CSE 151A: Machine learning

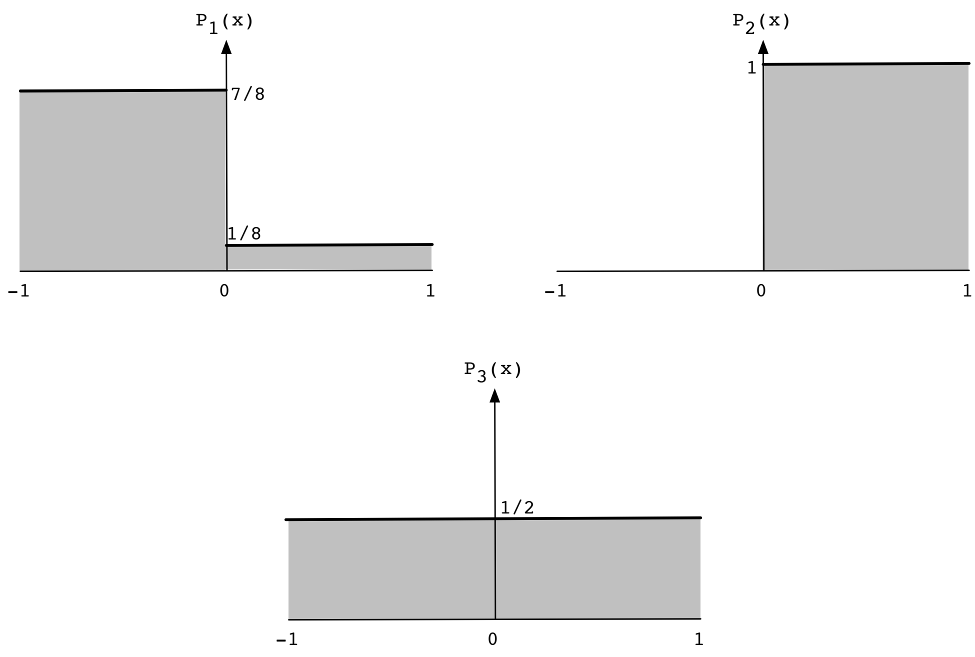
Homework 4

Instructions:

- You may discuss problems with your study group, but ultimately all your work (mathematical problems, code, experimental details) must be individual.
- Your solutions must be `typed up` and uploaded to Gradescope by 9.59PM on Thursday October 26. No late homeworks will be accepted under any circumstances, so you are encouraged to upload early.
- A subset of the problems will be graded.

Conceptual and mathematical problems

1. *A simple three-class scenario.* Suppose $\mathcal{X} = [-1, 1]$ and $\mathcal{Y} = \{1, 2, 3\}$, and that the individual classes have weights $\pi_1 = \frac{1}{3}, \pi_2 = \frac{1}{6}, \pi_3 = \frac{1}{2}$ and densities P_1, P_2, P_3 as shown below.



   What is the optimal classifier h^*? Specify it as a function from $\mathcal{X}$ to $\mathcal{Y}$. Don't worry about the specific prediction at $x = 0$.

2. Suppose we solve a classification problem with k classes by using a Gaussian generative model in which the jth class is specified by parameters π_j, μ_j, Σ_j. In each of the following situations, say whether the decision boundary is `linear`, `spherical`, or `other quadratic`.

(a) We compute the empirical covariance matrices of each of the k classes, and then set $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$ to the **average** of these matrices.

(b) The covariance matrices Σ_j are all **diagonal**, but no two of them are the same.

(c) There are two classes (that is, $k = 2$) and the covariance matrices Σ_1 and Σ_2 are multiples of the identity matrix.

3. *Example of regression with one predictor variable.* Consider the following simple data set of four points (x, y):

$$(1,1), (1,3), (4,4), (4,6).$$

   (a) Suppose you had to predict y without knowledge of x. What value would you predict? What would be its mean squared error (MSE) on these four points?

   (b) Now let's say you want to predict y based on x. What is the MSE of the linear function $y = x$ on these four points?

   (c) Find the line $y = ax + b$ that minimizes the MSE on these points. What is its MSE?

4. *Optimality of the mean.* One fact that we used implicitly in the lecture is the following:

   > If we want to summarize a bunch of numbers $x_1, \ldots, x_n$ by a single number s, the best choice for s, the one that minimizes the average squared error, is the **mean** of the x_i's.

   Let's see why this is true. We begin by defining a suitable loss function. Any value $s \in \mathbb{R}$ induces a mean squared loss (MSE) given by:

$$L(s) = \frac{1}{n} \sum_{i=1}^{n} (x_i - s)^2.$$

   We want to find the s that minimizes this function.

   (a) Compute the derivative of $L(s)$.

   (b) What value of s is obtained by setting the derivative dL/ds to zero?

5. We have a data set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$, where $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \mathbb{R}$. Suppose that we want to express y as a linear function of x, but the error penalty we have in mind is not the usual squared loss: if we predict $\widehat{y}$ and the true value is y, then the penalty should be the absolute difference, $|y - \widehat{y}|$. Write down the loss function that corresponds to the total penalty on the training set.

6. *Writing expressions in matrix-vector form.* Let $x^{(1)}, \ldots, x^{(n)}$ be a set of n data points in $\mathbb{R}^d$, and let $y^{(1)}, \ldots, y^{(n)} \in \mathbb{R}$ be corresponding response values. In this problem, we will see how to rewrite several basic functions of the data using matrix-vector calculations. To this end, define:

   - X, the $n \times d$ matrix whose rows are the $x^{(i)}$
   - y, the n-dimensional vector with entries $y^{(i)}$
   - $\mathbf{1}$, the n-dimensional vector whose entries are all 1

   Each of the following quantities can be expressed in the form cAB, where c is some constant, and A, B are matrices/vectors from the list above (or their transposes). In each case, give the expression.

   (a) The average of the $y^{(i)}$ values, that is, $(y^{(1)} + \cdots + y^{(n)})/n$.

   (b) The $n \times n$ matrix whose (i,j) entry is the dot product $x^{(i)} \cdot x^{(j)}$.

(c) The average of the $x^{(i)}$ vectors, that is, $(x^{(1)} + \cdots + x^{(n)})/n$.

(d) The empirical covariance matrix, assuming the points $x^{(i)}$ are centered (that is, assuming the average of the $x^{(i)}$ vectors is zero). This is the $d \times d$ matrix whose (i, j) entry is

$$\frac{1}{n} \sum_{k=1}^{n} x_i^{(k)} x_j^{(k)}.$$

7. In lecture, we asserted that in d-dimensional space, it is possible to perfectly fit (almost) any set of $d+1$ points $(x^{(0)}, y^{(0)}), (x^{(1)}, y^{(1)}), \ldots, (x^{(d)}, y^{(d)})$. Let's see how this works in the specific case where:

   - $x^{(0)} = 0$
   - $x^{(i)}$ is the ith coordinate vector (the vector that has a 1 in position i, and zeros everywhere else), for $i = 1, \ldots, d$
   - $y^{(i)} = c_i$, where $c_0, c_1, \ldots, c_d$ are arbitrary constants.

   Find $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $w \cdot x^{(i)} + b = y^{(i)}$ for all i. You should express your answer in terms of $c_0, c_1, \ldots, c_d$.

8. Keep the same set of $d+1$ points $(x^{(0)}, y^{(0)}), (x^{(1)}, y^{(1)}), \ldots, (x^{(d)}, y^{(d)})$ from the previous problem. As we saw, we can find w, b that perfectly fit these points; hence least-squares regression would find this "perfect" solution and have zero loss on the training set.

   Now, let us instead use ridge regression, with parameter $\lambda \geq 0$, to obtain a solution. We can denote this solution by w_λ, b_λ. Also define the squared training loss associated with this solution,

$$L(\lambda) = \sum_{i=0}^{d} (y^{(i)} - (w_\lambda \cdot x^{(i)} + b_\lambda))^2.$$

   (a) What is $L(0)$?

   (b) As λ increases, how does $\|w_\lambda\|$ behave? Does it increase, decrease, or stay the same?

   (c) As λ increases, how does $L(\lambda)$ behave? Does it increase, decrease, or stay the same?

   (d) As λ goes to infinity, what value does $L(\lambda)$ approach? Your answer should be in terms of the coefficients c_i.

Programming problems

9. *Discovering relevant features in regression.* The data file `mystery.dat` contains pairs (x, y), where $x \in \mathbb{R}^{100}$ and $y \in \mathbb{R}$. There is one data point per line, with comma-separated values; the very last number in each line is the y-value.

   In this data set, y is a linear function of just *ten* of the features in x, plus some noise. Your job is to identify these ten features.

   (a) Explain your strategy in one or two sentences. Hint: you will find it helpful to look over the routines in `sklearn.linear_model`.

   (b) Which ten features did you identify? You need only give their coordinate numbers, from 1 to 100.