CS188–Winter 2020 — Homework 1 Solutions

Hermmy Wang, SID 704978214

Collaborators: None

2. Twitter

(a) The first issue is to select a representative sample. He needs to select the samples that are representative of the whole population who use the public transit system. It is possible that many transit riders prefer other social media or do not use any kind of social media at all.

The second issue is the uncertainty of the data. Some comments contain mixed sentiments. A comment might say both positive and negative things about the LA transit system, while emphasizing more on the positive side, or vice versa. He needs to figure out an approach to correctly deal with this kind of comment.

The third issue is about confounding factors. Comments on Twitter related to the public transit system are highly subject to the person's mood on the given day. It is possible that the person encounters other events that make them happier than usual and therefore gives a more positive comment on the public transit.

3. Model Extensibility

(a) No. The model is only shown to perform well on the UK and USA dataset. It is not representative to detect breast cancer in Brazil. It is possible that breast cancer might have different signs or expressions on the patients from Brazil due to geological, cultural, or climate factors. Hence, it is inaccurate to use this model to detect breast cancer in Brazil.

4. Experiment Design

(a)
- student's year in school
- professor teaching the course
- course subject/topic/title
- quarter
- lecture time
- student's credit hours of the current quarter
- number of classes the student has for the lecture day
- transportation means to school
- commuting time to school
- whether student has medical issues

(b) We can use logistic regression to do the prediction. The output is a probability ranging from 0 to 1 that indicates how likely a given student will stop attending a lecture.
The labels of the training dataset will be binary such that a given student attending lecture is represented as 0 and not attending is represented as 1.

(c) We can use simple random sampling without replacement. If we are investigating for any particular lecture at UCLA, we can gather all student's IDs and randomly extract a sample from the whole population. We can survey by sending out emails to these students. To increase response rate, we could use incentives like a free prize draw for any student who take the survey.

5. True or False

(a) False. When data scientists want to investigate a problem without an existing dataset, they go out and conduct surveys to collect data.

(b) False. Data scientists use SQL, Python, R, and many other tools.

(c) False. The top tasks for data scientists are exploratory data analysis, data analysis to answer research question, communicate findings, data cleaning, and data visualization. Less than half of the data scientists spend the majority of their time developing new models.

(d) True. This is one of the drawbacks in mathematical models that they tend to reinforce historical trends.

(e) False. The data is categorical. The integers are discrete and represent ordered categories of different salary levels.

6.Probability

(a) P(X=0) means the probability of drawing no red marble when drawing two marbles randomly without replacement from the jar.

(b) P(X=0, Y=1) is the probability of drawing no red marble and exactly one green marble when drawing two marbles randomly without replacement from the jar.

7. Imputation

- Filling 0: The advantage is that it is easy and fast. The disadvantage is that it does not contain any useful information about the dataset and it affects the overall trend and the mean and median of the dataset. It works well on categorical data where 0 is not the category we are interested. It is not going to work well on numerical data as it affects the mean and median.
- Mode: The advantage is that it is not affected by outliers and augments the effect of the most frequent data. The disadvantage is that it is not well-defined when there are multiple modes. Mode of a small dataset can be misleading and not representative. It performs well on categorical data in larger datasets, but not on numerical data or smaller datasets.
- Mean: The advantage is that it is simple. The disadvantage is that mean is easily affected by outliers. It works better on normal distribution. It does not work that well on skewed data.
- Median: The advantage is that median is not affected by outliers. The disadvantage is that it is not representative of the entire dataset since it only uses one or two elements in the dataset. If the dataset changes on the left or right end, it is likely that median will not be able to reflect this change. It works better if the data is normally distributed. It does not work that well on skewed data.