CSE 250B: Machine learning **Winter 2018**

Homework 1 — Nearest neighbor and statistical learning

By the due date, upload the PDF of your `typewritten` solutions to `gradescope`.

1. *Risk of a random classifier.* A particular data set has 4 possible labels, with the following frequencies:

Label	Frequency
A	50%
B	20%
C	20%
D	10%

   (a) What is the error rate (risk) of a classifier that picks a label (A,B,C,D) uniformly at random?

   (b) One very simple type of classifier just returns the same label, always. What label should it return, and what will its error rate be?

2. *Discrete and continuous distributions.* In this class, we will deal with both discrete and continuous random variables. Let's look at examples of each.

   (a) A discrete random variable X is said to have Poisson distribution with parameter λ if it can take on values in $\{0, 1, 2, \ldots\}$, with

$$\Pr(X = i) \;=\; e^{-\lambda}\frac{\lambda^i}{i!}.$$

   You can check that these probabilities sum to 1 by looking at the Taylor series for e^λ. Can you give another example of a discrete distribution that assigns positive probabilities to infinitely many values?

   (b) A continuous random variable X has uniform distribution over $[a, b]$ if it has *density function*

$$f(x) \;=\; \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{elsewhere} \end{cases}$$

   This means that the probability that X lies in some interval $[a', b'] \subseteq [a, b]$ is

$$\int_{a'}^{b'} f(x)\, dx.$$

   What is the probability that X is exactly $(a + b)/2$?

3. *Complexity analysis for k-d tree with defeatist search.* Suppose a k-d tree is built on n data points in $\mathbb{R}^d$, by splitting until each leaf node has $\leq n_o$ points.

   (a) What is the time complexity of building the tree, as a function of n, d, and n_o? Justify your answer carefully.

   (b) What is the time complexity of answering a query using defeatist search?

4. *Properties of metrics.* Which of the following distance functions are metrics? In each case, either prove it is a metric or give a counterexample showing that is isn't.

   (a) $d_1 + d_2$, where d_1 and d_2 are each metrics.

   (b) Let's say Σ is a finite set and $\mathcal{X} = \Sigma^m$. The *Hamming distance* on $\mathcal{X}$ is
   $$d(x, y) = \#\text{ of positions on which } x \text{ and } y \text{ differ.}$$

   (c) Squared Euclidean distance on $\mathbb{R}^m$, that is,
   $$d(x, y) = \sum_{i=1}^{m} (x_i - y_i)^2.$$
   (It might be easiest to consider the case $m = 1$.)

5. *A joint distribution over data and labels.* A distribution over two-dimensional data points $X = (X_1, X_2) \in \mathbb{R}^2$ and their labels $Y \in \{0, 1\}$ is specified as follows:

   - The two labels are equally likely, that is, $\Pr(Y = 0) = \Pr(Y = 1) = 1/2$.
   - When $Y = 0$, the points X are uniformly distributed in the square $[-2, -1] \times [-2, -1]$.
   - When $Y = 1$, the points X are uniformly distributed in the square $[1, 3] \times [2, 4]$.

   (a) In a two-dimensional plane, sketch the regions where points (x_1, x_2) might fall. Label one of these regions with $y = 0$ and the other with $y = 1$.

   (b) What is the marginal distribution of X_1? Specify it exactly.

   (c) What is the marginal distribution of X_2?

6. *Two ways of specifying a joint distribution over data and labels.* Consider the following distribution over two-dimensional data points $X = (X_1, X_2)$ and their labels $Y \in \{0, 1\}$:

   - $\Pr(Y = 1) = 1/4$
   - When $Y = 0$, points X are uniformly distributed in the rectangle $[0, 3] \times [0, 1]$.
   - When $Y = 1$, points X are uniformly distributed in the rectangle $[-1, 1] \times [0, 1]$.

   Rewrite this distribution in the form of two functions: μ, the density function for X; and η, the conditional distribution of Y given X.

7. *Bayes optimality.* Consider the following setup:

   - Input space $\mathcal{X} = [-1, 1] \subset \mathbb{R}$.
   - Input distribution: $\mu(x) = |x|$.
   - Label space $\mathcal{Y} = \{0, 1\}$.
   - Conditional probability function
   $$\eta(x) = \Pr(Y = 1 | X = x) = \begin{cases} 0.2 & \text{if } x < -0.5 \\ 0.8 & \text{if } -0.5 \leq x \leq 0.5 \\ 0.4 & \text{if } x > 0.5 \end{cases}$$

   (a) What is the Bayes optimal classifier in this setting? What is the optimal risk R^*?

(b) Suppose we obtain the following training set of four labeled points:

$$(-0.8, 0), (-0.4, 1), (0.2, 1), (0.8, 0).$$

What is the decision boundary of 1-NN using this training set? What is the (true) error rate of this classifier, on the underlying distribution given by μ and η?

(c) In a binary setting, there are two possible errors: $0 \to 1$ (label is 0 but prediction is 1) or $1 \to 0$ (label is 1 but prediction is 0). Suppose these errors have different costs, c_{01} and c_{10}, respectively. We can then define the *cost-sensitive risk* of a classifier $h : \mathcal{X} \to \{0, 1\}$ as

$$R(h) = c_{01}\Pr(Y = 0, h(X) = 1) + c_{10}\Pr(Y = 1, h(X) = 0).$$

In the example above, what is the classifier that minimizes this cost-sensitive risk, if $c_{01} = 1$ and $c_{10} = 0.1$?

(c) Now consider a setting with $\mathcal{Y} = \{0, 1\}$ and with arbitrary $\mathcal{X}, \mu, \eta, c_{01}, c_{10}$. Write down an expression for the classifier with minimum cost-sensitive risk.