Discussion 6 – Loss, Transformations, and Correlation

Suraj Rampure

Friday, February 28th, 2020

Agenda

- Loss Functions
- Transformations
- Correlation

Some demos. As per usual, everything will be posted at

http://surajrampure.com/teaching/ds100.html

Loss Functions

- When creating a model for our data, we need some metric of how **good** our model is. This is what a loss function is.
- Here, as you did in lecture, we will consider the "constant model", meaning $\hat{y}_i = \theta$. This means that the **parameter** of our model is the constant θ – this is what we will try to find.

Recall the following from lecture: $(\text{actual} - \text{predicted})^2$

- L_2 loss for a single point: $(x_i - \theta)^2$
- Average L_2 loss for entire datase (i.e. our objective function): $\frac{1}{n}\sum_{i=1}^{n}(x_i - \theta)^2$
- Optimal value of θ: $\hat{\theta} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

There were two ways to arrive at this result – by calculus, and by an arithmetic technique (adding and subtracting $\bar{x}$ within the loss).

Let's now find the optimizing $\hat{\theta}$ when using L_1 loss.

$$L = \frac{1}{n} \sum_{i=1}^{n} |x_i - \theta|$$

$$|x_i - \theta| = \begin{cases} x_i - \theta & x_i > \theta \\ \theta - x_i & x_i \le \theta \end{cases}$$

$$\frac{dL}{d\theta} = \frac{1}{n} \left(1 + (-1) + (-1) + 1 + \cdots + (-1) \right)$$

$$\frac{d\,|x_i - \theta|}{d\theta} = \begin{cases} -1 & x_i > \theta \\ 1 & x_i \le \theta \end{cases}$$

$$= \frac{1}{n} \left[\sum_{x_i \le \theta} 1 + \sum_{x_i > \theta} (-1) \right]$$

$$= \frac{1}{n} \left[m_\theta \cdot 1 + (n - m_\theta)(-1) \right] = 0$$

$$m_\theta + m_\theta - n = 0$$

$$\Rightarrow \boxed{m_\theta = \frac{n}{2}} \Longrightarrow \boxed{\hat{\theta} = \text{median}(x)}$$

Transformations

Let's refer to the worksheet for this problem.

Correlation

The concept of correlation is intimately tied to the idea of simple linear regression.

$$r(x, y) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

r, denoted the **correlation coefficient**, is a value between -1 and 1.

- A value of 0 denotes absolutely no linear correlation.
- As r approaches 1 (or -1), the strength of the correlation between x and y increases.
- The sign of r tells us whether our correlation is positive (up and to the right) or negative (down and to the right).

"mean of the product of x and y, in standard units"

Correlation and Simple Linear Regression

$$r(x,y) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

Then, for the simple linear regression model $\hat{y}_i = a + bx_i$, we have

$$\hat{b} = r\frac{\sigma_y}{\sigma_x}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

When using L_2 loss!