Yale University
Department of Statistics and Data Science

Statistics 365/565

Issued: 02/12/2021 **Due:** 02/23/2021

Notes: Different losses.

Notation: $[k] = \{1, 2, \dots, k\}$. For a matrix $A \in \mathbb{R}^{m \times n}$ we will let $A_{(i,:)}$ denote the i^{th} row and $A_{(:,j)}$ denote the j^{th} column. **Both will be treated as column vectors.**

Problem 1: Suppose that we observe data that we wish to model as $x_i = \mu^* + w_i \in \mathbb{R}$ where w_i is the error (or noise) in our observations (not to be confused with residual error from least squares. context is king.) Assume that none of the x_i are exactly equal. We wish to estimate μ^*. One reasonable approach is to minimize the loss to estimate μ^* and call the estimate $\widehat{\mu}$

$$\widehat{\mu} = \text{argmin}_v \sum_{i=1}^{n} \ell(v, x_i)$$

We have already seen that the choice of $\ell(v, x) = (v - x)^2$ results in the solution $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Suppose that we take $\ell(v, x) = |v - x|$. What is the solution to the above optimization? Note that there is not a closed form solution, but there is a common name for the solution.

Hint: Take the derivative of $g(v) = |v|$ to be

$$g'(v) = \text{sign}(v) \equiv \begin{cases} +1 & \text{if } v > 0 \\ -1 & \text{if } v < 0 \\ 0 & \text{otherwise} \end{cases}$$

In the coding part you will explore some implications of this choice versus the ℓ_2 choice.

Problem 2: We have so far derived our optimization problems from the perspective of minimizing a loss. Another interpretation is a probabilistic one known as the maximum likelihood estimate. (Please see the notes mle.pdf). We will motivate Example 0 from lecture in this way.

Recall that a normally distributed random variable $V \sim N(\mu, \sigma^2)$ is a Gaussian (normally) distributed random variable with mean μ and variance σ^2. Its probability density function is

$$f_V(v) = \frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{(v - \mu)^2}{2\sigma^2}\right)$$

Suppose that we observe data $x_i \sim N(\mu^*, \sigma^2)$. One way to model this is that $x_i = \mu^* + w_i$ where w_i is $N(0, \sigma^2)$. We will assume that all of x_i are i.i.d. (identically and independently distributed). From the mle notes compute the log-likelihood and maximize over the choice of μ to obtain your estimate $\widehat{\mu}$. Your estimate should be $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Problem 3: Let $X \in \mathbb{R}^{n\times d}$ and $y \in \mathbb{R}^n$ be the feature matrix and observation vector, respectively. Recall the ordinary least squares problem is to solve

$$\widehat{\theta} = \text{argmin}_\theta \, \|X\theta - y\|_2^2$$

Prove that the vector of residual errors $e = X\widehat{\theta} - y$ is orthogonal to any column of X. As a reminder, two vectors v and w are orthogonal if $\langle v, w\rangle = 0$. Use this fact to establish that for any vector $v \in \text{span}(X)$ (which means any vector v in the span of the columns of X, equivalently in the column space of X, or also equivalently any $v = Xg$ for some arbitrary $g \in \mathbb{R}^d$). Then, $v^T e = 0$. That is to say that e is orthogonal to the column space of X.

Problem 4: Sometimes, it is useful to weight different observations in different ways. We define our risk as

$$L_d(\theta) = \frac{1}{n}\sum_{i=1}^{n} d_i \ell(\widehat{f}(x_i), y_i)$$

As an example we can consider weighted least squares

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

If $d_i = 1$ for all i then we know the solution to the above is simply the solution to ordinary least squares. However, changing d_i alters the solution. Find a closed form solution for $\widehat{\theta}$ defined as

$$\widehat{\theta} = \text{argmin}_\theta \, L_d(\theta)$$

Your solution should depend on d.