CS 4780/5780 Homework 6

Due: Tuesday 03/20/18 11:55pm on Gradescope

Problem 1: Optimization with Gradient Descent

(a) You have a univariate function you wish to minimize, $f(w) = 5(w - 11)^4$. Suppose you wish to perform gradient descent with constant step size $\alpha = 1/40$. Starting with $w_0 = 13$, perform 5 steps of gradient descent. What are $w_0, ..., w_5$? What is the value of $f(w_5)$?

(b) With the same function and starting point above, perform gradient descent until convergence, with the constant step size $\alpha = 1/80$. What are all of the $w_0, w_1, ...$ and all of the $f(w_0), f(w_1), ...$?

(c) Sketch a graph of the function f and draw how you expect the gradient descent algorithm will converge if we set the constant learning rate to $\alpha = 1/20$ or $\alpha = 1/160$.

Problem 2: Linear Regression

(a) Consider we have the following 1-d training set:

x	y
-2	7
-1	4
0	3
1	4
2	7

and our goal is to find a regression model that could regress x to our target value y. To do this, we are going to use a linear regression model. Namely, we are going to model our data by assuming the relationship

$$\begin{aligned} y &= w_1 x + w_0 + \epsilon \\ &= \vec{w}^T \phi(x) + \epsilon \end{aligned}$$

where $\phi(x) = [1, x]^T$. We call ϕ a feature mapping of x and this feature mapping allows us to absorb the bias w_0 into the vector $\vec{w}$.

(1) With this feature mapping, we can write down the design matrix as

$$X = [\phi(x_1)...\phi(x_n)]$$

Using the formula given in class, compute the closed form solution for $\vec{w}$. Even though you are not required to calculate the inverse by hand, we strongly encourage you to do so since we expect you to be able to calculate the inverse of a 2×2 and 3×3 matrix by hand.

(2) Recall that the loss function for linear regression is

$$\ell(\vec{w}) = \sum_{i=1}^{n} (y_i - \vec{w}^T \phi(x_i))^2$$

With the closed formed solution obtained in (a)(1), calculate the training loss.

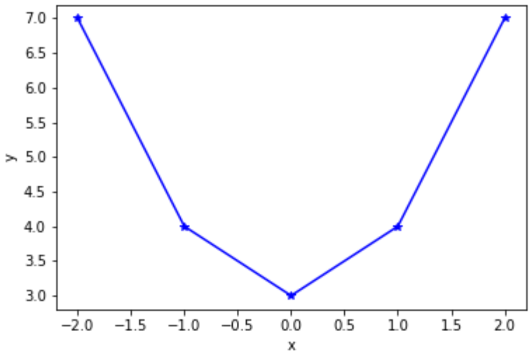

Figure 1: Plot of y against x

(b) In (a)(2), we realize that we could not attain a training loss of zero. To investigate this, we plot the data in Figure 1. It seems like the y is a quadratic function of x, instead of a linear function of x. Does it mean that we have no hope in getting a good linear regression model? The answer is no. We just need to be a little smart about our feature mapping ϕ. Now, let us define a new feature mapping $\phi_2(x) = [1, x, x^2]$. With this new feature mapping, what is the closed form solution $\vec{w}$? Can we attain a training loss of zero with this closed form solution?

Take-away: Feature extraction is essential in creating a powerful model. Although we will not cover how to design hand-crafted features, we will cover how to automatically learn a powerful representation of the data using kernel methods and possibly deep learning if time permits.

Problem 3: Weighted Ridge Regression

Suppose we have a dataset $\{(\vec{x}_1, y_1), ..., (\vec{x}_n, y_n)\}$. In a ridge regression setting, we assume that each example is equally important. Sometimes, certain examples are more important than others and we would want to assign a positive weight p_i to each training examples to indicate the level of importance of each training example. For instance, in the previous problem, if we care more about the predictions of our model on the examples with positive x, then we might assign higher weights to those training examples with positive x than those training examples with negative x.

In order to modify ridge regression to account for the weights of different training examples, we rewrite the loss function as

$$\ell(\vec{w}) = \sum_{i=1}^{n} (p_i(\vec{w}^T \vec{x}_i - y_i)^2) + \lambda \vec{w}^T \vec{w}$$

1. Suppose $X = [\vec{x}_1, ..., \vec{x}_n]^T$, $\vec{y} = [y_1, ..., y_n]^T$. Find a diagonal matrix P such that we can rewrite the loss function as

$$\ell(\vec{w}) = (X\vec{w} - \vec{y})^T P (X\vec{w} - \vec{y}) + \lambda \vec{w}^T \vec{w}$$

2. Using the rewritten loss function, derive a closed form solution for $\vec{w}$ by setting the gradient of the loss function equal to zero.
3. Show that the loss function is convex by showing the $\nabla^2 \ell$ is positive semidefinite for any $\vec{w}$. By showing the loss function is convex, we can conclude that the closed formed we derived in (2) is indeed a global minimum point.