COS 511: Foundations of Machine Learning

Homework #5 | Due: April 8, 2003

Problem 1

Suppose we use support-vector machines with the kernel:

$$K(x,u) = \begin{cases} 1 & \text{if } x = u \\ 0 & \text{otherwise.} \end{cases}$$

As we discussed in class, this corresponds to mapping each x to a vector $\psi(x)$ in some high dimensional space (that need not be specified) so that $K(x,u) = \psi(x) \cdot \psi(u)$.

As usual, we are given m examples $(x_1, y_1), \ldots, (x_m, y_m)$ where $y_i \in \{-1, +1\}$. Assume for simplicity that all the x_i's are distinct (i.e., $x_i \neq x_j$ for $i \neq j$).

a. Recall that the weight vector $\mathbf{w}$ used in SVM's has the form

$$\mathbf{w} = \sum_i \alpha_i y_i \psi(x_i).$$

Compute the α_i's explicitly that would be found using SVM's with this kernel.

b. Recall that the SVM algorithm outputs a classifier that, on input x, computes the sign of $\mathbf{w} \cdot \psi(x)$. What is the value of this inner product on training example x_i? What is the value of this inner product on any example x not seen during training? Based on these answers, what kind of generalization error do you expect will be achieved by SVM's using this kernel?

c. Recall that the generalization error of SVM's can be bounded using the margin δ (which is equal to $1/\|\mathbf{w}\|$), or using the number of support vectors. What is δ in this case? How many support vectors are there in this case? How are these answers consistent with your answer in part (b)?

Problem 2

Consider the problem of learning with expert advice when one of the experts gives perfect predictions. On some round t, let q be the fraction of surviving experts that predict 1. (A surviving expert is one that has not made any mistakes so far.) In class, we talked about the halving algorithm which predicts with the majority vote of the expert predictions, and we talked about the randomized weighted majority algorithm (with β set to zero) which predicts with one randomly selected expert.

In general, we can predict 1 with probability $F(q)$ and 0 with probability $1 - F(q)$ for some function F. For instance, for the halving algorithm, $F(q)$ is 1 if $q > 1/2$ and 0 if $q < 1/2$ (and arbitrary if $q = 1/2$). For the randomized weighted majority algorithm (again, with $\beta = 0$), $F(q) = q$.

Consider now a function $F : [0,1] \to [0,1]$ satisfying the following property:

$$1 + \frac{\lg q}{2} \leq F(q) \leq -\frac{\lg(1-q)}{2}. \tag{1}$$

a. Suppose we run an on-line learning algorithm that uses a function F satisfying (1) as described above. Show that the expected number of mistakes made by the learning algorithm is at most $(\lg N)/2$, where N is the number of experts. (Hint: consider the cases $y_t = 0$ and $y_t = 1$ separately. In each case, show that at most a fraction $4^{-\ell_t}$ of the surviving experts survive to the next round, where ℓ_t is the probability that the learning algorithm makes a mistake on round t.)

b. Show that the function

$$F(q) = \frac{\lg(1-q)}{\lg q + \lg(1-q)}$$

has range $[0, 1]$ and satisfies (1). (At the endpoints, we define $F(0) = 0$ and $F(1) = 1$ to make F continuous, but don't worry about these.)

c. **(Optional)** Suppose now that there are $k \geq 2$ possible outcomes rather than just 2. In other words, the outcome y_t is now in the set $\{1, \ldots, k\}$ (rather than $\{0, 1\}$ as we have considered up until now), and likewise, both experts and the learning algorithm make predictions in this set. Assume one of the experts makes perfect predictions. On some round t, let q_j be the fraction of surviving experts predicting outcome $j \in \{1, \ldots, k\}$. Suppose that the learning algorithm predicts each outcome j with probability

$$\frac{\lg(1-q_j)}{\sum_{i=1}^{k} \lg(1-q_i)}.$$

Show that the expected number of mistakes of this learning algorithm is at most $(\lg N)/2$.