CSE 151: Machine learning	**Winter 2019**

Homework 5

Submission instructions:

- Please type up your solutions.
- If a problem asks for a numerical answer, you need only provide this answer. There is no need to show your work, unless you would like to.
- Upload the PDF file for your homework to `gradescope` by 6pm on Tuesday February 12.

Part A: Regression and logistic regression

1. *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

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

2. 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 $c_0, c_1, \ldots, c_d$.

3. 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? You ranswer should be in terms of the coefficients c_i.

4. *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.

5. We identified *inherent uncertainty* as one reason why it might be difficult to get perfect classifiers, even with a lot of training data. In which of the following situations is there likely to be a significant amount of inherent uncertainty?

   (a) x is a picture of an animal and y is the name of the animal
   (b) x consists of the dating profiles of two people and y is whether they will be interested in each other
   (c) x is a speech recording and y is the transcription of the speech into words
   (d) x is the recording of a new song and y is whether it will be a big hit

6. A logistic regression model given by parameters $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ is fit to a data set of points $x \in \mathbb{R}^d$ with binary labels $y \in \{-1, 1\}$. Write down a precise expression for the set of points x with

   (a) $\Pr(y = 1|x) = 1/2$
   (b) $\Pr(y = 1|x) = 3/4$
   (c) $\Pr(y = 1|x) = 1/4$

7. Suppose that in a bag-of-words representation, we decide to use the following vocabulary of four words: (`is`, `flower`, `rose`, `an`, `a`). What is the vector form of the sentence "A rose is a rose is a rose"?

8. When using a logistic regression model with two labels, define the *margin* on a point x to be how far its conditional probability is from 1/2:
$$\text{margin}(x) = \left| \Pr(y = 1|x) - \frac{1}{2} \right|.$$
This is a number in the range $[0, 1/2]$.
For any $m \in [0, 1/2]$, define the following two quantities based on a **test set**:
   - $f(m)$: the fraction of test points that have margin $\geq m$
   - $e(m)$: the error rate on test points with margin $\geq m$

   As m grows, how will $f(m)$ and $e(m)$ behave? Would we expect them to increase/decrease? Will they necessarily increase/decrease?

Part B: Unconstrained optimization

1. We are given a set of data points $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$, and we want to find a single point $z \in \mathbb{R}^d$ that minimizes the loss function
$$L(z) = \sum_{i=1}^{n} \|x^{(i)} - z\|^2.$$
Use calculus to determine z, in terms of the $x^{(i)}$.

2. Given a set of data points $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$, we want to find the vector $w \in \mathbb{R}^d$ that minimizes this loss function:
$$L(w) = \sum_{i=1}^{n} (w \cdot x^{(i)}) + \frac{1}{2} c \, \|w\|^2.$$
Here $c > 0$ is some constant.
   (a) What is $\nabla L(w)$?
   (b) What value of w minimizes $L(w)$?

3. Consider the following loss function on vectors $w \in \mathbb{R}^4$:
$$L(w) = w_1^2 + 2w_2^2 + w_3^2 - 2w_3w_4 + w_4^2 + 2w_1 - 4w_2 + 4.$$
   (a) What is $\nabla L(w)$?
   (b) Suppose we use gradient descent to minimize this function, and that the current estimate is $w = (0, 0, 0, 0)$. If the step size is η, what is the next estimate?
   (c) What is the minimum value of $L(w)$?
   (d) Is there is a unique solution w at which this minimum is realized?

4. Consider the loss function for ridge regression (ignoring the intercept term):
$$L(w) = \sum_{i=1}^{n} (y^{(i)} - w \cdot x^{(i)})^2 \; + \; \lambda \|w\|^2$$
where $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \mathbb{R}$ are the data points and $w \in \mathbb{R}^d$. There is a closed-form equation for the optimal w (as we saw in class), but suppose that we decide instead to minimize the function using local search.

(a) What is $\nabla L(w)$?

(b) Write down the update step for gradient descent.

(c) Write down a stochastic gradient descent algorithm.