CS 229, Winter 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. **MLE Estimation of Gaussian Covariance Matrices**

The aim of this problem is to 1) practice taking the gradient of functions with respect to matrices and 2) consider a particular gradient that you will encounter later in the course with topics like Gaussian Discriminant Analysis and Gaussian Mixture Models. We would like to estimate the parameters of a Gaussian distribution:

$$p(x) = \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

In particular, we will consider the maximum liklihood estimation of the covariance matrix Σ given some data points $\{x^{(1)}, ..., x^{(n)}\}$.

(a) Let's begin by practicing the process of taking the gradient of a function with respect to a matrix. Derive an expression (in vectorized form) for $\nabla_X a^T X b$

Answer: Recall that the gradient of a function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is defined as

$$\nabla_A f(A) \in \mathbb{R}^{m\times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

i.e., an $m \times n$ matrix with

$$(\nabla_A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}}.$$

To find $\nabla_X a^T X b$, we first find an expression for $\frac{\partial}{\partial X_{ij}} a^T X b$

$$\begin{aligned} \frac{\partial}{\partial X_{ij}} a^T X b &= \frac{\partial}{\partial X_{ij}} \sum_{i=1}^{n}\sum_{j=1}^{d} a_i b_j X_{ij} \\ &= a_i b_j \end{aligned}$$

Thus $(\nabla_X a^T X b)_{ij} = a_i b_j$ so $\nabla_X a^T X b = ab^T$

To compute the maximum likelihood estimate of Σ, we will consider the log-likelihood function

$$\ell = \sum_{i=1}^{n} \log p(x^{(i)}) = \sum_{i=1}^{n} -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(x^{(i)}-\mu)^T\Sigma^{-1}(x^{(i)}-\mu)$$

In order to compute the maximum likelihood estimate, we will consider a change of variables $S = \Sigma^{-1}$. This function happens to be concave in $S = \Sigma^{-1}$, so by making this substitution we can maximize the log-likelihood of S by finding $\nabla_S \ell$ and setting it equal to 0. We can then recover the optimal Σ as $\Sigma = S^{-1}$ because our change of variables transformation $f(A) = A^{-1}$ is bijective and thus invertible.

Note: analyzing the convexity of this function with respect to Σ is NOT expected for this class and this step can be taken as a given. The goal is to practice taking gradients with respect to matrices and to see the MLE estimate of the covariance matrix of a Gaussian.

With the change of variables, we have that

$$\ell = \sum_{i=1}^{n} -\frac{k}{2}\log(2\pi) + \frac{1}{2}\log(|S|) - \frac{1}{2}(x^{(i)} - \mu)^T S(x^{(i)} - \mu)$$

This follows from the identity $|X^{-1}| = \frac{1}{|X|}$ for invertible X.

(b) Compute $\nabla_S \ell$ and set it equal to 0 to find a closed form solution for the maximum likelihood estimate of S. Then invert this estimate to find the maximum likelihood estimate of Σ.
Hint: The following identities (and the identity from (a)) will prove useful:

$$\nabla_X |X| = |X|(X^{-1})^T$$
$$(X^{-1})^T = (X^T)^{-1}$$

Answer:

$$\nabla_S \ell = 0$$
$$\nabla_S(\sum_{i=1}^{n} -\frac{k}{2}\log(2\pi) + \frac{1}{2}\log(|S|) - \frac{1}{2}(x^{(i)} - \mu)^T S(x^{(i)} - \mu)) = 0$$
$$\frac{1}{2}\sum_{i=1}^{n} \frac{1}{|S|}|S|(S^{-1})^T - (x^{(i)} - \mu)(x^{(i)} - \mu)^T = 0$$
$$\frac{1}{2}\sum_{i=1}^{n}(S^{-1} - (x^{(i)} - \mu)(x^{(i)} - \mu)^T) = 0$$

Simplifying this expression yields

$$S = \left(\frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \mu)(x^{(i)} - \mu)^T\right)^{-1}$$

and thus, since $S = \Sigma^{-1}$,

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

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:

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