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

Homework 2 — Regression, logistic regression, unconstrained optimization

This homework is not meant to be turned in. Try it on your own, and compare your answers to the solution set that will be released on Tuesday January 29.

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. 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.

4. 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.

5. *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.

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

7. *Minimizing absolute loss.* Given real values $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}$, we want to find the value v that minimizes the absolute loss

$$L(v) = \sum_{i=1}^{n} |x^{(i)} - v|.$$

What value of v achieves this? Justify your answer.

8. 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?

9. 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.

10. *Closed-form solution for ridge regression.* Consider the ridge regression loss function: for $w \in \mathbb{R}^d$,

$$L(w) = \sum_{i=1}^{n} (y^{(i)} - w \cdot x^{(i)})^2 \; + \; \lambda \|w\|^2.$$

(As usual, we need not worry about the b.) Let X denote the $n \times d$ matrix whose rows are the $x^{(i)}$ and let y be an n-dimensional vector whose entries are the $y^{(i)}$.

   (a) Rewrite $L(w)$ in matrix-vector form using X and y.

   (b) Derive a closed-form solution for the minimizer of $L(w)$.

11. *A case when Lasso finds a sparse solution.* For a given data set, let the least-squares cost function be written as $LS(w) = \|y - Xw\|^2$, where X is a matrix with one data point per row, y is a vector with one response value per row, and the intercept term is omitted.

   (a) As we saw in class, the optimal least-squares solution is $w^* = (X^T X)^{-1} X^T y$. Show that for any other vector $w \in \mathbb{R}^d$, the least-squares cost can be expressed as follows:

$$LS(w) = LS(w^*) + (w - w^*)^T X^T X (w - w^*).$$

   That is, it is the cost of w^* plus a suitable Mahalanobis distance between w and w^*. *Hint:* Write $\|y - Xw\|^2$ as $\|(y - Xw^*) + (X(w^* - w))\|^2$ and show that when this square is expanded, the cross-term is zero.

(b) Let LS^* denote $LS(w^*)$, where w^* is the optimal least-squares solution. The loss function for the Lasso is then

$$L(w) = LS^* + (w - w^*)^T X^T X (w - w^*) + \lambda \|w\|_1.$$

When minimizing this, we can ignore LS^* (which is a constant). This is equivalent to solving

$$\begin{aligned} \min \quad & (w - w^*)^T X^T X (w - w^*) \\ & \|w\|_1 \leq B \end{aligned}$$

for some value of B that depends on λ. Assume for simplicity that $X^T X = I$ and $B = 1$. Suppose w^* lies outside the ℓ_1 unit ball. Give an example of a two-dimensional ($d = 2$) situation where the solution to the above optimization is sparse, and another example where it is not sparse. In either case, make a diagram that includes the ℓ_1 unit ball and the location of w^*.

12. *Form of the squashing function.* For $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{1, 2\}$, consider a distribution over $\mathcal{X} \times \mathcal{Y}$ of the following form:

- $\Pr(y = 1) = \Pr(y = 2) = 1/2$
- The distribution of x given $y = 1$ is a spherical Gaussian $N(\mu_1, \sigma^2 I_d)$ and the distribution of x given $y = 2$ is $N(\mu_2, \sigma^2 I_d)$. Recall that the density of $N(\mu, \sigma^2 I_d)$ is given by

$$p(x) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left(-\frac{\|x - \mu\|^2}{2\sigma^2}\right).$$

Derive a closed-form formula for $\Pr(y = 1|x)$. How does it relate to the squashing function?