UDACITY

< Return to "Machine Learning Engineer Nanodegree" in the classroom

DISCUSS ON STUDENT HUB

Creating Customer Segments

REVIEW

HISTORY

Requires Changes

7 SPECIFICATIONS REQUIRE CHANGES

You have made a good start here, but a few tweaks are needed in order to meet all the specs. I have added links and suggestions to help you improve these sections and improve your understanding of the concepts. If you still have any issues or questions, you can use the “knowledge” platform on your classroom or discuss it on study groups. We look forward to the next submission, keep up the hard work!

Data Exploration

Three separate samples of the data are chosen and their establishment representations are proposed based on the statistical description of the dataset.

Good job commenting on the establishment that could be represented by each sample point by looking at the dataset statistics.

A prediction score for the removed feature is accurately reported. Justification is made for whether the removed feature is relevant.

Great work getting the R^2 score here and commenting on the feature's relevance.

Suggestion

Note that the score can change based on the value of the random state used. It might be a good idea to run the decision tree multiple times (you can loop over different values of random states, say 1-100) and average out the score from all the runs to get a more accurate score.

Student identifies features that are correlated and compares these features to the predicted feature. Student further discusses the data distribution for those features.

You haven't identified all the correlated features correctly. From the results of running `data.corr()`, which features appear to have a correlation coefficient of >0.5?

> Yes, it confirms my suspicions of Milk being related to other feature

But in Q2, you claimed that Milk is an important feature and cannot be predicted using the other features. Shouldn't the correlation deny your earlier suspicions?

Data Preprocessing

Feature scaling for both the data and the sample data has been properly implemented in code.

```
log_samples = log_data.sample(frac=0.1)
```

You need to take the log transform of the sample points here, which are stored in the variable `samples`. These contain the 3 indices you selected at the beginning for Q1.

Student identifies extreme outliers and discusses whether the outliers should be removed. Justification is made for any data points removed.

You need to identify and mention the indices of all double counted outliers in your answer. Your answer itself is not very clear, as you justify outlier retention in the second point, and then justify outlier removal in the third point. If you think outliers have a negative impact on clustering algorithms, why were they retained in the dataset?

Suggestions

- You could also use a Counter to find these points programatically.
- This paper on the impact of outlier removal on KMeans would be a good read on the topic: http://www.math.uconn.edu/~gan/ggpaper/gan2017kmor.pdf

Feature Transformation

The total variance explained for two and four dimensions of the data from PCA is accurately reported. The first four dimensions are interpreted as a representation of customer spending with justification.

Good work calculating the cumulative variance and interpreting each dimensions individually. You haven't interpreted the negative weighted features though.

Note that feature weights are interchangeable (ref.) (the signs of the PCA weights in a dimension can be flipped on running the code again). So a large weight in any direction (positive or negative) will indicate a higher purchase value for the feature. If two features are opposite in sign (in a dimension), it means that they are inversely correlated, meaning that spending on one increases as that on the other decreases. Both the of these features are important in determining customer spending for the dimension. Similarly, a low feature weight (in positive or negative direction) indicates that the customer buys lesser from this feature.

You can read on some examples of interpreting PCA dimensions from the following links:
https://onlinecourses.science.psu.edu/stat505/node/54
http://setosa.io/ev/principal-component-analysis/

PCA has been properly implemented and applied to both the scaled data and scaled sample data for the two-dimensional case in code.

Clustering

The Gaussian Mixture Model and K-Means algorithms have been compared in detail. Student's choice of algorithm is justified based on the characteristics of the algorithm and data.

Some key differences between the models:

Speed/Scalability

- K-Means is faster and more scalable.
- GMM is slower as it uses information about the data distribution — e.g., probabilities of points belonging to clusters.

Cluster assignment

- K-Means results in hard assignment of points to cluster (as it assumes clusters to be symmetrical spherical shapes)
- GMM results in soft assignment, as it uses more information about the data (it assumes the clusters to be elliptical in shape)

You can read more on the differences between the two models here:
https://www.quora.com/What-is-the-difference-between-K-means-and-the-mixture-model-of-Gaussian

Several silhouette scores are accurately reported, and the optimal number of clusters is chosen based on the best reported score. The cluster visualization provided produces the optimal number of clusters based on the clustering algorithm chosen.

The optimal cluster size has been determined by comparing the silhouette scores for different cluster sizes.

The establishments represented by each customer segment are proposed based on the statistical description of the dataset. The inverse transformation and inverse scaling has been properly implemented and applied to the cluster centers in code.

Nice job determining the establishments by looking at the dataset statistics.

Sample points are correctly identified by customer segment, and the predicted cluster for each sample point is discussed.

You need to revisit this section when the correct log transform of the sample points are taken and the subsequent sections are re-run. Note that there would only be 3 sample points then.

Conclusion

Student correctly identifies how an A/B test can be performed on customers after a change in the wholesale distributor's service.

For this question, you do not need to judge the impact of the delivery change on your own, but describe how an A/B test can be run to get definitive results from the analysis. There are two things to keep in mind when designing an A/B test:

1. You must use the **structured data** to identify which customers will be affected. So it requires you to use the cluster structure to conduct an A/B test.
2. For an A/B test to be effective, the experiment group has to be highly similar to the control group, before the treatment is applied to the experiment group. If they are dissimilar to each other, then the result of the A/B test might be to some variable other than the variable being tested.

You can read more on A/B testing from the following links:
https://en.wikipedia.org/wiki/A/B_testing
https://www.quora.com/When-should-A-B-testing-not-be-trusted-to-make-decisions/answer/Edwin-Chen-1

http://multithreaded.stitchfix.com/blog/2015/05/26/significant-sample/
http://techblog.netflix.com/2016/04/its-all-about-testing-netflix.html
https://vwo.com/ab-testing/

Student discusses with justification how the clustering data can be used in a supervised learner for new predictions.

You have correctly noted that the created customer segments can be used to turn this into a classification problem.

Comparison is made between customer segments and customer 'Channel' data. Discussion of customer segments being identified by 'Channel' data is provided, including whether this representation is consistent with previous results.

You need to revisit this section when the correct log transform of the sample points are taken and the subsequent sections are re-run.

RESUBMIT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

RETURN TO PATH

Rate this review