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

Discussion #3

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:

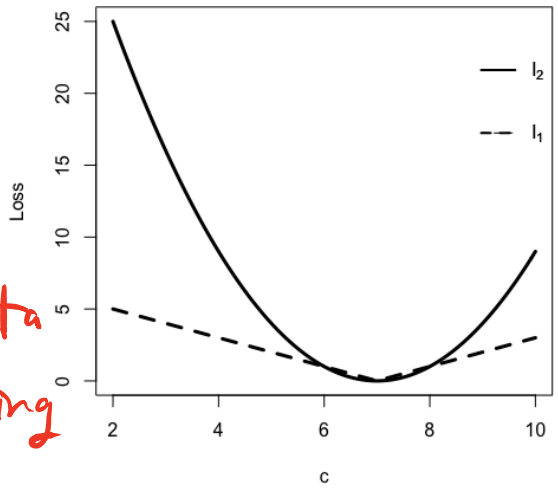
$$\min_c \frac{1}{n}\sum_{i=1}^{n}|x_i - c|$$

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?

L_1 : - doesn't penalize errors that much (grows linearly)

e.g. : - financial data
- disaster warning



L_2 : - smooth and differentiable everywhere
- penalizes wrong predictions significantly

e.g. : medical data

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 $x_i > c$

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

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.

$$L(x_i, c) = |x_i - c|$$

$$= \begin{cases} x_i - c & x_i \geq c \\ c - x_i & x_i < c \end{cases}$$

$$\frac{dL(x_i, c)}{dc} = \begin{cases} -1 & x_i \geq c \\ 1 & x_i < c \end{cases}$$

$$f(c) = \frac{1}{n} \sum_{i=1}^{n} L(x_i, c)$$

Empirical risk

$$f'(c) = \frac{1}{n} \sum_{i=1}^{n} \frac{dL(x_i, c)}{dc}$$

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

$$= \frac{1}{n} \left(1 \cdot (\#\, x_i < c) + (-1) \cdot (\#\, x_i \geq c) \right)$$

$$= 0$$

$$(\#\, x_i < c) - (\#\, x_i \geq c) = 0$$

$$\boxed{(\#\, x_i < c) = (\#\, x_i \geq c)}$$

$$\Rightarrow \quad \hat{c} = \text{median}$$

see solutions for answer
see my disc 3 slides for stratified sampling def'n

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?

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

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: 3, 3.5, 4, 5; $P(x)$: 4/15, 4/15, 5/15, 2/15	
$N =$ 15	$\mathbb{E}(X_i) =$ 3.73	$n =$ 3
Pop Mean = 3.73	$\mathbb{V}ar(X_i) =$ 0.396	$\bar{x} =$ 4
Pop Var = 0.396	$\mathbb{E}(\bar{X}) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n} \cdot n \cdot \mathbb{E}[X_1] = \mathbb{E}[X_1] = 3.73$	sample variance = $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{3}\left((3-4)^2 + (4-4)^2 + (5-4)^2\right) = \frac{2}{3}$
	$\mathbb{V}ar(\bar{X}) = \frac{N-n}{N-1}\frac{\mathbb{V}ar(X_1)}{n} =$	

note: each X_i has same distribution, but are not independent

correction factor

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.

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

pop and data design values overlap: not a coincidence: each store has equal prob.

$$\mathrm{Var}(X) = E\left[(X - \mu)^2\right]$$

$$= \sum_x (x-\mu)^2 \cdot P(X=x)$$

$$E[X] = \sum_x x \cdot P(X=x)$$

$$E[g(X)] = \sum_x g(x) \cdot P(X=x)$$