DS 100/200: Principles and Techniques of Data Science **Date: Feb 28, 2020**

Discussion #6

Name:

Loss Functions

1. The l_2 (squared) loss is the most commonly used loss function, in part because it has many nice properties, e.g.,

   - We can find the minimizer analytically, i.e., we can add and subtract the mean or we can differentiate.
   - In homework 1 exercise 1, you showed that the sample mean minimizes the average squared loss for the constant estimator.
   - The minimum average squared loss for the constant estimator corresponds to the sample variance, i.e.,

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (x_i - \theta)^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Data scientists sometimes use other loss functions when minimizing loss. Another popular loss function is the l_1 (absolute) loss.

Suppose that we have data $x_1, \ldots, x_n$. $y = \theta$

LOSS: We use a loss function to determine the loss resulting from a particular choice of model.

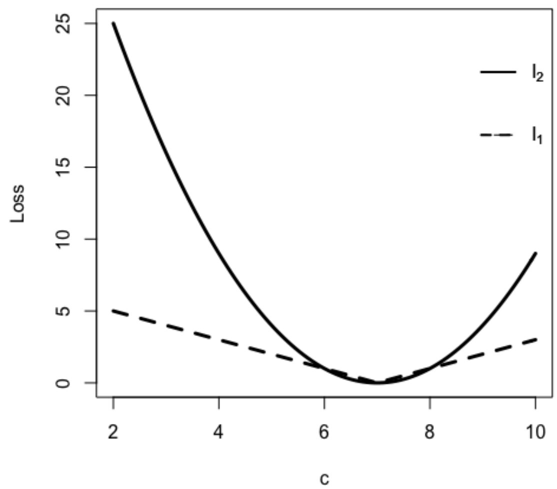
If θ is our predicted value and y is the actual value, then l_1 loss is defined as

$$l_1(\theta, y) = |y - \theta|$$

AVERAGE LOSS: We would like to find the value θ that minimizes the loss over all of our data. Specifically, we wish to minimize the average absolute loss:

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

We will heuristically derive the minimizer of the average l_1 loss for the constant estimator. But, before we do, examine the plot of the l_1 and l_2 loss functions below. These are expressed as functions of c, for $x = 7$. That is, we have plotted $|7 - \theta|$ and $(7 - \theta)^2$. Think about why might we prefer to use one loss function over another.



In our heuristic derivation, we will make two simplifying assumption: (a) all of the data values are unique and (b) there are an even number of data values. Follow the steps below to minimize the average absolute loss.

2. STEP 1: Split the summation into two summations, one for the $y_i \leq \theta$ and the other for the $y_i > \theta$

   $\min_\theta \frac{1}{n} \sum_{i=1}^{n} |y_i - \theta| =$

3. STEP 2: Rewrite $|y_i - \theta|$ in each summand so that it doesn't use absolute value.

4. STEP 3: Differentiate with respect to θ. (Don't worry about the dependence of the summation on θ - this is just a heuristic proof.)

5. STEP 4: Let m_θ represent the number of y_i that are less than or equal to θ. Set the derivative above to 0 and rewrite the two summands in terms of m_θ and n.

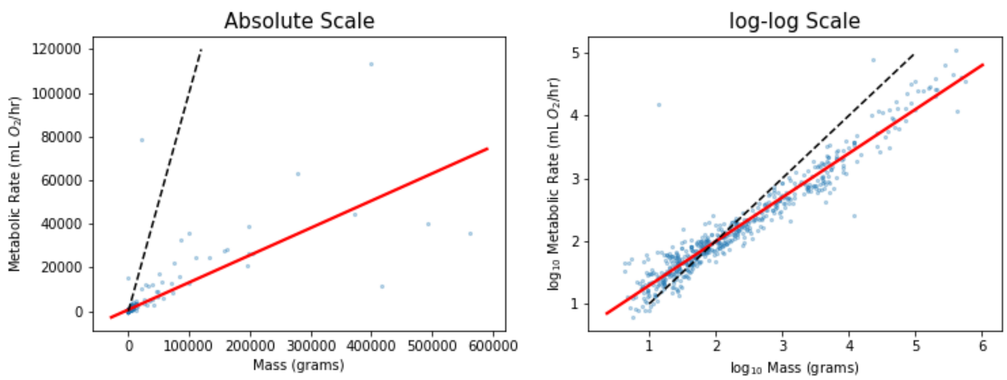
6. STEP 5: Explain why the minimizing value is the sample median.

SEE LAST PAGE FOR THE DERIVATION

Logarithmic Transformations

Bonus Question: can we use $-\log(x)$ transformations?

7. One of your friends at a biology lab asks you to help them analyze `panTHERIA`, a database of mammals. They are interested in the relationship between mass, measured in grams, and metabolic rate ("energy expenditure"), measured by oxygen use per hour. Originally, they show you the data on a linear (absolute) scale, shown on the left. You notice that the values on both axes vary over a large range with many data points clustered around the smaller values, so you suggest that they instead plot the data on a log-log scale, shown on the right. The solid red line is a "line of best fit" (we'll formalize this later in the course) while the black dashed line represents the identity line $y = x$.



(a) Let C and k be some constants and x and y represent mass and metabolic rate, respectively. Based on the plots, which of the following best describe the pattern seen in the data?

○ A. $y = C + kx$ ○ B. $y = C \times 10^{kx}$ ○ C. $y = C + k\log_{10}(x)$ ● D. $y = Cx^k$

log-log plot shows $\log(y) = k\log(x) + C \Rightarrow y = e^{k\log(x)+C} = e^{k\log(x)} e^C = (e^{\log(x)})^k e^C = x^k \cdot e^C = Cx^k$

(b) What parts of the plots could you use to make initial guesses on C and k?

C: y-intercept of red line in log-log plot

k: slope of red line in log-log plot

(c) Your friend points to the solid line on the log-log plot and says "since this line is going up and to the right, we can say that, in general, the bigger a mammal is, the greater its metabolic rate". Is this a reasonable interpretation of the plot?

yes; positive slope $\Rightarrow$ increase in x leads to increase in y

(d) They go on to say "since the slope of this line is less than 1, we see that, in general, mammals with greater mass tend to spend less energy per gram than their smaller counterparts". Is this a reasonable interpretation of the plot?

yes; $\frac{dy}{dx}$ shows the rate of change of y as x changes. In this case,

$\frac{dy}{dx} = \frac{d}{dx}[Cx^k] = Ckx^{k-1}$. Since $0<k<1$, x^{k-1} decreases as x increases.

Thus, $\frac{dy}{dx}$ decreases as k increases

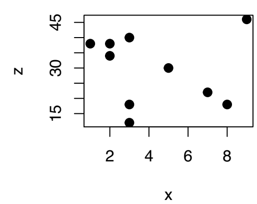
8. When making visualizations, what are some reasons for performing log transformations on the data?

- makes plots fit on a reasonable size axis
- makes exponential relationships linear (which is nice bc we can then use tools like linear algebra to analyze those relationships)

Regression Notions

9. When we have more than two variables, it can be difficult to discern relationships from pairwise plots. Here is an example. Consider the 3 variables x, y, and z. We have 10 observations. Suppose we are interested in predicting z.

x	y	z
2	17	38
1	18	38
9	14	46
7	4	22
8	1	18
2	15	34
3	17	40
3	3	12
5	10	30
3	6	18

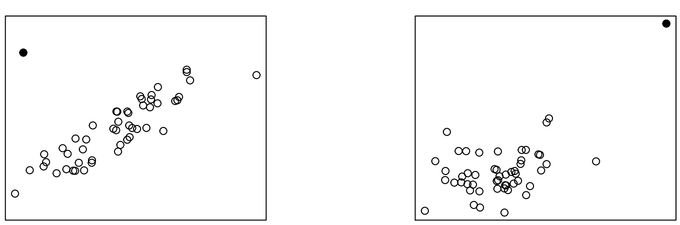


x alone is bad for predicting z values. But x and y together allow for perfect prediction of z values.

Takeaway: Just because the correlation between a target variable and a data variable is low does not mean that data variable is not useful for predicting the target variable.

The correlation between x and z is -0.07. The scatter plot reflects this weak relationship. It appears that we should not bother to include x in a linear model for predicting z. Examine x, y and z carefully, and in the space above, sketch a scatter plot to show that there is a useful linear relationship that involves x. $z = 2x + 2y$

10. Consider the two scatter plots below. For each scatter plot consider what happens to the correlation when the specially marked point is removed. Does the correlation get weaker, stronger, or stay about the same?



The approach for this problem is to draw a line of best fit with the black point included and without the black point included. If the absolute value of the slope of the second line is closer to 1 than the absolute value of the slope of the first line, then correlation increased.

1

Data points: $x_1 \dots x_n$

Target variable values: $y_1 \dots y_n$

Model: $y = \theta$ [θ is a constant]

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

Question: What is the best θ?

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

step 1: split the sum

$$L = \frac{1}{n}\left(\sum_{y_i \le \theta} |y_i - \theta| + \sum_{y_i > \theta} |y_i - \theta|\right)$$

when $y_i \le \theta$, $|y_i - \theta| = \theta - y_i$

when $y_i > \theta$, $|y_i - \theta| = y_i - \theta$

$$L = \frac{1}{n}\left(\sum_{y_i \le \theta} \theta - y_i + \sum_{y_i > \theta} y_i - \theta\right)$$

take derivative, set to 0, solve for minimizing θ

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

derivative of sum is the sum of derivatives

[e.g. $\frac{d}{dx}(x^2 + x) = \frac{d}{dx}(x^2) + \frac{d}{dx}(x)$]

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

$$\frac{dL}{d\theta} = \frac{1}{n}\left(\sum_{y_i \le \theta} 1 + \sum_{y_i > \theta} -1\right) = 0$$

pull negative out of second sum

$$\left(\sum_{y_i \le \theta} 1\right) - \left(\sum_{y_i > \theta} 1\right) = 0$$

$$\left(\sum_{y_i \le \theta} 1\right) = \left(\sum_{y_i > \theta} 1\right)$$

$\sum_{y_i \le \theta} 1 = \underbrace{1 + 1 + \dots + 1}_{\text{however many } y_i \le \theta}$

$$(\# \, y_i \leq \theta) = (\# \, y_i > \theta)$$

$$\theta = \text{median}(y)$$