CPSC 531H Machine Learning Theory (Term 2, 2013-14)
Assignment 2

Due: Tuesday February 25th, in class.

Question 1: [Mohri 6.1] Let $\mathcal{H}$ be a set of classifiers with VC-dimension d. Let $\mathcal{F}_t$ be the set of classifiers obtained by taking a weighted majority vote of t classifiers from $\mathcal{H}$, as in the AdaBoost algorithm. Prove that the VC-dimension of $\mathcal{F}_t$ is at most $O(td\log(td))$.

Note: You only need to prove an upper bound, not a lower bound.

Hint: It could be helpful to use the Sauer-Shelah lemma.

Question 2: [Mohri 6.3] Assume that the main weak learner assumption of AdaBoost holds (i.e., under any distribution, there exists a base learner with error strictly better than 1/2). Let h_t be the base learner selected at round t. Show that the base learner h_{t+1} selected at round $t+1$ must be different from h_t.

Question 3: Prof. Marge Innizwut proposes the following simple kernel function:

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

- **(a):** Prove this is a legal kernel. You may assume the instance space X is finite. Specifically, describe a mapping $\Phi : X \to \mathbb{R}^m$ (for some value m) such that $K(x, x') = \Phi(x)^\mathsf{T}\Phi(x')$.
- **(b):** Marge likes this kernel because in the range of Φ, any labeling of the points in X will be linearly separable. So, this should be perfect for learning any desired target function just run a kernelized version of Perceptron or SVM. Why is any assignment of labels to points linearly separable?
- **(c):** What is the problem with Marge's reasoning — why does this kernel not necessarily make the learning task easy?

Question 4 is on the reverse side.

Question 4: $(1-\epsilon)$**-approximation to maximum margin via Perceptron**

The simple Margin-Perceptron algorithm from Lecture 10 gave us a 1/3-approximation to the maximum margin. In this exercise, let's derive the variant of Margin-Perceptron that gives a $(1-\epsilon)$-approximation.

The basic algorithm takes the training data, ane arbitrary parameter γ as input, and our desired approximation error ϵ as input. Let us assume that $\|x_i\| = 1$ for all i.

Margin-Perceptron

- **Input:** $(x_1, y_1), \ldots, (x_m, y_m)$, $\gamma \in [0,1]$, $\epsilon \in [0,1]$.
- Initialize $w_0 \leftarrow 0$ and $t \leftarrow 0$
- Repeat
  - Find any i with either

$$\textbf{Misclassification:} \quad y_i \neq \operatorname{sign}(w_t^{\mathsf{T}} x_i)$$
$$\textbf{Poor margin:} \quad |w_t^{\mathsf{T}} x_i| / \|w_t\| \leq (1-\epsilon)\gamma$$

  - If such an i is found, set $w_{t+1} \leftarrow w_t + y_i x_i$ and $t \leftarrow t+1$.
- Until no such i exists
- Output $w_t / \|w_t\|$

(a): Suppose that there exists a linear threshold function $x \mapsto \operatorname{sign}(\bar{w}^{\mathsf{T}} x)$ with $\operatorname{margin}(\bar{w}) \geq \gamma$. Prove that

$$\|w_t\| \geq t\gamma \qquad \text{for all } t \geq 0.$$

Hint: Use Cauchy-Schwarz.

(b): Prove that

$$\|w_{t+1}\| \leq \|w_t\| + (1-\epsilon)\gamma + \frac{1}{2\|w_t\|}.$$

Hint: Use the Taylor approximation of $\sqrt{x}$ at $x = 1$.

(c): Prove that

$$\|w_t\| \leq \frac{2}{\epsilon\gamma} + (1-\epsilon/2)\gamma t \qquad \text{for all } t \geq 0.$$

Hint: Consider separately the cases $\|w_t\| < 1/(\epsilon\gamma)$ and $\|w_t\| \geq 1/(\epsilon\gamma)$. In the former case use a trivial bound, and in the latter case use part (b).

(d): Assume the existence of $\bar{w}$ as in part (a). Conclude that, after at most $4/(\epsilon\gamma)^2$ iterations, Margin-Perceptron outputs a classifier with margin at least $(1-\epsilon)\cdot\gamma$. *Hint:* Combine the lower bounds and upper bounds on $\|w_t\|$.

(e): Let $\gamma^* = \max_w \operatorname{margin}(w)$ be the maximum margin of any linear classifier on the given examples. Design a new function Margin-Maximizer which takes as input the labeled examples and the parameter ϵ. The new function can call Margin-Perceptron at most $O(\log(1/\gamma^*)/\epsilon)$ times. It must output a classifier with margin at least $(1-2\epsilon)\gamma^*$.