CS 229, Spring 2023
Section #1 Solutions: Linear Algebra, Least Squares, and Logistic Regression

---

1. **Least Squares Regression**

   Many supervised machine learning problems can be cast as optimization problems in which we either define a cost function that we attempt to minimize or a likelihood function we attempt to maximize. These functions are often called *Objective Functions*. Assuming you successfully defined an objective function that is either convex (to minimize) or concave (to maximize), you can find the optimal point with either of the following approaches:

   (a) Find a closed form solution for setting the gradient equal to 0 (i.e. $\nabla_\theta J(\theta) = 0$)

   (b) Find the gradient of the objective function w.r.t. the parameters and do gradient descent.

   Most of the time, finding a closed form solution for $\nabla_\theta J(\theta) = 0$ is impossible, so we attempt to use gradient descent instead.

   (a) Here, let us consider the original least-squared regression problem:

$$
\begin{aligned}
J(\theta) &= \frac{1}{2}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})^2 \\
&= \frac{1}{2}(X\theta - \vec{y})^T(X\theta - \vec{y})
\end{aligned}
$$

   where X is the design matrix with each row as a example in our data, θ are the parameters, and $\vec{y}$ is the vector of ground truth values we want to predict.
   Here are some useful formulas:

$$
\begin{aligned}
\frac{\partial x^T A x}{\partial x} &= (A + A^T)x \\
\frac{\partial x^T y}{\partial x} &= \frac{\partial y^T x}{\partial x} = y
\end{aligned}
$$

   i. Derive the gradient $\nabla_\theta J(\theta)$

Answer:

$$
\begin{aligned}
J(\theta) &= \frac{1}{2}(X\theta - \vec{y})^T(X\theta - \vec{y}) \\
&= \frac{1}{2}(\theta^T X^T - \vec{y}^T)(X\theta - \vec{y}) \\
&= \frac{1}{2}(\theta^T X^T X\theta - \vec{y}^T X\theta - \theta^T X^T\vec{y} + \vec{y}^T\vec{y}) \\
&= \frac{1}{2}(\theta^T X^T X\theta - 2\theta^T X^T\vec{y} + \vec{y}^T\vec{y}) \\
\nabla_\theta J(\theta) &= \frac{1}{2}[(X^T X + X^T X)\theta - 2X^T\vec{y}] \\
&= \frac{1}{2}[2X^T X\theta - 2X^T\vec{y}] \\
&= X^T X\theta - X^T\vec{y}
\end{aligned}
$$

This solution may be used to perform gradient descent on the least squares objective with the formula

$$\theta^{(t+1)} := \theta^{(t)} - \alpha\nabla_\theta J(\theta)$$

or to find a closed form solution (see part ii).

ii. Find a closed form solution for θ^* (the parameters that minimize the loss function). You may assume that $X^T X$ is invertible.
Answer:

$$
\begin{aligned}
\nabla_\theta J(\theta) &= 0 \\
X^T X\theta^* - X^T y &= 0 \\
X^T X\theta^* &= X^T y \\
\theta^* &= (X^T X)^{-1}X^T y
\end{aligned}
$$

(Optional) As mentioned in lecture, $X^T X$ is invertible if and only if X is both full rank and $n \geq d$ (X is "skinny"). This is not the point of our discussion of least squares so you may assume that $X^T X$ is invertible if you are not familiar with this terminology.

2. **Logistic regression and Classification**

First, let's review logistic regression. The objective function of logistic regression is

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)}\log(\sigma(\theta^T x^{(i)})) + (1-y^{(i)})\log(1-\sigma(\theta^T x^{(i)}))$$

Where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function.

(a) As a review, please derive the gradient of the objective function w.r.t. the parameters θ ($\nabla_\theta J(\theta)$) and write out the gradient descent update formula.

Answer: We know that $\frac{\partial \sigma(x)}{\partial x} = \sigma(x)(1-\sigma(x))$

So we have

$$\begin{aligned}
\nabla_\theta J(\theta) &= -\frac{1}{m}\sum_{i=1}^{m} y^{(i)}\frac{1}{\sigma(\theta^T x^{(i)})}\sigma(\theta^T x^{(i)})(1-\sigma(\theta^T x^{(i)})x^{(i)} - (1-y^{(i)})\frac{1}{1-\sigma(\theta^T x^{(i)})}\sigma(\theta^T x^{(i)})(1-\sigma(\theta^T x^{(i)}))x \\
&= -\frac{1}{m}\sum_{i=1}^{m} y^{(i)}(1-\sigma(\theta^T x^{(i)}))x^{(i)} - (1-y)\sigma(\theta^T x^{(i)})x^{(i)} \\
&= -\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - \sigma(\theta^T x^{(i)}))x^{(i)}
\end{aligned}$$

Plugging this into the gradient descent update formula yields

$$\begin{aligned}
\theta^{(t+1)} &:= \theta^{(t)} - \alpha\nabla_\theta J(\theta) \\
&:= \theta^{(t)} + \alpha\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - \sigma(\theta^T x^{(i)}))x^{(i)}
\end{aligned}$$

(b) (PSET question) Prove that the objective function is convex by showing the Hessian matrix is positive semi-definite. That is, show that $z^T H z \geq 0$ for all z

Answer: Discussed hessian/basic setup but no solution provided (on HW).

(c) Consider a problem where we are given labels that are either 1 or -1 instead of 1 or 0. How would you be able to cast this problem as a logistic regression problem with sigmoid activation?

Answer: There are many possible answers. One approach is to simply re-scale all the labels by using $(y+1)/2$. Another approach is to modify the objective function as follow:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\log(\sigma(y^{(i)}\theta^T x^{(i)}))$$

3. **Basic probability review**

   Bayes rule is defined as follows:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

   Show the following is true:

$$P(Y|X,E) = \frac{P(X,Y|E)}{P(X|E)}$$

   Answer:

$$\begin{aligned} P(Y|X,E) &= \frac{P(Y,X,E)}{P(X,E)} \\ &= \frac{P(Y,X|E)P(E)}{P(X|E)P(E)} \\ &= \frac{P(Y,X|E)}{P(X|E)} \\ &= \frac{P(X|Y,E)P(Y|E)}{P(X|E)} \end{aligned}$$