CAP6615 Midterm I
Fall 2004
Tuesday, Oct. 26^{th} 2004
6:15PM-12:15AM

1. **[25 points] Fisher linear discriminants:** In a two class problem, the Fisher linear discriminant is a useful technique for determining the one dimension along which the patterns $\{\mathbf{x}_i \in \mathbb{R}^D, \forall i \in C_1\}$ and $\{\mathbf{x}_j \in \mathbb{R}^D, \forall j \in C_2\}$ have the largest value of the Fisher linear discriminant (FLD) objective function $J_W(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$, where $S_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T$ and $S_W = \sum_{i \in C_1}(\mathbf{x}_i - \mathbf{m}_1)(\mathbf{x}_i - \mathbf{m}_1)^T + \sum_{j \in C_2}(\mathbf{x}_j - \mathbf{m}_2)(\mathbf{x}_j - \mathbf{m}_2)^T$. The means for classes C_1 and C_2 are $\mathbf{m}_1$ and $\mathbf{m}_2$ respectively. The single dimension along which the patterns are projected to is $y_k = \hat{\mathbf{w}}^T \mathbf{x}_k$ where $\mathbf{x}_k$ is a pattern that could be from either class. Here it should be understood that $\hat{\mathbf{w}}$ is the weight vector that maximizes $J_W(\mathbf{w})$. Unless otherwise specified, if the index k is used as a pattern index, it means that the index is applicable to all patterns and not merely patterns from either C_1 or C_2.

It is often desirable to visualize the patterns in 2D or in 3D. Since the patterns usually live in a high dimensional space, this is impossible at the present time. The aim of this question is to come up with a mathematical criterion by which we have a "best" way of looking at the patterns in 2D or in 3D.

To this end, for 2D visualization, we would like to recover a $D \times 2$ weight matrix $\mathbf{W}$ rather than a weight vector $\mathbf{w}$ and for 3D visualization, we would like to recover a $D \times 3$ weight matrix $\mathbf{W}$ rather than a weight vector $\mathbf{w}$. After obtaining the weight matrix $\hat{\mathbf{W}}$, the new 2D or 3D feature vector corresponding to $\mathbf{x}_k$ is $\mathbf{y}_i = \hat{\mathbf{W}}^T \mathbf{x}_k$. The set of new feature vectors $\{\mathbf{y}_k\}$ can be visualized in a 2D plane or in a 3D volume depending on whether D is two or three.

(a) [10 points] We would like you to set up a Fisher linear discriminant to recover the best two- (for 2D) or best three- (for 3D) dimensional feature vectors. Give a conceptual level approach for achieving this aim. In particular, please ensure that you conceptually handle the issue that the different columns (2 for 2D and 3 for 3D) of $\hat{\mathbf{W}}$ have to be different for the visualization to be meaningful. Why is this necessary?

(b) [10 points] A simple way of achieving this aim is to repeatedly run the Fisher linear discriminant while making sure that the second (and/or the third) weight vector are not the same as the first. You are **NOT** allowed to use this recursive approach. Instead, try and design a "batch"

objective function $J_W(\mathbf{W})$ which when maximized, yields $\hat{\mathbf{W}}$. What constraints must you impose on $\mathbf{W}$ such that you can avoid the problem of identical columns.

(c) [5 points] Repeat **(b)** above for the Kernel Fisher discriminant. Each feature vector $\mathbf{x}$ is mapped to a new vector $\Phi(\mathbf{x})$. Design a Kernel Fisher discriminant to find W.

2. **[25 points] Support Vector Machines:** Two-dimensional patterns are placed at $\mathbf{x}_1 = (1, 0)$, $\mathbf{x}_2 = (-1, 0)$, $\mathbf{x}_3 = (0, 1)$ and $\mathbf{x}_4 = (0, -1)$. You are given the classification $\mathbf{x}_1 \in C_1$, $\mathbf{x}_2 \in C_2$, $\mathbf{x}_3 \in C_1$ and $\mathbf{x}_4 \in C_2$. In addition, we also add $N - 4$ 2D patterns obeying the criterion

$$\begin{aligned} \text{If } x^{(1)} + x^{(2)} &> 1, \text{ then } \mathbf{x} \in C_1, \\ \text{If } x^{(1)} + x^{(2)} &< -1, \text{ then } \mathbf{x} \in C_2, \\ & -1 \leq x^{(1)} + x^{(2)} \leq 1 \text{ is impossible} \end{aligned}$$

with the understanding that $-1 \leq x^{(1)} + x^{(2)} \leq 1$ is only allowed for the above four "privileged" patterns.

(a) [10 points] Find a solution to the SVM optimization problem

$$\alpha_0 = \arg\max_{\alpha} W(\alpha) = \arg\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \tag{1}$$

subject to the constraints

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \; \alpha_i \geq 0, \; \forall i \in \{1, \ldots, N\}. \tag{2}$$

Since the patterns are linearly separable, the notation $(\mathbf{x}_i \cdot \mathbf{x}_j)$ denotes a standard vector-vector dot product.

(b) [5 points] Once you have found the optimal $\alpha_0 = \{\alpha_i^{(0)}\}$, determine the best b_0 and $\psi_0 = \sum_{i=1}^{N} \alpha_i^{(0)} y_i \mathbf{x}_i$.

(c) [5 points] Having obtained ψ_0, compute ϕ_0.

(d) [5 points] Compute the margin $[c_1(\phi_0) - c_2(\phi_0)]/2$, $W(\alpha_0)$ and $\psi_0 \cdot \psi_0$.

3. **[25 points] Kernels:** Assume standard feature vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$. If the feature vectors are projected into a higher dimensional space, the resulting new feature vectors can be written as $\Phi(\mathbf{x})$ and $\Phi(\mathbf{y})$ corresponding to $\mathbf{x}$ and $\mathbf{y}$ respectively with $\Phi(\cdot)$ denoting the mapping.

(a) [5 points] If the mapping is from a D dimensional space to a K dimensional space with $K > D$, specify the nature and number of functions involved in the mapping Φ. For example, if the mapping is from 2D to 3D, you need three functions each of which maps $\mathbb{R}^2 \rightarrow \mathbb{R}$.

(b) [8 points] Assume that K is finite. Give a conceptual level answer to the following question. Since K is finite, we can directly compute inner products between $\Phi(\mathbf{x})$ and $\Phi(\mathbf{y})$ using the standard vector space formula for an inner product. That is, we can think of $\Phi(\mathbf{x})$ and $\Phi(\mathbf{y})$ as vectors $\mathbf{u}$ and $\mathbf{v}$ with the inner product $\mathbf{u} * \mathbf{v} \stackrel{\text{def}}{=} \sum_i u_i v_i$. Is there a contradiction between using this direct "vector space" inner product and the kernel inner product formula? Explain using an example.

(c) [7 points] Again assuming that K is finite, derive a relationship between the kernel and the functions in $\Phi(\mathbf{x})$ such that the direct inner product $\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})$ gives the same answer as $k(\mathbf{x}, \mathbf{y})$ where $k(\cdot, \cdot)$ is the kernel function used.

(d) [5 points] Show that the polynomial kernel of degree two $(\mathbf{x} \cdot \mathbf{y})^2$ satisfies your criterion. You'll have to define $\Phi(\mathbf{x})$ such that this works.

(e) [**Extra credit:** 15 points] Is it possible to show that taking inner products using $\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})$ is the same as using a kernel inner product $k(\mathbf{x}, \mathbf{y})$ for a polynomial kernel of non-negative integer degree d—$k(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y})^d$? You have to show this for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$. There is no partial credit for this question.

4. **[25 points] Relational Clustering**: The K-means clustering algorithm can be seen as minimizing the following objective function

$$\min_{(M,\mathbf{y})} E_{\text{cluster}}(M, \mathbf{y}) = \min_{(M,\mathbf{y})} \sum_{i=1}^{N} \sum_{a=1}^{K} M_{ia} ||\mathbf{x}_i - \mathbf{y}_a||^2 \tag{3}$$

subject to the constraints $\sum_{a=1}^{K} M_{ia} = 1$ and $M_{ia} \in \{0, 1\}$ where $\{\mathbf{x}_i\}$ is the data, $\{\mathbf{y}_a\}$ the set of cluster centers and $\{M_{ia}\}$ the set of memberships of data points in clusters. The chicken and egg problem of computing the memberships and cluster centers with both being unknown is now cast as an optimization problem wherein we have two sets of unknowns—the memberships $\{M_{ia}\}$ and the cluster centers $\{\mathbf{y}_a\}$.

(a) [3 points] Show that the solution for $\{M_{ia}\}$ while keeping the clusters centers $\{\mathbf{y}_a\}$ fixed is equivalent to choosing $M_{ia} = 1$ for the cluster center $\mathbf{y}_a$ which is closest to $\mathbf{x}_i$.

(b) [2 points] Show that the solution for $\{\mathbf{y}_a\}$ while keeping the memberships $\{M_{ia}\}$ fixed is equivalent to choosing $\mathbf{y}_a$ to be the centroid of all data points $\{\mathbf{x}_i\}$ which "belong" to it.

We now move from point clustering to relational clustering. In the previous K-means clustering case, we assigned the membership of a data point to the current nearest cluster center. Instead of doing this, we now wish to assign the membership of data point "i" by also examining the membership of nearby data point "j" and how close "j" is to cluster "a". To do this, we first parse the data and generate a nearest neighbor graph $\{G_{ij}\}$. If $G_{ij} = 1$, it implies that "i" and "j" are neighbors. For the sake of simplicity, assume that this graph is symmetric. That is, if "i" is a neighbor of "j", then "j" is a neighbor of "i".

(c) [10 points] Given this graph G, design a new *pairwise* relational clustering objective function somewhat similar to (3). The new objective function should have the following properties. i) It should only have quadratic terms linking $\{\mathbf{x}_i\}$ to $\{\mathbf{y}_a\}$ [for example $\| \mathbf{x}_i - \mathbf{y}_a \|^2$], ii) it should cover the pairs of data $\{\mathbf{x}_i\}$ which are neighbors in G, iii) it should use both the memberships of data point "i" in "a" and of data point "j" in "a" for neighbors "i" and "j" in G. Give a conceptual level explanation followed by a mathematical one.

(d) [5 points] Derive the solution for M_{ia} which should depend on the current value of $\mathbf{y}_a$ and the memberships M_{ja} of "i"s neighbors "j" in G.

(e) [5 points] Derive the solution for $\mathbf{y}_a$ which should depend on the current value of $\{M_{ia}\}$.