DS 100/200: Principles and Techniques of Data Science Date: Sep 11, 2019

Discussion #3 Solutions

Name:

Error, Loss, and Risk

The l_2 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 as shown in lecture or we can differentiate as shown in discussion last week.
- The minimum empirical risk corresponds to the sample variance, i.e.,

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

- For a SRS from a population, the expected value of the sample mean (which minimizes the empirical risk) equals the population average (which minimizes the risk). This property also holds for many probability models.

Data scientists sometimes use other loss functions when minimizing risk. Another popular loss function is the l_1 loss. We will derive the minimizer of the average l_1 loss as a way to review the concepts of error, loss, and risk.

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

ERROR: If we summarize the data with the value c then we incur errors. The error for x_1 is $x_1 - c$, for x_2 it is $x_2 - c$ and so on.

LOSS: We want to translate these errors into a loss. The loss is the "cost" of making an error. We use a loss function to determine this cost. The cost is nonnegative and typically grows with the error.

l_1 loss, also known as *absolute loss* is defined as

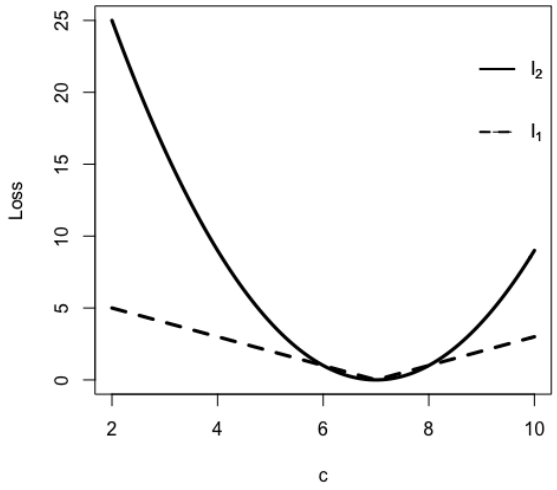
$$l(x - c) = |x - c|$$

EMPIRICAL RISK: We would like to find the value c that minimizes the loss over all of our data. Specifically, we wish to minimize the average loss, i.e empirical risk:

$$\min_c \frac{1}{n}\sum_{i=1}^{n}|x_i - c|$$

We will heuristically derive the minimizing value for average absolute loss. 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 $|7 - c|$ and $(7 - c)^2$.

1. 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 $x_i \leq c$ and the other for the $x_i > c$

   $\min_c \frac{1}{n} \sum_{i=1}^{n} |x_i - c| =$

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

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

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

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

Solution:

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

Differentiate with respect to c and set to 0:

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

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

The World, Data Design, and the Sample

In 2000, Hayes investigated whether there were any differences in prices for grocery stores in different areas of New York City. The goal was to determine whether prices were higher in the poorer areas of the city.

The sampling frame they used consisted of 1408 food stores with at least 4000 square feet of retail space. The price of a "market basket" of goods was determined for each store.

As in this study, we are often concerned about differences between groups in our population. For this reason, rather than take a SRS of the 1408 stores in the city, Hayes divided the stores up into 3 groups according to the median household income in the store's zip code. Then a SRS of stores was taken in each group. This approach is called a *Stratified Random Sample*. That is, the population is divided into non-overlapping groups and a SRS is taken from each group, independently.

See Hayes (2000) *Are prices higher for the poor in New York City?* in the Journal of Consumer Policy for more information. One finding from the study was that the overall food basket is cheaper in poor areas but cereal, orange juice, apples and bananas are significantly costlier.

7. Can you think of a reason why the researcher carried out a stratified random sample rather than a simple random sample?

Solution: Taking a SRS with in each group of stores gives us assurance of the accuracy of the sample average for each group because we can divide the sample size up between groups to make sure that we get a reasonable sampling error for each group.

When a group is small in comparison to other groups, if we stratify then we are certain to get enough samples from the small group. We can take more samples, proportionally from the small group. This is called oversampling. It is a common practice.

More advanced idea: When the group means are quite different, we can get a more accurate estimate of the population mean by stratifying.

8. Why do you think the researcher defined the sampling frame to be food stores with at least 4000 square feet?

Solution: Small food stores are notorious for their high prices and they might not have all of the items in the market basket available for sale. Restricting the sampling frame in this way better lets us compare like with like.

However, one problem is that we are assuming that different income groups have similar access to these super markets. If they do not, this sampling frame could skew our analysis. In Hayes' sampling frame, 26% of the population is categorized as low income, however only 13% of the super markets are categorized as low income.

The World, Data Design, and the Sample

In lecture, we examined the data generation process for a Simple Random Sample (SRS). Here we draw connections between three notions: the population, random variables, and the sample/empirical data.

We use a simplification of the previous study of the price of a market basket of goods. We examine a hypothetical city with 15 grocery stores and take a SRS of 3 stores. The value for each story is the price of the market basket.

Below is a diagram to help draw the distinction between the "world" that we want to generalize to, the data generation process used to obtain our data (aka data design), and the data that we got.

9. Fill in the missing information

POPULATION	DATA DESIGN	EMPIRICAL
3, 3, 5, 4, 4, 3.5, 3, 4, 3.5, 3.5, 5, 3.5, 4, 4, 3	SRS of 3 stores	Data: $x_1 = 4, x_2 = 3, x_3 = 5$
Histogram of Population	X_1 = price for first store sampled	Histogram of Sample
	x | | | | | $P(x)$ | | | | |	
$N =$	$\mathbb{E}(X_i) =$	$n =$
		$\bar{x} =$
Pop Mean =	$\mathbb{V}ar(X_i) =$	sample variance =
	$\mathbb{E}(\bar{X}) =$	$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 =$
Pop Var =	$\mathbb{V}ar(\bar{X}) = \frac{N-n}{N-1}\frac{\mathbb{V}ar(X_1)}{n} =$	

Solution: For the population, $N = 15$, Pop Mean = 3.73, Pop Var = 0.396. The histogram should have bars the same width above 3, 3.5, 4, and 5 (1/2 wide) with heights of 4, 4, 5, and 2.

For the data design, the probability distribution table for a draw from the box, X, is:

x	3	3.5	4	5
$P(x)$	4/15	4/15	5/15	2/15

$\mathbb{E}(X_i) = 3.73$

$\mathbb{V}ar(X_i) = 0.396$

$\mathbb{E}(\bar{X}) = 3.73$

$\mathbb{V}ar(\bar{X}) = \frac{N-n}{N-1}\frac{\mathbb{V}ar(X_1)}{n} = 0.113$

For the empirical (the sample), $n = 3$, $\bar{x} = 4$, sample variance $= 2/3$. The sample histogram will have three bars of the same height, one each at 3, 4, 5.

10. Compare the population, data design, and sample. Make 4 observations about the similarities and differences. For example, the probability distribution of X_1 matches the population distribution.

Solution: With the probability model, we know that we can expect the samples to look like the population. More specifically:

1. The probability distribution matches the population distribution.
2. The expected value of X equals the population average.
3. The variance of X equals the population variance.
4. The expected value of the random average, $\bar{X}$ equals the population average. This means that it is an unbiased estimate of the population average.
5. The variance of $\bar{X}$ is smaller than the variance of X because we are averaging 3 random variables so we can expect the variability to decrease. Also, since we are sampling without replacement, the variability will decrease.

In addition, the sample looks somewhat like the population, but clearly not exactly like the population.

6. Our particular sample has an average of 4, which is close to 3.73, but definitely not spot on.
7. The empirical histogram is similar in shape to the population histogram, but definitely not the same.
8. The variance of the sample, is the mean square error, the average squared loss, the empirical risk. It is an estimate itself of the population variance.