

CS281B/Stat241B Homework Assignment 1 (due February 1, 2008, at noon, in the box outside Soda 723)

1. **(Multi-class pattern classification)** In the two-class pattern classification problem, we have seen that we can express the excess risk of a classifier f (that is, the amount by which its risk exceeds the Bayes risk) as

$$R(f) - R^* = \mathbb{E}(\mathbf{1}[f(X) \neq f^*(X)] |2\eta(X) - 1|).$$

It is easy to check that we can write this as

$$R(f) - R(f^*) = \mathbb{E} \left(\max_{y \in \mathcal{Y}} P(Y = y|X) - P(Y = f(X)|X) \right). \quad (1)$$

Consider the following multi-class problem: \mathcal{X} is the observation space, $\mathcal{Y} = \{1, \dots, k\}$ is the outcome space, and P is a probability distribution on $\mathcal{X} \times \mathcal{Y}$. Write $\eta_y(x) = P(Y = y|X = x)$. The Bayes decision function $f^* : \mathcal{X} \rightarrow \mathcal{Y}$ (that is, the classifier that minimizes risk, $R(f) = P(Y \neq f(X))$) is given by

$$f^*(x) = \arg \max_{y \in \mathcal{Y}} \eta_y(x).$$

- (a) Show that, for any $f : \mathcal{X} \rightarrow \mathcal{Y}$, (1) is also true in this case.
(b) Given estimates $\hat{\eta}_y$ of the conditional probability functions η_y , the *plug-in decision function* is

$$\hat{f}(x) = \arg \max_{y \in \mathcal{Y}} \hat{\eta}_y(x).$$

Derive an upper bound on the excess risk, $R(\hat{f}) - R^*$, in terms of the function

$$x \mapsto \max_{y \in \mathcal{Y}} |\eta_y(x) - \hat{\eta}_y(x)|.$$

2. **(Perceptron algorithm)** Implement the perceptron algorithm. Consider the following two data sets, which are labelled according to a linear threshold function. (Define $e_i \in \{0, 1\}^d$ as the unit vector with the i th component equal to 1.)

- (a) Choose $x_i \in \{0, 1\}^{d+1}$ with $x_i = e_i + e_{d+1}$, $y_i = 1$ for $i = 1, \dots, d$, and $x_{d+1} = e_{d+1}$, $y_{d+1} = -1$.
(b) Construct a set $S_n \subset \{0, 1\}^{2n}$ as follows. Set $S_1 = \{01, 10, 11\}$, and for $i \geq 1$,

$$S_{i+1} = \{x01 : x \in S_i\} \cup \{11 \dots 10, 00 \dots 011\}.$$

For each element $b = (b_1 \dots b_{2n})$ of S_n , define an associated boolean value $v_b \in \{0, 1\}$ as

$$v_b = b_{2n} \wedge (b_{2n-1} \vee (b_{2n-2} \wedge (b_{2n-3} \vee (\dots (b_2 \wedge b_1) \dots))),$$

where \wedge denotes conjunction ($b_1 \wedge b_2$ is 1 only for $b_1 = b_2 = 1$) and \vee denotes disjunction ($b_1 \vee b_2$ is 0 only for $b_1 = b_2 = 0$). Finally, set

$$\{(x_1, y_1), \dots, (x_{2n+1}, y_{2n+1})\} = \{(b, 1), 2v_b - 1\} : b \in S_n\}.$$

Plot the number of updates made by the perceptron algorithm with these two data sets, as a function of n for $d = 2n$. Comment on the difference. Explain why it occurs.

3. **(Lower bounds on risk in pattern classification)** We say that a class F of $\{\pm 1\}$ -valued functions defined on \mathcal{X} *shatters* a set $\{x_1, \dots, x_n\} \subseteq \mathcal{X}$ if

$$\{(f(x_1), \dots, f(x_n)) : f \in F\} = \{\pm 1\}^n,$$

that is, if F can compute all 2^n dichotomies of the set. The Vapnik-Chervonenkis dimension of F is

$$d_{VC}(F) = \max \{n : F \text{ shatters some } \{x_1, \dots, x_n\} \subseteq \mathcal{X}\}.$$

We have seen the following minimax lower bound on expected risk for the class of linear threshold functions on \mathbb{R}^d .

Theorem 1 For any classification rule f_n and any $n > 1$, there is a probability distribution P on $\mathcal{X} \times \{\pm 1\}$ for which some $f \in F$ has $L(f) = 0$ but

$$\mathbb{E}L(f_n) \geq \left(\frac{\min(n, d) - 1}{2n} \right) \left(1 - \frac{1}{n} \right)^n.$$

- (a) Show that this result remains true for F an arbitrary set of functions with $d_{VC}(F) \geq d$.
 (b) Hence prove minimax lower bounds on expected risk for the following classes of binary-valued functions.
- i. *Decision stumps* on \mathbb{R}^d ,

$$F = \{ \theta_{H(i, a, s)} : 1 \leq i \leq d, a \in \mathbb{R}, s \in \{\pm 1\} \},$$

where

$$\theta_S(x) = \begin{cases} 1 & \text{if } x \in S, \\ -1 & \text{otherwise,} \end{cases}$$

and $H(i, a, s)$, for $1 \leq i \leq d$, $a \in \mathbb{R}$ and $s \in \{\pm 1\}$, is the halfspace

$$H(i, a, s) = \{ x \in \mathbb{R}^d : s(x_i - a) > 0 \}.$$

- ii. *Indicators of unions of intervals* in \mathbb{R} ,

$$F = \{ \theta_U : U \text{ is a union of up to } k \text{ intervals} \}.$$

- iii. *Indicators of convex sets* in \mathbb{R}^2 ,

$$F = \{ \theta_S : S \subseteq \mathbb{R}^2 \text{ and } S \text{ convex} \}.$$