CS 688 – Fall 2010

Homework 3 – Due Nov. 23

Instructor: Carlotta Domeniconi

Problem 1 Consider the following kernel function: $K(\mathbf{x}_i, \mathbf{x}_j) = (<\mathbf{x}_i, \mathbf{x}_j>)^2$.

Verify that for each of the following two mappings ϕ, it holds $K(\mathbf{x}_i, \mathbf{x}_j) =< \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) >$. Show your calculations.

1. $\phi : \Re^2 \rightarrow \Re^3$, $\phi(\mathbf{x}) = \frac{1}{\sqrt{2}} \begin{pmatrix} x_1^2 - x_2^2 \\ 2x_1x_2 \\ x_1^2 + x_2^2 \end{pmatrix}$

2. $\phi : \Re^2 \rightarrow \Re^4$, $\phi(\mathbf{x}) = \begin{pmatrix} x_1^2 \\ x_1x_2 \\ x_1x_2 \\ x_2^2 \end{pmatrix}$

Problem 2 Consider the simple linear SVM classifier $(w_1x + w_0)$, and the non-linear SVM classifier $(\mathbf{w}^t\phi + w_0)$, where $\phi : \Re \rightarrow \Re^2$ is defined as: $\phi(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$

(a) Provide three input points x_1, x_2, and x_3 and their associated labels (-1 or $+1$) such that they cannot be separated with the simple linear classifier, but are separable by the non-linear classifier with $\phi = (x, x^2)^t$.

(b) Mark your three points x_1, x_2, and x_3 as points in the *feature space* with their associated labels. Draw the decision boundary of the non-linear SVM classifier that separates the points *in the feature space* obtained with $\phi = (x, x^2)^t$.

Problem 3 Consider a two-class classification problem in one dimension. Suppose that the two classes have equal prior probabilities: $P(c_1) = P(c_2) = 1/2$, and distributions as follows:

$$p(x|c_1) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

$$p(x|c_2) = \begin{cases} 2 - 2x & 0 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

(a) Plot the two functions $p(x|c_1)$ and $p(x|c_2)$ (within the same Cartesian plane).

(b) Plot the Bayes decision boundary and give its equation.

(c) What is the Bayes decision rule?

Problem 4 (**Clustering**) Consider the 3-means algorithm on a set S consisting of the following six two-dimensional points: $a = (0, 0)$, $b = (8, 0)$, $c = (16, 0)$, $d = (0, 6)$, $e = (8, 6)$, $f = (16, 6)$. The algorithm uses the Euclidean distance to assign each point to the nearest centroid; ties are broken in favor of the centroid to the left/down. A starting configuration is a subset of three starting points from S that form the initial centroids. A 3-partition is a partition of S into 3 subsets; thus, {a,b,e}, {c,d}, {f} is a 3-partition. Clearly, any 3-partition induces a set of three centroids in a natural way. A 3-partition is stable if repetition of the 3-means iteration with the induced centroids leaves it unchanged.

1. How many starting configurations are there?
2. What are the stable 3-partitions?
3. What is the number of starting configurations leading to each of the stable 3-partitions.
4. What is the maximum number of iterations from any starting configuration to its stable 3-partition?

Note: Although it is not required, you may want to implement the K-means algorithm, and run it on the data given above to verify your answers.

Instructions Complete the homework by November 23. You will not turn in the solutions to me. Students will solve the problems on the board during class on November 23. Attendance is required on that day.