Homework 1

CSE 4820/5819, Spring 2021

Released: January 28, 2021
Due: February 16, 2021

Note: You may use any programming language and plotting software to complete Exercise 2. However, Python is recommended. **Please submit your code for full credit (answers and plots without code will only receive half credit).**

Exercise 1 [40 points]. Mathematical Background

1. (5 points) Let $x \in \mathbb{R}^n$ and $A, B \in \mathbb{R}^{n \times n}$. What is an equivalent expression to $(x^T AB)^T$?
2. (10 points) Consider the set of points $\{x \in \mathbb{R}^2 : \|x\|_p = 1\}$.
   (a) Plot this set of points for $p = 2$. This is the unit sphere in L^2.
   (b) On the same figure, plot the unit sphere in L^1.
   (c) Again on the same plot, do the same for $p = \infty$.
   (d) What do you notice about the unit spheres as p increases from 1 to ∞?
3. (10 points) Let $x, y \in \mathbb{R}^n$.
   (a) Express $\|x - y\|_2^2$ using the dot product.
   (b) Simplify your answer above as the sum of three terms.
4. (10 points) Consider the multivariate quadratic function

$$f(x) = \frac{1}{2} x^T \begin{bmatrix} 1 & 4 \\ 0 & 1 \end{bmatrix} x + \begin{bmatrix} 1 \\ 2 \end{bmatrix}^T x + 1$$

   (a) What is the gradient ∇f?
   (b) What is the Hessian $\nabla^2 f$?
5. (5 points) Suppose we are solving a minimization problem. In gradient descent, at every iteration we need to calculate the gradient ∇f evaluated at x_n in order to compute the next guess $x_{n+1} = x_n - \alpha \nabla f(x_n)$. When using Newton's method to minimize a function f by finding its critical points where $\nabla f = 0$, at every iteration in addition to $\nabla f(x_n)$ what additional information about f at x_n do we need to calculate in order to compute x_{n+1}?

Exercise 2 [60 points]. Logistic Regression

1. (10 points) Recall that in logistic regression, we have the model

$$P(y = 1|x; \theta) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

   (a) Suppose $x, \theta \in \mathbb{R}$. Plot $h_\theta(x) = g(\theta^T x)$ for $\theta = 1$, 2, and 5 on the same plot. How does the shape of g change as θ increases? What is the shape of $h_\theta(x) = g(\theta^T x)$ in the limit $\theta \to \infty$?

   (b) When the training data are linearly separable (there exists a line/hyperplane that perfectly separates the labeled data into different classes), we maximize the likelihood of the data by choosing θ so that any positive distance from the decision boundary corresponds to $P(y = 1|x; \theta) = 1$ and any negative distance corresponds to $P(y = 1|x; \theta) = 0$. Another way to say this is that we want to find θ such that $g(\theta^T x) = 1$ whenever $\theta^T x > 0$, and $g(\theta^T x) = 0$ whenever $\theta^T x < 0$. What θ gives this model? (Hint: see plots above.)

   - **Note this shows logistic regression does not converge when the training data are linearly separable.** In practice, data are rarely linearly separable, but there are also different ways to modify logistic regression to take the separable case into account, which we will return to when we discuss regularization.

2. (20 points) Recall that for m datapoints $(x^{(i)}, y^{(i)})$, where $x \in \mathbb{R}^n$ is the feature vector, $y \in \{0, 1\}$ is the class label, and $i = 1, ..., m$ is the index of the datapoint, the log likelihood for logistic regression is

$$\ell(\theta) = \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))$$

   where $h_\theta(x^{(i)}) = g(\theta^T x^{(i)})$.

   (a) Let $m = 1$ so that we consider only the log likelihood of the single datapoint (x, y). Use the known fact that $g'(z) = g(z)(1 - g(z))$ to show that

$$\frac{\partial}{\partial \theta_j} \ell(\theta) = (y - h_\theta(x)) x_j$$

   Note that this gives us the j^{th} element of the gradient $\nabla_\theta \ell$.

   (b) What is the expression for the gradient for arbitrary m?

3. (30 points) Download the files `xvals.dat` and `yvals.dat` included in the homework. The first contains the features and the second contains the labels for 99 datapoints.

   (a) Save the data from `xvals.dat` in a matrix, X. Each row of `xvals.dat` corresponds to a datapoint, and each column corresponds to a feature. In this dataset, how many features does each datapoint have?

(b) Something we did not go into detail in class is that in general, the decision boundary separating the two classes does not pass through the origin. Therefore, the linear model $\theta^T x = 0$ for the decision boundary actually implicitly includes an intercept term θ_0. Let $[x_1, x_2]$ be the feature vector describing a given datapoint. Then in logistic regression, we add a constant feature 1 to the feature vector in order to get an intercept term:

$$\theta^T x = \begin{bmatrix} \theta_0 & \theta_1 & \theta_2 \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

Add this constant feature to your matrix X.

(c) Use your answer from part 2(a) above to write code for finding θ^* that maximizes the log likelihood using **gradient ascent**. Start from an initial guess of $\theta = 0$ (the vector of all zeros), and stop when $\|\nabla_\theta \ell\|_2 \leq 10^{-6}$ (the tolerance we're using to approximate the condition $\nabla_\theta \ell = 0$).

i. Use a learning rate of $\alpha = 0.01$. How many steps did you need to take?

ii. Now try using $\alpha = 0.1$. What happens?

iii. How many steps do you need for convergence when $\alpha = 0.001$?

(d) Plot the data, using different colors for each label, and the decision boundary that you found in part 3(c) above. Recall the decision boundary is the line where $\theta_0^* + \theta_1^* x_1 + \theta_2^* x_2 = 0$.

(e) Now let's make a prediction on a new datapoint. According to your trained model in part 3(c) above, what is $P(y = 1 | x = \begin{bmatrix} 2 & 1 \end{bmatrix}^T; \theta^*)$? On the other hand, what is $P(y = 0 | x = \begin{bmatrix} 2 & 1 \end{bmatrix}^T; \theta^*)$? Therefore, what class label y do we predict for $x = \begin{bmatrix} 2 & 1 \end{bmatrix}^T$?