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

Discussion #6 Solutions

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 $x_1, \ldots, x_n$.

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

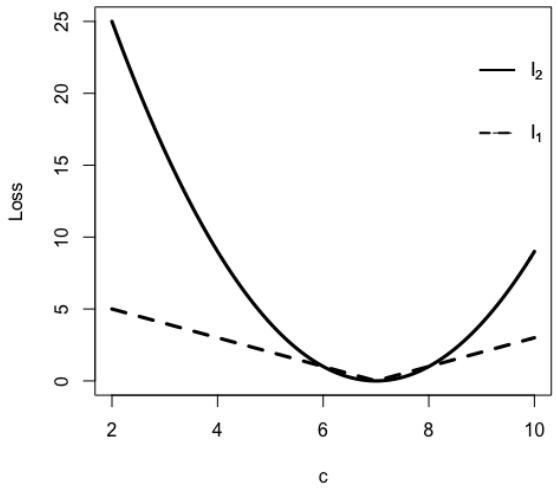
   If θ is our predicted value and y is the actual value, then l_1 loss is defined as

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

   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.



Solution: Less sensitive to large errors.

Comparing l_2 and l_1 loss in the figure, we can see that l_2 would be useful in situations where a large error is catastrophically worse than a small error. Medical diagnoses come to mind.

On the other hand, situations where larger errors are just linearly worse that small ones might be in investing. We would use l_1 error in these situations.

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 $x_i \leq \theta$ and the other for the $x_i > \theta$

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

3. STEP 2: Rewrite $|x_i - \theta|$ 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 x_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.

Solution:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}|x_i-\theta| &= \frac{1}{n}\sum_{x_i\le\theta}|x_i-\theta|+\frac{1}{n}\sum_{x_i>\theta}^{n}|x_i-\theta| \\ &= \frac{1}{n}\sum_{x_i\le\theta}(\theta-x_i)+\frac{1}{n}\sum_{x_i>\theta}(x_i-\theta)\end{aligned}$$

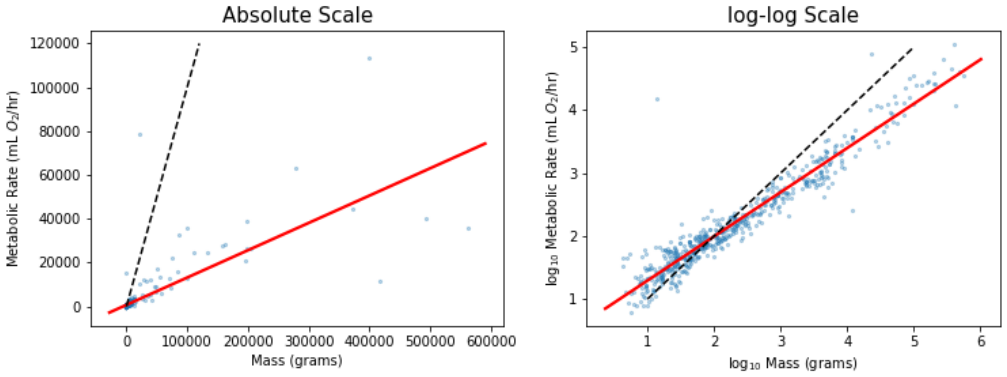
Differentiate with respect to θ and set to 0:

$$\begin{aligned}0 &= \frac{1}{n}\sum_{x_i\le\theta}1+\frac{1}{n}\sum_{x_i>\theta}^{n}-1 \\ &= \frac{m_\theta}{n}-\frac{n-m_\theta}{n}\end{aligned}$$

This is minimized when the number of x_i below θ equals the number above, i.e., for the median.

Logarithmic 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. $y = Cx^k$

Solution: Starting with $y = Cx^k$, we can take the $\log_{10}$ of both sides to find the relationship between $\log_{10}(y)$ and $\log_{10}(x)$.

$$\begin{aligned}\log_{10}(y) &= \log_{10}(Cx^k) \\ &= \log_{10}(C) + \log_{10}(x^k) \\ &= \log_{10}(C) + k\log_{10}(x)\end{aligned}$$

Thus, $\log_{10}(y)$ and $\log_{10}(x)$ are linearly related, which matches what the log-log plot shows above.

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

Solution:

- C: 10^b, where b is the y-intercept of the solid red line in the log-log plot.
- k: slope of the solid red line 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?

Solution: Yes, the observation is equivalent to saying that the slope is positive, which means increases in x correspond to increases 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?

Solution: Yes, a slope between 0 and 1 means that k is likely between 0 and 1. Looking at $\frac{dy}{dx}$, we see that for these values of k, as x grows, its effect on y diminishes. In this case, it means that gram-for-gram larger mammals spend less energy than their smaller counterparts.

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

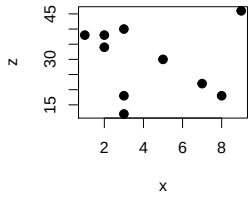
Solution: Comparing orders of magnitude, when the underlying effects seems to be multiplicative and not additive. One heuristic is that "trimming outliers" doesn't seem to be helping the scale of the plot, i.e., new "outliers" appear when you truncate the data.

You have some domain knowledge about the variable, e.g., intensities measured on 16-bit scale.

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



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.

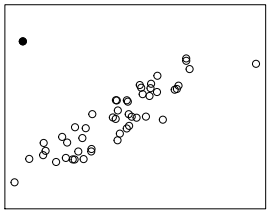
Solution: Although there is no relationship between x and z, if we know y then we can use x to perfectly predict z. That is,

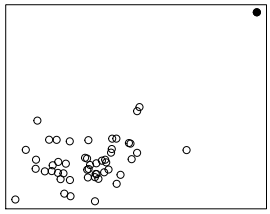
$$z = 2x + 2y$$

A scatter plot of $(2x + 2y, z)$ shows that all points fall on a line.

Scatter plots are limited in that they can't always reveal a linear relationship between three or more variables.

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?





Solution: When we drop the point in the left scatter plot, the correlation will get stronger. The correlation increases from 0.60 to 0.90, when the point is dropped.

When we drop the point in the right scatter plot, the correlation will get weaker. For these data, the correlation drops from 0.60 to 0.30.