S&DS 365 Homework 2 Solutions

Yale University, Department of Statistics

Feb 23, 2021

1 Problem 1:

The function we wish to minimize is

$$f(\mu) = \sum_{i=1}^{n} |\mu - x_i|$$

taking the derivative of this function,

$$\frac{d}{d\mu} f(\mu) = \sum_{i=1}^{n} \text{sign}(\mu - x_i)$$

and we will find the minimizer when we have this derivative equal to zero.

Re order the points in increasing order $x_1 \leq x_2 \leq \cdots x_n$ and assume no two points are the same. If the number of points is even then pick $\hat{\mu}$ to be in the middle of the two middle points,

$$\hat{\mu} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

then there are $\frac{n}{2}$ points above and bellow $\hat{\mu}$ and thus the sum of the signs is 0. If n is odd, then $\hat{\mu}$ is the $\frac{n+1}{2}$ data point. This results in the same conclusion as above in that the number of data points above and bellow $\hat{\mu}$ are equal and thus the signs cancel. This proves the median as the minimizer under absolute error loss.

2 Problem 2:

We will write down the log likelihood and take it's derivative to find the MLE. Since the data points are all independent, the joint density of the data is the product of the individual densities

$$\begin{aligned} f(x_1, \cdots, x_n; \mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}} \end{aligned}$$

and the log likelihood is logarithm of this function

$$l(\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}$$

we then take the derivative in μ and set it equal to zero to solve for the MLE, note the first term has no dependence on μ

$$\frac{d}{d\mu} l(\mu) = \frac{\sum_{i=1}^{n}(x_i - \mu)}{\sigma^2} = 0$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

3 Problem 3:

We will provide two solutions below. One requires that X is full rank, so that $(X^T X)^{-1}$ exists. The other solution does not. Both solutions receive full credit.

Approach 1: We do not assume that $X^T X$ is invertible.

Recall that we aim to optimize $L(\theta) = \|X\theta - y\|_2^2$ over θ for ordinary least squares. Let $\widehat{\theta} \in \arg\min_\theta \|X\theta - y\|_2^2$. We know from the optimality conditions that for $\widehat{\theta}$ then it must satisfy

$$\begin{aligned}\nabla L(\theta) \mid_{\theta=\widehat{\theta}} &= \nabla L(\widehat{\theta}) \\ &= 2X^T(X\widehat{\theta} - y) \\ &= 0\end{aligned}$$

Therefore, we know that $X^T(X\widehat{\theta} - y) = 0$. Let $e = X\widehat{\theta} - y$. Thus, $X^T e = 0$. Furthermore, let $v = Xg$ for an arbitrary g. That is, v is in the column space X or span(X).

$$\begin{aligned}\langle v, e &= v^T e \\ &= g^T X^T e \\ &= 0 \quad \text{since } X^T e = 0.\end{aligned}$$

Approach 2: We assume $X^T X$ is invertible.

The least squares estimator is $\hat{\theta} = (X^T X)^{-1} X^T y$.

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

Thus, $X^T e = 0$. Let $v = Xg$ be any vector in the column space of X. Then,

$$\begin{aligned}v^T e &= g^T X^T e \\ &= g^T 0 \\ &= 0\end{aligned}$$

4 Problem 4:

Let $D = \text{diagonal}(d_1, \cdots, d_n)$ be the diagonal matrix with entries d_i along the diagonal and 0's elsewhere. Then we have

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} d_i(x_i^T\theta - y_i)^2 \tag{1}$$

$$= \|D^{\frac{1}{2}}(X\theta - y)\|_2^2 \tag{2}$$

we can work with either the sum expression or the matrix form. Working with the sum expression (1)

$$\begin{aligned}\nabla_\theta L(\theta) &= \frac{2}{n}\sum_{i=1}^{n} d_i(x_i^T\theta - y_i)x_i \\ &= \frac{2}{n}\sum_{i=1}^{n} d_i(x_i^T\theta x_i - y_i x_i) \\ &= \frac{2}{n}X^T DX\theta - \frac{2}{n}X^T Dy\end{aligned}$$

setting this equal to zero we must have

$$\theta = (X^T DX)^{-1}X^T Dy$$

note that if $d_i = 1$ for all i this is the usual least squares estimator.

We can also derive the gradient from the norm expression (2) and get the same result

$$\begin{aligned}\nabla_\theta L(\theta) &= \nabla_\theta \|D^{\frac{1}{2}}(X\theta - y)\|_2^2 \\ &= \frac{2}{n}X^T D^{\frac{1}{2}}(D^{\frac{1}{2}}X\theta - D^{\frac{1}{2}}y) \\ &= \frac{2}{n}(X^T DX\theta - X^T Dy)\end{aligned}$$

which is the same expression as above.