Mandatory individual assigment

Maria Barrett - ITU username: mbarrett

March 11, 2014

1 Questions

In this assignment I want to find relations between tiredness of snow, self-reported programming skills and operating system.

1. What are frequent patterns of the above mentioned features?
2. How do the data points cluster using these features?
3. Can tiredness of snow be predicted from operating system and age?

2 Preprocessing

The following methods are used: replacing missing values, standardization and binarization.

2.1 Programming skills - ordinal data

All values that are not floatable or are not between 1.0 and 10.0 are replaced with a random[1] number between 1.0 and 10.0. This dataset was used for Apriori. For KNN and K-means I standardized the data using z-score to zero mean and unit variance.

2.2 Tiredness of snow - binary data

All values that did not contain 'yes', 'no' or was 1 or 0 were considered outliers. Outliers were replaced with a random draw from the same distribution as the valid replies. A dataset where the values were represented by 'y' and 'n' was used for apriori. A dataset where values represented by 0 and 1 was used for K-means and KNN.

[1] All random draws were seeded so that my results are reproducible.

2.3 Type of operating system - discrete data

Valid answers: lowercased answers containing “win”, “ubu”, “lin”, “mac” or “osx”. These are assigned the values “w”, “l” or “m” respectively for Apriori. For K-means and KNN these values were binarized using the one hot method (i.e. 0,0,1 for windows, 0,1,0 for linux-based operating systems and 1,0,0 for Mac OSX.)

2.3.1 Problems

Discrete values takes more hardcoding than the other cleaning functions. The function I made can only be used for answers to this specific question, because I look for allowed subpatterns. The other functions (for cleaning ratio and binary data) could be used on other columns without modifications. I could have chosen to make the cleaning of discrete values more general and e.g just try to ensure that answer was letters below 20 characters, but that would increase number of categories and also increase ambiguity. For such a small task it makes most sense to hard code allowed patterns.

2.4 Problems

It’s always suboptimal to replace outliers with random accepted values - preserving the distribution or not. In this case I judged it to be acceptable compared to the alternatives.

3 Apriori

Apriori works best with discrete values. It also wants unique values for each feature.

The longest possible frequent pattern above minimum support threshold of 0.1 were: [’8.0’, ’w’, ’y’], [’7.0’, ’w’, ‘y’]. From these I formulated rules by taking all possible permutations of 2 and deleting those where a value occurred in both i.e (7.0)(7.0) or (7.0),(7.0,y) Or at least I tried. Despite deleting a lot of pointless permuations automatically, some sticked and ended up being calculated. All rules with a confidence above 0.5 can be seen in 7. By use of human sense I removed all rules with duplicate values. This final result can be seen below:

(‘8.0’,)(’y’,): 0.5833, (’7.0’,)(’y’,): 0.8, (’w’,)(’y’,): 0.5854, (’8.0’, ’y’)(’w’,): 0.7143, (’7.0’, ’w’)(’y’,): 0.7143, (’y’,)(’w’,): 0.5854, (’8.0’, ’w’)(’y’,): 0.5556, (’8.0’,)(’w’,): 0.75

3.1 Problems

I encountered a problem with the mixed input types. Since there were more possible values for the programming skills feature the support had to be pretty

low (0.1) to get a value from this feature in order to get a frequent pattern of length 3 (shorter frequent patterns would be boring for making rules).

4 K-means clustering

The result of K-means clustering on the normalized data set is: k = 2 Mean of cluster 1: [-1.23114294, 0.76190476, 0.04761905, 0.19047619, 0.57142857] Mean of cluster 2: [0.56204352, 0.54347826, 0.19565217, 0.26086957, 0.63043478] The order of the features are: [age, 3 values for operating system, tiredness of snow]

5 K-nearest neighbor

The data set was split in a train set (50 datapoints) and a test set (remaining 17 datapoints). I tried to predict tiredness of snow from the remaining features. The prior distribution of the classes are yes: 0.55, no: 0.45. k = 3 Error on test set: 0.352941176471 The error rate is pretty high. The conclusion is that age and operating system poorly predicts whether or not people were tired of snow.

6 Other observations

A simple visual inspection concludes that the first answers were cleaner than the later. An explanation can be that until the nature of the questionaire revealed itself to the student who filled out from the top (and did not read ahead), the respondent may have felt obliged to answer truthfully.

7 Appendix 1

All calculated rules where some contain duplicate values. ('7.0',)('7.0', 'y'): 0.8, ('8.0',)('y',): 0.5833, ('7.0',)('y',): 0.8, ('w', 'y')('y',): 1.0, ('8.0', 'y')('8.0', 'w'): 0.7143, ('7.0', 'w')('w',): 1.0, ('w',)('y',): 0.5854, ('8.0', 'y')('w',): 0.7143, ('7.0', 'w')('y',): 0.7143, ('y',)('w',): 0.5854, ('8.0', 'w')('y',): 0.5556, ('8.0',)('w',): 0.75, ('8.0', 'w')('w',): 1.0, ('y',)('w', 'y'): 0.5854, ('7.0', 'w')('w', 'y'): 0.7143, ('8.0', 'w')('w', 'y'): 0.5556, ('8.0',)('8.0', 'y'): 0.5833