Data Profiling Exam Exercises - Solutions

2020-01-13

- 1 - True. Box plots give a rough idea of the shape of the distribution of a variable by showing where the probability mass lies within the domain of the variable.
- 2 - False. The shape of the distributions of two variables tells us nothing about their correlation.
- 3 - False. We can clearly see that almost all the probability mass lies in the value 1.
- 4 - False. Variable E shows no outliers whatsoever.
- 5 - True. Correlation is a measure of linear dependence between two variables, and because scatter plots depict the *joint* distribution between two variables, we can eyeball how strong their linear relationship is.

2020-01-27

- 1 - False. First, note the significant difference in the relative position of the median within the box. Also, D has an almost nonexistent bottom whisker, showing a great concentration of probability mass on a very small interval, possibly even a single value.
- 2 - True. The data in the different plots are not on the same scale.
- 3 - True. They are depicted by circles in the box plot.
- 4 - False. Box plots literally have specific graphical notation to depict outliers, while in histograms they may be bunched up with other values if bins are too large.
- 5 - False. Histograms allow for greater granularity in observing the distribution of data points.

2021-01-19

- 1 - False. There are more records than dimensions, and we can say that we only face the curse of dimensionality when the number of dimensions is greater than the number of records.
- 2 - True. A false predictor is a variable that is strongly correlated with the output class, but to which we will not realistically have access to at prediction time. In this case we are trying to predict whether a patient will have arrhythmia, so obviously we won't have access to the knowledge of variable Z.
- 3 (**Doubts**) - We have no reason to say that A18 and A30 are redundant. However, I don't think we can say whether A19 and A22 are redundant either, as they have a few records with values different the majority, which could end up being highly predictive of the target variable.
- 4 - True. There is a suspiciously large number of data points at zero, which may suggest using zero as a placeholder for missing values for that variable.
- 5 (**Doubts**) - False. A14 has a large number of missing values. On the other hand, there seems to be nothing wrong with A228; is there?

2021-02-05

- 1 - False. If we look at the histograms, some variables are clearly binary.
- 2 - True. There are only 165 records and some box plots show a relatively large number of outliers.
- 3 - True. It provides the most information gain according to the decision tree.
- 4 (**Doubts**) - I don't think we can say anything about the intrinsic dimensionality of a given dataset without having access to the support of the probability density.
- 5 - (**Doubts**) - How can we tell?

2022-02-10

- 1 - True. Because months are sequential, there is an implicit order between them.
- 2 (**Doubts**) - I don't see why it wouldn't be relevant, even if extremely unbalanced.

- 3 (**Doubts**) - The answer says that there are no outliers, but from the numerical definition used in box plots, they exist. However, if you look at the histogram, it doesn't seem like there are outliers. Should we consider only the numerical definition of outlier or should we also look at the histogram?
- 4 - False. From looking at the correlation matrix, we can't say anything about the relevance of variables.
- 5 - False. The concept of outlier doesn't exist in categorical variables since there is no distance defined for them.

2022-02-26

- 1 (**Doubt**) - Looking at the histogram, we can tell that it is ordinal. However, should we not consider the possibility that it is actually a categorical variable (with no notion of order or distance) that has been incorrectly encoded into integers?
- 2 (**Doubt**) - I have no idea how to answer this.
- 3 (**Doubt**) - I would think that it is relevant because it provides a lot of information gain (as per the decision tree). What does "relevant" mean then?
- 4 (**Doubt**) - In the histogram of 'gg_pos' there are values greater than 30, are those not outliers?
- 5 - False. There are only 16 variables compared to 1000 records.

2023-01-23

- 1 - False. There is no concept of order between regions.
- 2 - False. The correlation between them is very low.
- 3 - True. It correlates somewhat strongly with the output variable.
- 4 - False. At first glance you'd think that there are lots of outliers; however, consider that we have 1000 records available. With that in mind, we can see that only 'bmi' and 'claim' could be considered as having a large number of outliers, but they are a minority of the numeric variables.
- 5 - True. The concept of data leakage is directly related to that of a false predictor, and consists of using information in the model training process which will not realistically be available at prediction time. Re-read the statement until you understand why 'claim' fits that description.

2023-02-10

- 1 - True. There is an implicit order in that variable: it makes sense to say that "two vehicles were involved in the accident" is somehow *greater than* "one vehicle was involved in the accident".
- 2 - False. The correlation between them is only 0.68.
- 3 - False. It does not perfectly split the samples into their respective classes.
- 4 - False. We have no reason to believe that there are a lot of outliers.
- 5 - False. First of all, the variable as described would be extremely sparse and therefore basically useless for most models (unless using a language model). More importantly, 'description' is a false predictor (though I guess that doesn't stop us from using it in *training*).