DS 100/200: Principles and Techniques of Data Science Date: Sep 4, 2019

Discussion #2 Solutions

Name:

Working with 0-1 Data

Often the data we work with are indicator variables that *indicate* whether a quality exists or not in individuals. Examples include: whether a voter voted for Trump in the 2016 election; whether someone released from Broward prison committed a crime within two years of release; whether 8^{th} boys in a school district out-perform the girls on a math test; whether women who have been married more than 5 years are having affairs.

These variables can be represented by 0-1 values, where a 1 denotes the individual has the characteristic and a 0 that they don't. This is a common way to represent a qualitative variable. Aside from being a common occurrence, 0-1 data are special in that they have features that make them easy to work with.

Let's explore these aspects of 0-1 data. We begin with a simple example. There are 20 people in Deb's extended family (parents, siblings, spouses, nieces and nephews) who were of voting age in 2016. From oldest to youngest, here's how they voted, where 1 stands for a Trump vote.

$1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0$

1. If we want to summarize the support in her family for Trump, how would we do this?

Solution: Any of the following answers is OK:

- The number of 1s
- the fraction of 1s
- the percentage of 1s

2. More generally, we can refer to these data as $x_1, x_2, \ldots, x_n$, where each x_i is a 0 or 1. Show that the average of the x_i, i.e., $\bar{x}$, is the proportion of 1s in the data. (In your proof, assume that m of the n values are 1s).

Solution:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} x_i &= \frac{1}{n}\sum_{1s} x_i + \frac{1}{n}\sum_{0s} x_i \\ &= \frac{m}{n}\end{aligned}$$

3. In class, we saw that for general values of x_i, the sample average minimizes the empirical risk for l_2 loss:

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

Confirm that the minimizer of the empirical risk in the special case of 0-1 data is the proportion of 1s in the sample. To do this, rewrite the summation as two summations, one for the xs that are 0 and the other for 1s. Let m be the number of 1s.

Solution:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 = \frac{1}{n}\sum_{1s}(1-c)^2 + \frac{1}{n}\sum_{0s}(0-c)^2 = \frac{1}{n}\sum_{1s}(1-c)^2 + \frac{1}{n}\sum_{0s} c^2$$

$$= \frac{m}{n}(1-c)^2 + \frac{n-m}{n}c^2$$

By setting this equal to zero and taking the derivative, we can find the value of c that minimizes the empirical risk:

$$\frac{d}{dc}\quad \frac{m}{n}(1-c)^2 + \frac{n-m}{n}c^2 = \frac{-2m}{n}(1-c) + \frac{2(n-m)}{n}c$$

$$\frac{-2m}{n}(1-\hat{c}) + \frac{2(n-m)}{n}\hat{c} = 0 \implies$$

$$-m(1-\hat{c}) + (n-m)\hat{c} = 0 \implies -m + n\hat{c} = 0 \implies \hat{c} = \frac{m}{n}$$

Note that for the l_2 loss, if you set the derivative of the loss to 0 and find a local minimum, then the local minimum is also a global minimum.

Additional Note: You can also do this by taking the derivative of the original expression with respect to c and solve for the optimal value of c.

Working with 0-1 Random Variables

Often we want to generalize our findings beyond the set of values that we have observed. For example, we might want to generalize to all voters, all parolees, all married women, all school districts, etc. To do this we need to understand how the data we have observed were generated. As you saw in Data 8, it can be problematic to generalize from your data to a larger population.

We describe on approach here that depends on a random process where we can compute the probability of an individual winding up in our sample.

Suppose that an individual chosen at random from the population has a chance p of having the characteristic, i.e., a chance p of being 1. And chance $1 - p$ of being 0.

Let X_1 – capital X, denote the 0-1 value of the first individual chosen at random according to this random process. Similarly define $X_2, \ldots, X_n$. We use upper case letters here to denote that these are random quantities that represent the possible outcome that the chance process might yield. This is not the same as our observed data values.

4. Provide a probability distribution table for X_1

Solution:

x	0	1
$P(x)$	$1-p$	p

5. Provide a probability distribution table for X_n

Solution: X_n should have the same distribution table as X_1.

6. What is the expected value of X_1?

Solution:

$$\mathbb{E}(X_1) = \sum_{x \in \{0,1\}} xP(x) = 0(1-p) + 1(p) = p$$

7. If our goal is to find an estimator for the probability distribution that minimizes the expected squared loss, i.e.,

$$\mathbb{E}[(X - c)^2]$$

show that the value of c that minimizes the expected squared loss is $\mathbb{E}(X)$.

Solution: One way to show this is to write $X - c = X - \mu + \mu - c$ where $\mu = \mathbb{E}(X)$. Squaring both sides,

$$\mathbb{E}[(X-c)^2] = \mathbb{E}[(X-\mu)^2 + (\mu - c)^2 + 2(X-\mu)(\mu - c)]$$

Now using linearity of expectation and pulling out the constants,

$$\begin{aligned}\mathbb{E}[(X-c)^2] &= \mathbb{E}[(X-\mu)^2] + (\mu-c)^2 + 2\underbrace{\mathbb{E}[X-\mu]}_{=0}(\mu - c) \\ &= \mathrm{Var}(X) + (\mu - c)^2.\end{aligned}$$

Since $c = \mu$ minimizes the second term of the above expression, it also minimizes expected squared loss as the first term is constant with respect to c.

Note: You can also take the derivative with respect to c in the original expression and solve for the optimal value.

8. Note that with this data generation process, $\frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ can be used to estimate $p = \mathbb{E}(X_1)$. And we can show that

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

Hence, the data generation/design process is crucial to our ability to make a good estimate for p.

Solution:

$$\mathbb{E}(\bar{X}) = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n} p = \frac{1}{n}np = p$$

Population, Sampling Frame, and Bias

Hite Report

Shere Hite published the Hite Report in 1987. The book reported findings from a survey of 4,500 women. Some of these findings were quite sensational. We'll focus on one in particular:

> 70% who are married 5+ years are having sex outside their marriage

To carry out the survey, 100,000 questionnaires were mailed to such organizations as professional womens groups, counseling centers, church societies. Identify the following aspects of the sampling process.

9. Target Population – complete collection of individuals we want to generalize to

> **Solution:** Women married at least 5 years in the US

10. Question. – the focused question that we are trying to answer about the population.

> **Solution:** The proportion of women married who are "faithful".

11. Sampling Frame – collection of individuals that might have been chosen for the sample

> **Solution:** Women who belong to one of the Women's group that received the questionnaires. (We don't know if the respondents are actually married for at least five years.)

12. Design – technique used to survey individuals in the sampling.

> **Solution:** This was not a probability sample. The respondents were those in the frame who took the time to complete the survey. It's not clear from the information given how the women's groups were selected.

13. Sources of Bias – potential sources of bias that might be introduced by the difference between the population and the sampling frame and the sampling method/design. It looks like the sample is representative of the population in terms of race and location (see tables below), but they might differ in other important ways.

Location	Study	U. S.
Large city/urban	60%	62%
Rural	27%	26%
Small town	13%	12%

Race	Study	U. S.
White	82.5%	83%
Black	13%	12%
Hispanic	1.8%	1.5%
Asian	1.8%	2%

Solution: Despite the demographics of the respondents matching the US population, there are many potential sources of bias. We list a few here along with the expected direction of the bias.

- sampling frame - women who do not belong to a group were not reached. We expect this would bias the results toward a lower rate of faithfulness because they might be happier in their relationship than those in a group.
- self selection - the women chose to complete the survey, which was very lengthy. We expect this would bias the results toward lower rate of faithfulness because they have a compelling reason to complete the survey.
- question - We haven't provided this information, but the survey had several confusingly written and/or leading questions, which could also bias the results.

Boys Outperform Girls in Math

In June 2018, the Upshot published an article about the performance of 8^{th} grade girls and boys in school districts across the country. (`https://www.nytimes.com/interactive/2018/06/13/upshot/boys-girls-math-reading-tests.html`) We will simplify the scenario and consider only whether or not the boys outperformed the girls on average in a school district. (The actual data was how much more advanced in their math studies was one group over the other). From the article, we know that "the study included test scores from the 2008 to 2014 school years for 10,000 of the roughly 12,000 school districts in the United States," and we can assume that the data collected were a census since there was no sampling involved.

Identify the following aspects of the sampling process:

14. Target Population – CAUTION - an individual is not necessarily a person. What is being studied here?

Solution: School districts in 2018 in the US

15. Question –

Solution: In a school district, do 8^{th}-grade boys outperform 8^{th}-grade girls on the standardized math test?

16. Sampling Frame –

Solution: School districts that administer the standardized test. This might exclude private schools.

17. Design –

Solution: The design appears to be a census, where all school districts that administer the standardized test reported their results.

18. Confounders – What are some possible ways that we may break down the data to compare more homogenous groups? How do you think the proportion might change in these smaller groups?

Solution: The NY Times story investigated several potential confounding factors. Two of interest were:

- Income - it was found that in school districts with higher income parents, that the difference between boys and girls was the largest (in favor of the boys).
- Race - It was found that school districts that had higher proportions of LatinX and African American students, the difference reversed in favor of the girls.