CAP6615
Midterm I
Fall 2002

30th October 2002

1. **[30 points] Fisher linear discriminant:** In a two class problem, the Fisher linear discriminant is a useful technique for determining the one dimension along which the patterns $\{\mathbf{x}_i, \forall i \in C_1\}$ and $\{\mathbf{x}_j, \forall j \in C_2\}$ have the largest value of $J_W(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$, 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 class 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 $y_k = \hat{w}^T \mathbf{x}_k$ where $\mathbf{x}_k$ is a pattern that could be from either class. Here it should be understood that $\hat{w}$ 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 C_1 or C_2.

One drawback of the Fisher linear discriminant is that all of the feature vectors are collapsed into a one dimensional space. Quite often, there is useful information not just in the single best dimension of discrimination, but in the second best, the third best etc.

a) [10 points] We would like you to set up the Fisher discriminant in order to recover not just the single best dimension $\{y_k\}$, but also the second best dimension $\{y_k^{(2)}\}$. Give a conceptual level approach for achieving this objective. You are free to conceptually incorporate the hint given in the next question. (A purely mathematical answer with no conceptual elaboration will not be given many points.)

b) [5 points] One way of achieving this is the following. After first obtaining $\hat{w}$, project all the features into a subspace that is orthogonal to the first dimension. That is, given the original patterns, find a way of obtaining a new set of patterns $\{\mathbf{x}_k^{(2)}\}$ which all have the property that $\hat{w}^T \mathbf{x}^{(2)} = 0$. [Hint: This can be achieved by setting $\mathbf{x}^{(2)} = \mathbf{x} - \mathbf{z}y$ for an appropriate choice of the vector $\mathbf{z}$]. Show how this projection procedure can be used to obtain the second best dimension $\{y_k^{(2)}\}$ of discrimination.

c) [5 points] Now, let's move over to a (possibly infinite-dimensional) Hilbert space. Each feature vector $\mathbf{x}$ is mapped to a new vector $\Phi(\mathbf{x})$. A Kernel Fisher discriminant is used to find the single best dimension y. What is the solution for $\hat{w}$ in this Kernel Fisher setup? Under what circumstances is it computable?

d) [10 points] Now determine the second best dimension $\{y_k^{(2)}\}$ by setting $\Phi^{(2)}(\mathbf{x}^{(2)}) = \Phi(\mathbf{x}) - \mathbf{z}y$ and requiring the inner product between $\hat{w}$ and $\Phi^{(2)}(\mathbf{x}^{(2)})$ to be zero. As before, determine $\mathbf{z}$ and $\{y_k^{(2)}\}$.

2. **[20 points] Multi-layer perceptrons:** Assume a two-layer perceptron (with one hidden layer). The equations for the input-to hidden layer are

$$\begin{aligned} a_j &= \sum_i w_{ji} x_i \\ z_j &= g^{(1)}(a_j) \end{aligned}$$

and the equations for the hidden-to-output layer are

$$\begin{aligned} a_k &= \sum_j w_{kj} z_j \\ z_k &= g^{(2)}(a_k) \end{aligned}$$

a) [10 points] Show that the multi-layer perceptron is equivalent to a single-layer perceptron if the hidden layer unit z_j is a linear function of its input a_j. Write down the equation for the equivalent single-layer perceptron.

b) [3 points] What happens if $z_j = \sum_{j'} u_{jj'} a_{j'}$ where $\{u_{jj'}\}$ is a further set of weights? Do you still get an equivalent single-layer perceptron?

c) [7 points] Argue at a conceptual level, the minimum requirement for keeping the higher layers of a multi-layer perceptron from "crashing down" and giving you an effective single layer perceptron. (Once again, a purely mathematical answer with no conceptual elaboration will not be given many points.)

3. **[20 points] SVM learning:** Assume that you have optimized the SVM cost function $W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$ and that you have in your possession a set $\{\hat{\alpha}_i,\ i \in \{1, \ldots, N\}\}$. You may also assume that only a subset $\{\hat{\alpha}_s,\ s \in S\}$ are greater than zero. Please note that no kernel is being used here.

a) [7 points] What inequalities must the solution for b satisfy? Explain. You may assume that the patterns are linearly separable.

b) [7 points] Now assume that the patterns are not linearly separable due to a small subset of patterns Ω which can be regarded as outliers. Assume that you have the global maximum of the above cost function $W(\alpha)$ subject to the constraints $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{N} \alpha_i y_i = 0$. How do you identify which patterns are outliers based on the obtained solution? What inequalities must the solution for b *now* satisfy? Explain.

c) [6 points] Calculate the margin $\rho(\phi)$ and $c_1(\phi)$ and $c_2(\phi)$ in terms of $\{\hat{\alpha}_i, \mathbf{x}_i\}, b$ and $\{y_i\}$. You may assume that the patterns are linearly separable for this sub-question.

4. **[30 points] Obtaining a distance from an inner product kernel:** If we have two feature vectors $\mathbf{x}$ and $\mathbf{y}$, the square of the Euclidean distance between the feature vectors is written as $||\mathbf{x} - \mathbf{y}||^2 = \sum_{k=1}^{D} (x_k - y_k)^2$ where it should be understood that the feature vectors $\mathbf{x}$ and $\mathbf{y}$ live in a $D-$dimensional space. Now, if the feature vectors are projected into a (usually higher dimensional) Hilbert 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) [10 points] If the squared distance between $\mathbf{x}$ and $\mathbf{y}$ is defined as $d(\mathbf{x}, \mathbf{y}) \overset{def}{=} ||\Phi(\mathbf{x}) - \Phi(\mathbf{y})||^2$, evaluate $d(\mathbf{x}, \mathbf{y})$ for a i) polynomial kernel $k_{Poly}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^2$ and for a ii) Gaussian radial basis function kernel $k_{GRBF}(\mathbf{x}, \mathbf{y}) = \exp\{-\frac{1}{2}\frac{||\mathbf{x}-\mathbf{y}||^2}{\sigma^2}\}$.

b) [5 points] This next question concerns a "walk" from a feature vector $\mathbf{x}$ to a feature vector $\mathbf{x} + \Delta\mathbf{x}$ where $\Delta\mathbf{x}$ should be read as a very small displacement. Equipped with the formulas for the squared distance $d(\mathbf{x}, \mathbf{y})$ for the two kernels above, evaluate the distances once again but now for $d(\mathbf{x}, \mathbf{x} + \Delta\mathbf{x})$.

c) [15 points] Taking the limit as $\Delta\mathbf{x} \to 0$, derive a general formula for $d(\mathbf{x}, \mathbf{x} + \Delta\mathbf{x})$ in terms of the partial derivatives of the kernel. This quantity is called the metric and plays an extremely important role in analysis.