CS188–Winter 2020 — Homework 1 Solutions

Eugene Choi, SID 905368197

Collaborators: Luca Matsumoto (UID :204726167), Trevor Holt (UID: 204794180)

1. Twitter

(a) Many of the riders utilizing the public transit system might not own a twitter account. Therefore, the data might be strongly biased toward people with twitter accounts leading to an inaccurate representation of overall sentiment.

The model that does sentiment analysis may include hashtag keywords from twitter posts that are unrelated to public transit because the unrelated posts may share the same hashtags as the posts about public transit. This would corrupt our data from the inclusion of topics that are unrelated to the people's sentiment of the public transit system.

Sarcasm may also influence the data because the model is only taking consideration the key words. A person may not actually hate the public transit system, but may just joke about it being horrible. This would make our data unreliable when trying to come to a consensus about what the people think about public transit.

2.

(a) Although the model performs well when trained on both the UK and US data sets, this can just be a coincidence. Therefore, we cannot assume that the pre-trained model will perform well like it did with the UK and the US data sets when it is trained on the UK and Brazil data sets.

3.

(a) The professor's bruinwalk rating, the number of students that have taken another course with the same professor, availability of bruincast, previous coursework of student within the same department of the class,and number of classes the student is taking.

(b) Professor's bruinwalk rating: Float value from one to five

Number of classes the student is taking: Int value

Availability of bruincast: Binary value from zero to one (no or yes)

Number of Students that have taken another course with the same professor: Int value

Previous coursework of student within the same department of the class: Take course number and use it as an int value

(c) Professor's bruinwalk rating: We can web scrape the bruinwalk website

Number of classes the student is taking: Receive data from UCLA student registration

Availability of bruincast: Web scrape CCLE to see if bruincast if offered

Number of Students that have taken another course with the same professor: Receive data from UCLA database

Previous coursework of student within the same department of the class: Run a query on UCLA student database

4.

(a) False: Data scientists start with a question or problem and must decide on a method to use to collect data that is relevant to solving the question or problem at hand.

(b) False: Data scientists use many different tools other than Python, such as SQL and R.

(c) False: Data scientists use most of their time cleaning and analysing their data rather than developing new models.

(d) True: When making decisions about the future, we tend to look at the past first to see the types of actions and trends that either worked or did not work. By inspecting historical data to make a decision about the future, we introduce a historical bias by allowing the results we found to dictate our choice.

(e) False: The income data is quantitative because it is categorical.

5.

(a) (1/2) * (3/5) = (1/5) chance that P(X=0)

(b) ((2/6) * (1/5)) + ((1/6) * (2/5)) = (2/15 chance that P(X=0, Y=1))

6.

(a) Advantage:

Filling with a constant value: This method of imputation would be simple and easy on categorical data. On data sets that have only a few null values, we can choose constants that fit the data the best.

Filling with mean: Filling the missing values with the mean is a very simple method to use and to understand. This method of imputation would be good when the data set does not have large amounts of null values and when correlation between the imputed values is low.

Filling with median: This method of imputation is very fast and simple when filling in missing values. Similar to the advantages with the mean, this method of imputation would be good with data sets that do not have more than a handful of data missing. Less missing data being changed leads to less of an impact on the model, lower skew, and lower bias.

Deleting: This method of augmenting your data set is very simple if your data set has a lot of statistical power. The model we train may not require the data we are missing. This is also good on data sets that are missing lots of categorical data because there won't be a skew.

Disadvantage:

Filling with a constant value: By changing the null values to a constant value like zero, the constant value will have more precedence in the data and will most likely cause a skew. It would also be time consuming to figure out which contant values best fit. This method of imputation would be bad on data sets with many null values.

Filling with mean: This method of imputation would be bad on data with many outliers because it will cause the model to skew. Filling with the mean can also not be done on categorical data. This method of imputation can also cause the standard error to be biased.

Filling with median: The disadvantages are similar to filling with the mean. This method introduces bias when the amount of missing data is large (bad on data with many outliers), can affect the estimation of variance by underestimating, and affects the correlation between variables.

Deleting: If the difference between the incomplete data and the complete data is not random and there is a large difference between them, deleting might not be a good idea because you can introduce bias into your models. Bad on data sets that are missing a handful of data because you may be losing data that affects your model. Those types of data should be imputed because they will not introduce too much bias.