**CSE 250B: Machine learning** **Winter 2019**

# Homework 3 — Regression, logistic regression, unconstrained optimization

1. *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?

2. *Lines through the origin.* Suppose that we have data points $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$, where $x^{(i)}, y^{(i)} \in \mathbb{R}$, and that we want to fit them with a line that passes through the origin. The general form of such a line is $y = ax$: that is, the sole parameter is $a \in \mathbb{R}$.
   (a) The goal is to find the value of $a$ that minimizes the squared error on the data. Write down the corresponding loss function.
   (b) Using calculus, find the optimal setting of $a$.

3. Suppose that $y = x_1 + x_2 + \cdots + x_{10}$, where:
   - $x_1, \ldots, x_{10}$ are independent, and
   - the $x_i$ each have a Gaussian distribution with mean 1 and variance 1.

   (a) We wish to express $y$ as a linear function of just $x_1, \ldots, x_5$. What is the linear function that minimizes MSE?
   (b) What is the mean squared error of the function in (a)?

4. We have a data set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$, where $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \mathbb{R}$. 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.

5. We have $n$ data points in $\mathbb{R}^d$ and we want to compute all pairwise dot products between them. Show that this can be achieved by a *single* matrix multiplication.

6. *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.

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

7. 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) $\Pr(y = 1|x) = 1/4$

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

9. 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 $x^{(i)}$.

10. 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 $\eta$, 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?

11. 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.