Machine learning@Data science, 2016–Exercises.

Set 3. Due February 26, 2016

Problem 12 Consider the class $\mathcal{A}$ of all sets of the form

$$A_\alpha = \{x \in \mathbb{R} : \sin(\alpha x) > 0\}$$

where $\alpha > 0$. What is the VC dimension of $\mathcal{A}$? (Note that $\mathcal{A}$ has one free parameter.)

Problem 13 Let $\mathcal{A}_1, \ldots, \mathcal{A}_k$ be classes of sets, all of the with VC dimension at most V. Show that the VC-dimension of $\cup_{i=1}^k \mathcal{A}_i$ is at most $4V \log_2(2V) + 4k$. You may use the fact that for $a \geq 1$ and $b > 0$, if $x \geq 4a \log(2a) + 2b$ then $x \geq a \log x + b$.

Can you bound the VC-dimension of the class of all sets of the form

$$A_1 \cup \cdots \cup A_k \quad \text{with} \quad A_1 \in \mathcal{A}_1, \ldots, A_k \in \mathcal{A}_k?$$

Problem 14 (PERCEPTRON CONVERGENCE.) Consider the "normalized" version of the perceptron algorithm in which one starts with a nonzero vector $w_0 \in \mathbb{R}^d$ and, cycling through the data, one sets, for $t = 1, 2, \ldots$

$$w_t = \begin{cases} w_{t-1} & \text{if } w_{t-1}^T X_t Y_t \geq 0 \\ w_{t-1} + Y_t X_t / \|X_t\| & \text{otherwise.} \end{cases}$$

Suppose that the data are linearly separable. This means that there exists $w_* \in \mathbb{R}^d$ such that $w_*^T X_i Y_i / \|X_i\| \geq 1$ for all $i = 1, \ldots, n$. Prove that the algorithm finds a classifier that separates the data in at most $\|w_* - w_0\|^2$ updates.
Hint: Prove that whenever w_t is updated (i.e., $w_{t-1}^T X_t Y_t < 0$), one has $\|w_t - w_*\|^2 \leq \|w_{t-1} - w_*\|^2 - 1$.

Problem 15 Let g_n be an arbitrary (data-dependent) classifier. The leave-one-out error estimate is defined as

$$R_n^{(D)}(g_n) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{g_{n-1}(X_i, D_{n,i}) \neq Y_i\}}$$

where $D_{n,i} = ((X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n))$. Show that the estimate is nearly unbiased in the sense that

$$\mathbf{E} R_n^{(D)}(g_n) = \mathbf{E} R(g_{n-1}) \ .$$

Use this to derive a bound for the expected risk of a perceptron classifier of the previous exercise when the data are linearly separable (i.e., $L^* = 0$ and the Bayes classifier is linear).

Problem 16 Consider the majority classifier

$$g_n(x, D_n) = \begin{cases} 1 & \text{if } \sum_{i=1}^n Y_i \geq n/2 \\ 0 & \text{otherwise} \end{cases}$$

(Thus, g_n ignores x and the X_i's.) Assume that n is odd. What is the expected risk $\mathbf{E} R(g_n) = \mathbf{P}\{g_n(X) \neq Y\}$ of this classifier? Study the performance of the leave-one-out error estimate. Show that for some distributions $\text{Var}(R_n^{(D)}(g_n)) \geq c/\sqrt{n}$ for some constant c. *Hint:* Strange things happen when the number of 0's and 1's is about the same in the data.