Problem 1 Prove that if $X \in \mathbb{R}^d$ and $Y \in \{0,1\}$ are independent random variables, then the Bayes risk equals $R^* = \min(p, 1-p)$ where $p = \mathbf{P}\{Y = 1\}$. Is there an example in which X and Y are not independent and yet $R^* = \min(p, 1-p)$?

Problem 2 Let $X' = (X_1, X_2)$ and let $R^*_{X_1}$ denote the Bayes risk for classifying the binary random variable Y based on the observation X_1. Similarly, let $R^*_{X'}$ denote the Bayes risk based on the joint observation of X_1 and X_2. Prove that $R^*_{X_1} \geq R^*_{X'}$ and, if X_2 is independent of (X_1, Y), then $R^*_{X_1} = R^*_{X'}$.

Problem 3 (UNBALANCED COST FUNCTION.) Consider the two-class classification problem in which the cost of misclassification is given by a loss function $\ell : \{0,1\} \times \{0,1\} \to [0, \infty)$. Thus, the risk of a classifier $g : \mathcal{X} \to \{0,1\}$ is $R(g) = \mathbf{E}\ell(g(X), Y)$. Determine the optimal classifier g^* that minimizes the risk and derive a formula for the Bayes risk $R^* = R(g^*)$ in terms of η and ℓ.

Problem 4 (CLASSIFICATION WITH REJECT OPTION.) In some classification problems, one is allowed to say "I don't know". Formally, a classifier may take three possible values: $0, 1$, and "I don't know". The probability of rejection is $\mathbf{P}\{g(X) = \text{"I don't know"}\}$, and the probability of error is $\mathbf{P}\{g(X) \neq Y | g(X) \neq \text{"I don't know"}\}$. Let $0 < c < 1/2$ and define

$$g_c(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 + c \\ 0 & \text{if } \eta(x) \leq 1/2 - c \\ \text{"I don't know"} & \text{otherwise.} \end{cases}$$

Show that this classifier is optimal in the sense that if for some classifier g

$$\mathbf{P}\{g(X) = \text{"I don't know"}\} \leq \mathbf{P}\{g_c(X) = \text{"I don't know"}\},$$

then

$$\mathbf{P}\{g(X) \neq Y | g(X) \neq \text{"I don't know"}\} \geq \mathbf{P}\{g_c(X) \neq Y | g_c(X) \neq \text{"I don't know"}\}.$$

Problem 5 Show that for any two classifiers $|R(g_1) - R(g_2)| \leq \mathbf{P}\{g_1(X) \neq g_2(X)\}$. When do we have equality?

Problem 6 (ADMISSIBILITY OF THE 1-NN RULE.) Show an example of a distribution for which the expected risk of the 1-nearest neighbor classifier is strictly smaller, for any sample size, than that of the k-nearest neighbor classifier for any odd integer $k > 1$.

Problem 7 (SCALE-INVARIANT NEAREST NEIGHBOR RULE?) The components of the observation vector $x \in \mathbb{R}^d$ often correspond to incomparable quantities. In such cases the Euclidean distance (or ℓ_1, ℓ_∞ distances) are not natural to use because a simple rescaling of a component can drastically change distances (and nearest neighbors). It would be desirable to have a distance measure that is invariant of any (possibly non-linear) re-scaling of the coordinates. Can you propose a (possibly data-based) way of assigning nearest neighbors that does not change by any monotone transformation of the coordinate exes?

Problem 8 (CHERNOFF BOUND FOR BINOMIAL RANDOM VARIABLES.) Let B be a binomial random variable with parameters n and p. Show that for all $t > np$,

$$\mathbf{P}\{B > t\} \leq e^{t - np - t\ln(t/(np))} .$$

Problem 9 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

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

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 for some distributions $\mathrm{Var}(R_n^{(D)}(g_n)) \geq c/\sqrt{n}$ for some constant c. *Hint:* Strange things happen when the number of 0's and 1's is about the same in the data.

Problem 10 What is the VC dimension of the class of all squares in the plane? (A square is a set of the form $\{(x, y) \in \mathbb{R}^2 : x \in [a, a+h], y \in [b, b+h]\}$ for some $a, b \in \mathbb{R}$ and $h \geq 0$.) What is the VC dimension of the class of all squares with a fixed side length, say $h = 1$? Can you generalize this to $\mathbb{R}^d$?

Problem 11 Let $\mathcal{A}_k$ be k the class of all sets in $\mathbb{R}$ that can be written as a union of k closed intervals. Determine the VC dimension of $\mathcal{A}_k$.

Problem 12 A *neural network with one hidden layer* is a function $\mathbb{R}^d \to \mathbb{R}$, that can be written as

$$\psi(x) = c_0 + \sum_{i=1}^k c_i \sigma(\psi_i(x)),$$

where $x = (x^{(1)}, \ldots, x^{(d)})^T$, is the input vector,

$$\sigma(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \end{cases}$$

and

$$\psi_i(x) = b_i + \sum_{j=1}^d a_{i,j} x^{(j)}.$$

(See Figure 1.)

Consider the class $\mathcal{A}$ of sets containing all sets of form $\{x : \psi(x) > 0\}$ where the parameters $a_{i,j}, b_i$ and c_i may take arbitrary real values. Estimate the shatter coefficients and/or VC dimension of the class $\mathcal{A}$. *Hint:* to obtain an upper bound recall that n points in $\mathbb{R}^d$ may be split at most n^{d+1} different ways by linear hyperplanes. For the lower bound try to show that kd points may be shattered when k is even. Take kd points in general position. to pick any subset of size $\leq kd/2$, divide these points in $k/2$ groups of size at most d. Fit a hyperplane to each group...

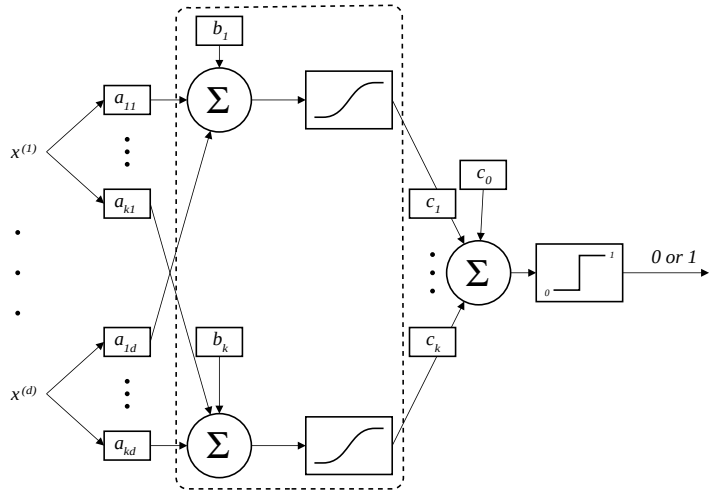

Figure 1: A neural netwok.

Problem 13 Let $(x_1, y_1), \ldots, (x_n, y_n)$ be data in $\mathbb{R}^d \times \{-1, 1\}$. Suppose that the data are *linearly separable*, that is, there exists a $w \in \mathbb{R}^d$ such that $y_i w^T x_i > 0$ for all $i = 1, \ldots, n$. The *margin* of such a vector is

$$\gamma(w) = \min_{i=1,\ldots,n} \frac{y_i w^T x_i}{\|w\|} .$$

Formulate a *convex optimization problem* whose solution is a vector w^* that classifies the data correctly (i.e., $y_i {w^*}^T x_i > 0$ for all $i = 1, \ldots, n$) and maximizes the margin. Show that the optimal solution w^* lies in the vector space spanned by the examples x_i for which the margin $\frac{y_i {w^*}^T x_i}{\|w^*\|}$ is minimal among all examples.

Problem 14 Consider the cost functional $A(f) = \mathbf{E}\phi(-f(X)Y)$ where $\phi : \mathbb{R} \to \mathbb{R}_+$ is a positive, increasing, strictly convex cost function, $f : \mathcal{X} \to \mathbb{R}$ is a real-valued function and $Y \in \{-1, 1\}$. Determine the function f^* that minimizes $A(f)$. Show that the classifier $g(x) = \operatorname{sgn}(f^*(x))$ is the Bayes classifier.

Problem 15 Give estimates for VC dimension of the following classes of sets: (1) convex polygons in the plane; (2) convex k-gons in the plane.

Problem 16 Suppose that data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, 1\}$ are such that each $x_i \in \{0, 1\}^d$ (i.e., the x_i have binary components) and for each $i = 1, \ldots, n$, $y_i = 1$ if and only if at least one component of x_i is 1. Show that the data are linearly separable. What is largest achievable margin (i.e., the smallest distance of all data points to the separating hyperplane) in the worst case?

Repeat the exercise in the case when for each $i = 1, \ldots, n$, $y_i = 1$ if and only if the sum of the components of x_i is at least $d/2$ (assume here that d is odd).

Problem 17 (VOTING OF DECISION STUMPS.) A *decision stump* in $\mathbb{R}^d$ is a classifier of the form

$$g(x) = \begin{cases} -1 & \text{if } x^{(i)} \le a \\ 1 & \text{if } x^{(i)} > a \end{cases} \quad \text{or} \quad \begin{cases} -1 & \text{if } x^{(i)} \ge a \\ 1 & \text{if } x^{(i)} < a \end{cases}$$

where $x^{(i)}$ is the i-th component of $x \in \mathbb{R}^d$ and $a \in \mathbb{R}$ is arbitrary. Consider the set $\mathcal{F}$ of real-valued functions on $\mathbb{R}^d$ that are convex combinations of decision stumps and classifiers of the form $\operatorname{sgn}(f(x))$ for some $f \in \mathcal{F}$.

What kind of classifiers do we obtain? For $d = 2$, show that a 2×2 "checkerboard" (i.e., $\operatorname{sgn}(x^{(1)}x^{(2)})$ for $x = (x^{(1)}, x^{(2)}) \in [-1, 1]^2$) cannot be realized by such a classifier.

Give an upper bound for the Rademacher complexity

$$\mathbf{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i)$$

where the σ_i are i.i.d. Rademacher random variables.