UDACITY

‹ Back to Machine Learning Engineer Nanodegree

Creating Customer Segments

REVIEW

HISTORY

Requires Changes

5 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.

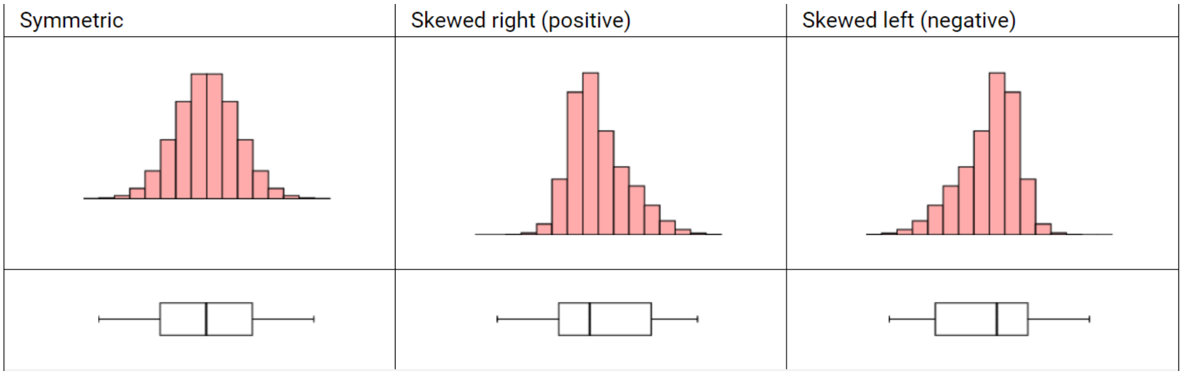
A prediction score for the removed feature is accurately reported. Justification is made for whether the removed feature is relevant.

The score is not very correct, as we expect a higher score for Detergents paper (the correlation matrix in the next section will show that). This is probably because of the use of `cross_val_score`. Note that it splits the data into train/test set of its own for doing cross validation. Hence, using train_test_split before it doesn't make sense (it reduces the dataset size passed to cross_val_score by 75%). You should simply pass in `X` and `y` to this function.

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

The correlated features are correctly identified, but there are a couple of issues in your answer:

- The data is not normally distributed for any feature. If you look at the plot along the diagonals, the distributions do not seem to follow a bell curve. Are they skewed?



- > it is quite evident that Detergents_Paper exhibits a high degree of correlation with Milk and Grocery. Thus, making it very difficult to predict and very relevant.

This is not correct. If a feature has high correlation with the other predictor variables, does it mean that it can be derived by the other features of the dataset? If yes, then is this feature really necessary considering that the information stored in it can easily be obtained by the rest of the data?

On the other hand, if the feature is not that strongly correlated, would it mean that the feature contains unique information that is not captured by the other features? Would this feature be redundant or useful?

The answer to the above questions should help you in deciding the relevance of the selected feature.

Data Preprocessing

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

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

You haven't identified all double counted outliers here, as there were a total of 5 (and not 4). You also need to mention the criteria for removing the two selected outliers (for eg: did you remove all outliers that were 5*IQR away from the first and third quartiles?).

Suggestions

- You could also use a Counter to find these points programatically.
- You should also discuss why outliers should be removed in the first place. How do they negatively impact PCA and clustering algorithms? 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 should also expand on what the inverse correlation between the dominant features (especially in dimensions 3 and 4) imply about the customer spending pattern of those dimensions.

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 mention that all points should be assigned to cluster 0, but they are actually assigned to cluster 1. Doesn't this make the clusterer's results inconsistent with your expectations?

You also need to explicitly compare the values of the sample points with the cluster centers (from Q8) to better evaluate which segment best represents them.

Conclusion

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

Creating a new clustering technique is not recommended, as we have already segmented customers. How would you use this segmented data to run an A/B test?

The A/B test needs to be explained in detail. How many such tests are needed? How will you divide the customers into control and variation groups?

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.

Even though there is some overlap in the central region, the overall alignment is pretty good!

☑ RESUBMIT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

RETURN TO PATH

Rate this review