Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 3
Due: March 14, 2006

Problem 1: Support Vector Machines

Let S be a training sample of size m.

(1) Assume that S has been generated according to some probability distribution $D(x, y)$, where $(x, y) \in X \times \{-1, +1\}$.

(a) Define the Bayes classifier $h^* : X \to \{-1, +1\}$. Show that the kernel K^* defined by $K^*(x, x') = h^*(x)h^*(x')$ for any $x, x' \in X$ is positive definite symmetric. What is the dimension of the natural feature space associated to K^*?

(b) Give the expression of the solution obtained using SVMs with this kernel. What is the number of support vectors? What is the value of the margin? What is the generalization error of the solution obtained? Under what condition is the data linearly separable?

(c) Let $h : X \to \mathbb{R}$ be an arbitrary real-valued function. Under what condition on h is the kernel K defined by $K(x, x') = h(x)h(x')$, $x, x' \in X$, positive definite symmetric?

(2) Download the files "negative.dat" and "positive.dat" which contain an equal number of negative and positive examples. Download and install the libsvm software library from:

`http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

(a) Let N_{SV} be the number of support vectors obtained when training an SVM. At most how many times do you need to train on a sample of size $m-1$ in order to compute the leave-one-out error?

(b) The leave-one-out error may have high variance and is costly to compute. Instead, you can do ten-fold cross validation: split the data in ten equal parts, ten times train on 9 and test on one, and compute the average error and standard deviation.

Use SVMs with polynomial kernels with samples of size 2000, 4000, 6000, 8000 and 10000, and plot the test error as a function

of the sample size for different values of the polynomial degree, $d = 1, \ldots, 4$.

Problem 2: Kernel Methods

Let $X = \{a, c, g, t\}$. To classify DNA sequences using SVMs, we wish to define a kernel between sequences defined over X. We are given a finite set $I \subset X^*$ of non-coding regions (introns). For $x \in X^*$, denote by $|x|$ the length of x and by $F(x)$ the set of factors of x, i.e., the set of subsequences of x with contiguous symbols. For any two strings $x, y \in X^*$ define $K(x, y)$ by

$$K(x, y) = \sum_{z \in (F(x) \cap F(y)) - I} \rho^{|z|}, \tag{1}$$

where $\rho \geq 1$ is a real number.

(1) Show that K is a rational kernel and that it is positive definite symmetric.

(2) Give the time and space complexity of the computation of $K(x, y)$ with respect to the size s of a minimal automaton representing $X^* - I$.

(3) Long common factors between x and y of length greater than or equal to n are likely to be important coding regions (exons). Modify the kernel K to assign weight $\rho_2^{|z|}$ to z when $|z| \geq n$, $\rho_1^{|z|}$ otherwise, where $1 \leq \rho_1 \ll \rho_2$. Show that the resulting kernel is still positive definite symmetric.