CS148 Homework 1

Prithvi Kannan
UID: 405110096

Collaborators: Vanessa Wang

Problem 1:

- Highly skewed responses. People are more likely to post about their experience on twitter if they had a really good or bad experience versus neutral.
- Twitter audience is not necessarily representative of LA Dept of Health patrons. People on Twitter are likely younger, so this method may ignore older patients.

Problem 2:

- Features
  - past GPA – decimal
  - percent A in professor's class – decimal
  - student interest level – scale of 1-5
  - past experience in subject – scale of 1-5
  - past history of attendance – decimal
- We can create a survey which we can randomly distribute to students to learn about personal information. To ensure coverage of all student segments we distribute the survey over email, since all students have access and must use email to get class information. We can find out past grade distributions via Bruinwalk records.

Problem 3:

- Replace with median – this is one of the easiest ways to handle null values. This can be a great method if there are very few null values, and the data follows a normal distribution. However, on sparse datasets, our median may not be representative of the underlying distribution.
- Replace with a constant value – here we replace all null fields with a predetermined constant or 0. This method works well for categorical features but can introduce bias and doesn't consider any underlying correlations between features. This method works well on data where we cannot compute median/mean.
- Hot deck – in this method we replace null fields with a randomly chosen value from similar individuals in sample. This ensures that the imputed value will be bounded by the true values in the sample while still providing variance in the imputations. However, the value that we impute with may not make sense in all cases. This method is good for preserving randomness within the model.
- Regression – instead of picking an arbitrary value to impute with, we can train a model to predict the missing value given the remaining features. For example, if we are missing weight, but have height, age, and waist measurement, we can train a model to predict

weight using the complete fields. This works well when we have enough additional features related to the feature we are trying to impute. This may not work so well if the data is sparse or if the columns are not related. In regression, we lose variability around predicted values.

Problem 4:

a. It would not make sense to one-hot-encode heartrate. While BPM is discrete, we can use these values as integers and use them as continuous data.
b. No, it would not make sense to one-hot-encode health categories. While values are discrete, the meaning of 1 vs 5 is important, so we use this as continuous data.
c. Yes, it would make sense to one-hot-encode different fashion brands. These values are discrete and there is no way to create an ordering, so this is a perfect candidate for one-hot.
d. Yes, it would make sense to one-hot-encode states if the number of states are low. We cannot create a meaning for "difference" between two states and therefore cannot represent ordinality. However, if we are considering all 50 states, we would be adding 50 different axes to our model. We may want to do something different to group states before one hot encoding, such as categorizing as west coast/central/east coast.
e. Yes. There are a total of 6 combinations (2 directions and 3 times), but there is no ordinal relationship between the different options. Therefore, one-hot-encoding makes sense to understand each combination separately.

Problem 5:

a. False. A very small p-value (< 0.05) provides evidence against the null and in favor of the alternative hypothesis.
b. False. Sometimes all of the necessary features are already represented in the dataset, so augmenting is unnecessary. For example, if trying to predict BMI given height and weight, no other features are needed since the true relationship is that BMI = weight / (height^2).
c. False. If the question being asked is novel enough, then we may also need to create the dataset using standard data collection methods.
d. False. One-hot-encoding adds a new axis for each distinct categorical value of a field, so if there are many possible values then many axis will be added.
e. True. If we are building systems based on biased data, then our new outputs will carry over those biases in predictions.

Problem 6:

a. Either you can draw red first then non-red or non-red first then red. Therefore, $P(X=1) = 2/6 * 4/5 + 4/6 * 2/5$
b. Either you can draw white first then blue or blue first then white. So, $P(X=0, Y=1) = 3/6 * 1/5 + 1/6 * 3/5$