CSE 251A: Machine learning **Winter 2021**

Homework 1 — Nearest neighbor and statistical learning

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, regardless of the input. Suppose we would like to pick the most accurate classifier of this type. 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*

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

   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. *Properties of metrics.* Recall that d is a distance metric if and only if it satisfies the following properties:

   (P1) $d(x, y) \ge 0$

   (P2) $d(x, y) = 0 \iff x = y$

   (P3) $d(x, y) = d(y, x)$ (symmetry)

(P4) $d(x,z) \leq d(x,y) + d(y,z)$ (triangle inequality)

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 $m = 1$.)

4. *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?

5. *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.

6. *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

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

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. What would the classifier with the minimum cost-senstive risk look like for the distribution described in the beginning of the problem when $c_{01} = 10$ and $c_{10} = 1$?

7. *Error rate of 1-NN classifier.*

(a) Give an example of a data set with just three points (x, y) for which the 1-NN classifier does *not* have zero training error (that is, it makes mistakes on the training set).

(b) Is 1-NN classification necessarily consistent in cases where the Bayes risk R^* is zero?

8. *Bayes optimality in a multi-class setting.* In lecture, we discussed the setup of statistical learning theory in binary classification. We will now generalize this to situations in which the label space $\mathcal{Y}$ is possibly larger, though still finite.

Suppose $|\mathcal{Y}| = \{1, 2, \ldots, \ell\}$, where $\ell > 2$. We will replace our earlier conditional probability function η by a set of ℓ such functions, denoted $\eta_1, \ldots, \eta_\ell$. Each η_i is a function from $\mathcal{X}$ to $[0, 1]$ and has the following meaning:

$$\eta_i(x) = \Pr(Y = i | X = x).$$

In particular, therefore, $\sum_i \eta_i(x) = 1$ for any x.

What is the Bayes-optimal classifier – that is, the classifier with minimum error – in this case? Specify it precisely, in terms of the η functions.

9. *Classification with an abstain option.* As usual, we can factor a distribution over $\mathcal{X} \times \mathcal{Y}$, with $\mathcal{Y} = \{0, 1\}$, into a marginal distribution μ on $\mathcal{X}$ and a conditional probability function $\eta(x) = \Pr(Y = 1 | X = x)$.

In some situations, it is useful to allow a classifier to *abstain* from predicting on instances x on which it is unsure. Such instances can then be treated separately. Suppose the cost structure is set up so that:

- If the classifier makes a prediction (either 0 or 1), it incurs no cost if the prediction is correct and a cost of 1 if the prediction is wrong.
- If the classifier abstains, it incurs a fixed cost θ, which is some real number between 0 and 1/2.

What classifier $h : \mathcal{X} \to \{0, 1, \text{abstain}\}$ has minimum expected cost? You should write $h(x)$ as a function of $\eta(x)$ and θ.

10. *The statistical learning assumption.* In each of the following cases, say whether or not you feel the statistical learning assumption would hold. If not, explain the nature of the violation (for instance, μ is changing but not η, or η is changing, or the sampling is not independent and random). The answers may be subjective, so explain your position carefully.

(a) A music studio wants to build a classifier that predicts whether a proposed song will be a commercial success. It builds a data set of all the songs it has considered in the past, labeled according to whether or not that song was a hit; and it uses this data to train a classifier.

(b) A bank wants to build a classifier that predicts whether a loan applicant will default or not. It builds a data set based on all loans it accepted over the past ten years, labeled according to whether or not they went into default. These are then used to train the classifier.

(c) An online dating site uses machine learning prediction techniques to decide whether a pair of people are likely to be compatible with each other. Their classifier has worked well on the west coast, and now they decide to take it to the national level.

11. Recall that a classifier is a function that takes a feature vector x in a vector space X and maps it to a discrete label y. Two classifiers C_1 and C_2 are said to be equal if for all $x \in X$, $C_1(x) = C_2(x)$.

    Now suppose we have a training dataset S, and we have two 1-nearest neighbor classifiers C_1 and C_2, both of which are trained on S. State whether C_1 and C_2 are equal in the following two cases, and justify your answer.

    (a) C_1 uses the L_1-distance (to measure distance between training examples and the test point) and C_2 uses the L_2-distance.

    (b) C_1 uses the L_2-distance (to measure distance between training examples and the test point) and C_2 uses the square of the L_2-distance. (Note that the "distance" used in nearest neighbors does not always have to obey the triangle inequality.)