\mathrm{Tr}\, \Phi_k$ will favor a smaller δ, thus more likely to obtain canonical embeddings ($\delta = 0$). Other regularizations, such as logdet P, are also interesting to consider.



[FIG4] **Error rates of soft-margin MRSE for the vowel recognition example.**

CLASSIFICATION RULES

Suppose in the training phase, we have solved the soft-margin MRSE problems (8) for each class $k \in \{1, \ldots, m\}$. Let the optimal solutions be $\Phi_k^\star$, $\rho_k^\star$, $k = 1, \ldots, m$ (the optimal slack variables will not be used), and assume $\rho_k^\star > 1$ for all k. They are used in the testing or classification phase as follows. Given a new data point $x \in \mathbf{R}^d$, we first let $z = (x, 1) \in \mathbf{R}^{d+1}$ and compute

$$\rho_k = z^T \Phi_k^\star z^T, \quad k = 1, \ldots, m,$$

then label the data with the class

$$\hat{k} = \arg\min_k (\rho_k - 1)/(\rho_k^\star - 1).$$

For the variant (12) and all-versus-all approach (14), the classification rule is simply

$$\hat{k} = \arg\min_k z^T \Phi_k^\star z^T. \qquad (15)$$

As an application example, we apply the soft-margin MRSE approach to the vowel recognition problem described in [12]. The data in this example has ten dimensions and 11 possible classes. There are 528 training samples and 462 testing samples. In the experiment, we used the formulation (12) and classification rule (15). Figure 4 shows both the training error rates and the testing error rates by varying the separation ratio ρ from 1.1 to 10. The error rates of the maximum-likelihood (ML) approached are also plotted for reference. At $\rho = 2$, the soft-margin MRSE approach gives training error rate 0.002 and testing error rate 0.44, which are comparable with the best results obtained with other approaches as listed in Table 12.3 of [12].

CONCLUSIONS

The soft-margin MRSE problem and its variants can be obtained in two ways: either by relaxation, i.e., dropping the log-determinant term in the otherwise complete probabilistic generative models or, by adding positive semidefinite constraints to generic SVM with quadratic kernels, to capture essential characteristics of Gaussian models. This reflects an interesting tradeoff between modeling and optimization: while generic SVMs use simple discriminative models that allow efficient (convex) optimization, generative models impose more structure on the problem (i.e., adding more constraints) and often make the optimization intractable. A key step in developing effective large-margin methods for discriminative training of generative models is to strike a balance between the model complexity and computational efficiency. This often requires appropriate convex reformulation or relaxation of the training criteria and parameter constraints. This article shows one concrete example of such reformulation.

The soft-margin approach described in this article is an alternative to the conventional discriminative methods; e.g., [2], [3], and [13]. The importance of discriminative learning in speech recognition has been elaborated in [14] and [15].

AUTHORS

Lin Xiao is a researcher in the Machine Learning group at Microsoft Research, Redmond, Washington.

Li Deng is a principal researcher in the Speech Technology group at Microsoft Research, Redmond, Washington.

REFERENCES

[1] F. Glineur, "Pattern separation via ellipsoids and conic programming," (in English), Master's thesis, Faculté Polytechnique de Mons, Belgium, 1998.

[2] H. Jiang and X. Li, "Parameter estimation of statistical models using convex optimization," *IEEE Signal Processing Mag.*, vol. 27, no. 3, pp. 115–127, May 2010.

[3] F. Sha and L. K. Saul, "Large margin training of hidden Markov models for automatic speech recognition," in *Proc. Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hofmann, Eds. Cambridge, MA: MIT Press, 2007, pp. 1249–1256.

[4] T.-H. Chang, Z.-Q. Luo, L. Deng, and C.-Y. Chi, "A convex optimization method for joint mean and variance parameter estimation of large-margin CD-HMM," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'08)*, 2008, pp. 4053–4056.

[5] E. R. Barnes, "An algorithm for separating patterns by ellipsoids," *IBM J. Res. Develop.*, vol. 26, no. 6, pp. 759–764, Nov. 1982.

[6] G. Calafiore, "Approximation of n-dimensional data using spherical and ellipsoidal primitives," *IEEE Trans. Syst., Man, Cybern. A*, vol. 32, no. 2, pp. 269–278, 2002.

[7] P. Liu and F. Soong, "A quadratic optimization approach to discriminative training of CDHMMs," *IEEE Signal Processing Lett.*, vol. 16, no. 3, pp. 149–152, 2009.

[8] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[9] J. Shawe-Taylor and N. Cristianini, *Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[10] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[11] L. Vandenberghe, S. Boyd, and S.-P. Wu, "Determinant maximization with linear matrix inequality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 19, no. 2, pp. 499–533, 1998.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2001.

[13] X. He, L. Deng, and C. Wu, "Discriminative learning in sequential pattern recognition," *IEEE Signal Processing Mag.*, vol. 25, no. 5, pp. 14–36, Sept. 2008.

[14] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, "Research developments and directions in speech recognition and understanding, Part 1," *IEEE Signal Processing Mag.*, vol. 26, no. 3, pp. 75–80, May 2009.

[15] J. Baker, L. Deng, S. Khudanpur, C.-H. Lee, J. Glass, and N. Morgan, "Updated MINDS report on speech recognition and understanding," *IEEE Signal Processing Mag.*, vol. 26, no. 4, pp. 78–85, July 2009.

[SP]