PROJECT

Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW	CODE REVIEW	NOTES

Requires Changes

1 SPECIFICATION REQUIRES CHANGES

SHARE YOUR ACCOMPLISHMENT

Great job with this project. It looks like your really took the feedback and ran with it. Nice. Unfortunately, there is still one little thing to fix.

For PCA check out:

- http://www-bcf.usc.edu/~gareth/ISL/
- Modeling with Data by Ben Klemens
- Pattern Recognition and Machine Learning by Bishop

Also read up on Singular Value Decomposition, especially in Introduction to Linear Algebra by Strang.

For Bayes:

- Feature Selection for High-Dimensional Data
- Learning Bayesian Models With R
- Machine Learning by Mitchell
- Machine Learning for Audio, Image, and Video Analysis by Camastra, Vinciarelli

Data Exploration

> ✓ Three separate samples of the data are chosen and their establishment representations are proposed based on the statistical description of the dataset.
>
> Great seaborn box plot! Nice visualization of your samples!

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

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

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.

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.

> ✓ 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.
>
> Wow! Great plots. Very nice job visualizing :D

> 🔄 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.
>
> Are you sure the best number of components is 3? From your data frame, it looks like 2 has a higher silhouette score.

> ✓ 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.

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

Conclusion

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

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

> ✓ 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.

RESUBMIT PROJECT

⬇ DOWNLOAD PROJECT

Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

⊙ Watch Video (3:01)

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.

RETURN TO PATH

Rate this review
☆☆☆☆☆