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 11.59PM on Thursday May 1. 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. *Identical spherical Gaussians.* Suppose we have a classification task where the data lies in $\mathcal{X} = \mathbb{R}^d$ and there are two classes, $\mathcal{Y} = \{1, 2\}$. We use a Gaussian generative model and fit the data from each class with *spherical Gaussians having the same covariance.* So, class 1 has weight π_1 and density $N(\mu_1, \sigma^2 I_d)$ while class 2 has weight π_2 and density $N(\mu_2, \sigma^2 I_d)$.

   The decision boundary in this case is *linear*, of the form $w^T x = b$. Obtain precise expressions for w and b in terms of the model parameters $\mu_1, \mu_2, \sigma, \pi_1, \pi_2$. (When calculating this, recall that if u, v are vectors then $\|u + v\|^2 = \|u\|^2 + \|v\|^2 + 2u \cdot v$.)

2. *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?

3. *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?

4. 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 squared loss*: if we predict $\widehat{y}$ and the true value is y, then we want the penalty to be the *absolute difference* $|y - \widehat{y}|$. Write down the loss function $L(w, b)$ that corresponds to the total penalty on the training set.

5. *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

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

Programming problems

Before beginning these problems, download `hw4.zip` from Piazza and uncompress it.

6. *Experiments with Gaussian generative models.* Look through the provided notebook `gaussian-generative.ipynb`. It takes a given Gaussian generative model in $\mathbb{R}^2$ and plots the decision boundary as well as a few points sampled from the model. Look through the code to understand what it is doing.

   The initial code sets the covariance matrices of each class to the identity; as a result, the shown boundary is linear. Play around with the covariance matrices (*while leaving the means fixed*) to get other types of boundary. Show an example of each of the following:

   - A linear boundary that is not parallel to the coordinate axes
   - A spherical boundary
   - A boundary that is either elliptical or parabolic

   (And if you have time, can you get a hyperbolic boundary? This is not for turning in.)

7. *Experiments with least-squares regression.* Look through the notebook `california-housing.ipynb`. It loads a data set of over 20,000 points where each data point has 8 predictor variables x (describing characteristics of a locality) and one response variable y (the median house value in that locality).

   (a) Start working through the notebook. When you get to the correlation matrix between the features, answer the following questions:

- Which other feature is most highly correlated with `MedHouseVal`?
- Which pair of features are most positively correlated?
- Which pair of features are most negatively correlated?

(b) Now continue to the section where the training and test sets are created. Then answer the following questions:

- How many points are there in the training set? The test set?
- If we were to predict y *without looking at* x, what would be the single best value to predict for the test set? What would be the resulting MSE on the test set?

(c) Continuing through the notebook, fit a linear regressor to the training data using least-squares. Give the coefficients of the linear model (make sure you indicate which feature corresponds to each coefficient). Also give the mean-squared error (MSE) on the test data.

(d) Next, fit a linear regressor using just the two features `Latitude` and `Longitude`. What is the MSE in this case?

(e) Suppose we want a linear model that is based on a single predictor variable. Which variable is the best one to choose? What is the resulting MSE?

8. *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.