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

Homework 2 — Statistical learning and generative models

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

1. *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?

2. *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.

3. *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.

4. *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.

5. *Conditional probability.* A particular child is always in one of two possible moods: `happy` and `sad`. The prior probabilities of these are:

$$\pi(\texttt{happy}) = \frac{3}{4}, \quad \pi(\texttt{sad}) = \frac{1}{4}.$$

One can usually judge his mood by how talkative he is. After much observation, it has been determined that:

- When he is happy,

$$\Pr(\text{talks a lot}) = \frac{2}{3}, \quad \Pr(\text{talks a little}) = \frac{1}{6}, \quad \Pr(\text{completely silent}) = \frac{1}{6}$$

- When he is sad,

$$\Pr(\text{talks a lot}) = \frac{1}{6}, \quad \Pr(\text{talks a little}) = \frac{1}{6}, \quad \Pr(\text{completely silent}) = \frac{2}{3}$$

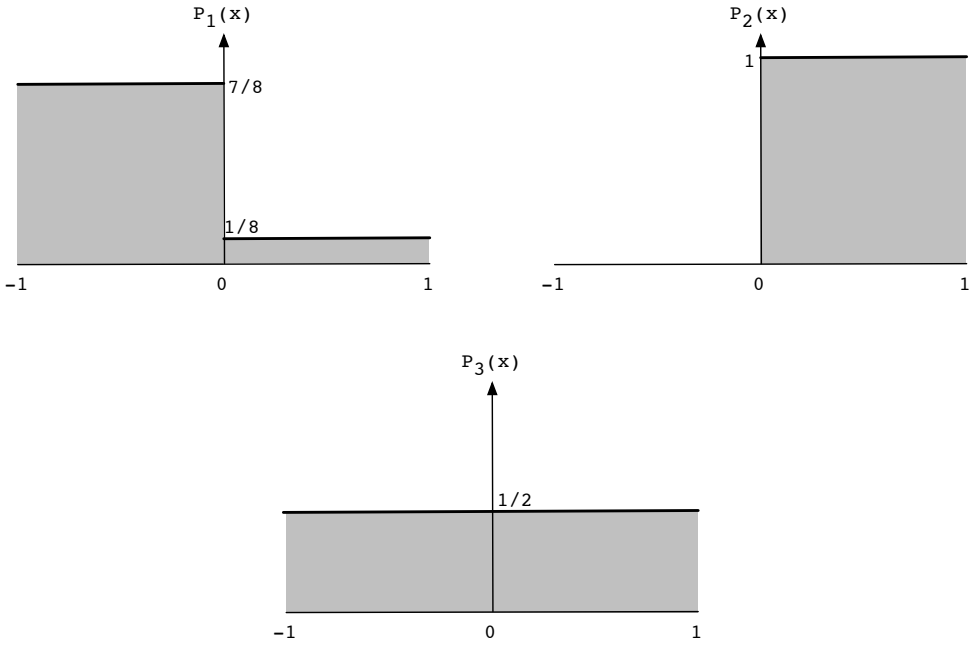
(a) Tonight, the child is just talking a little. What is his most likely mood?

(b) What is the probability of the prediction in part (a) being incorrect?

6. *Bayes optimal classifier.* Suppose $\mathcal{X} = [-1, 1]$ and $\mathcal{Y} = \{1, 2, 3\}$, and that the individual classes have weights

$$\pi_1 = 1/3, \pi_2 = 1/6, \pi_3 = 1/2$$

and densities P_1, P_2, P_3 as shown below.



(a) What is the best (Bayes-optimal) classifier h^*? Specify it exactly, as a function from $\mathcal{X}$ to $\mathcal{Y}$.

(b) What is the error rate of h^*?

7. *Covariance and correlation.* Random variable X has mean zero and standard deviation 10. Random variable Y is defined by $Y = 2X$.

   (a) What is the covariance matrix of the joint distribution of (X, Y)?

   (b) What is the correlation coefficient between X and Y?

8. *Bivariate Gaussians.* Each of the following scenarios describes a joint distribution (x, y). In each case, give the parameters of the (unique) bivariate Gaussian that satisfies these properties.

   (a) x has mean 2 and standard deviation 1, y has mean 2 and standard deviation 0.5, and the correlation between x and y is -0.5.

   (b) x has mean 1 and standard deviation 1, and y is equal to x.

9. *More bivariate Gaussians.* Roughly sketch the shapes of the following Gaussians $N(\mu, \Sigma)$. For each, you only need to show a representative contour line which is qualitatively accurate (has approximately the right orientation, for instance).

   (a) $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix}$

   (b) $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix}$

   Now check your sketches by using your computer to plot 100 random samples from each of these Gaussians.

10. *Qualitative appraisal of Gaussian parameters.* A Gaussian over $\mathbb{R}^2$ has covariance matrix $\begin{pmatrix} a & b \\ b & c \end{pmatrix}$. Give precise characterizations, in terms of a, b, c, of when the following are true.

    (a) The two variables are negatively correlated.

    (b) The two variables are uncorrelated.

    (c) One variable is a linear function of the other.

    (d) One of the variables is a constant.