UDACITY

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

Creating Customer Segments

REVIEW

HISTORY

Meets Specifications

Data Exploration

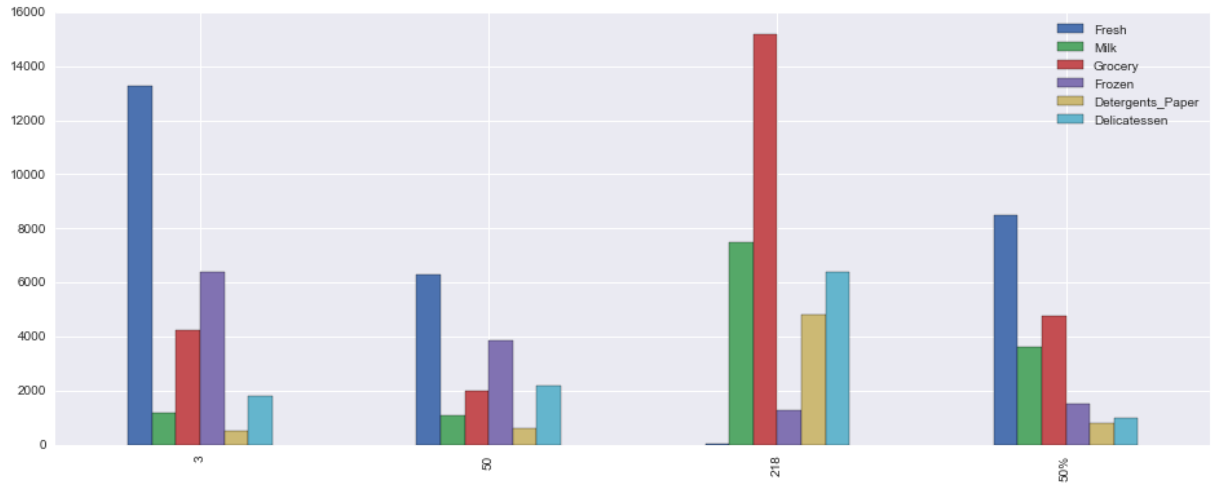
Three separate samples of the data are chosen and their establishment representations are proposed based on the statistical description of the dataset.

Visualize the comparison

Further improvement could be done by comparing the sample points with the dataset's mean.
Please look at the following code blocks:

```
import seaborn as sns
samples_bar = samples.append(data.describe().loc['mean'])
samples_bar.index = indices + ['mean']
_ = samples_bar.plot(kind='bar', figsize=(14,6))
```

You would get the plot similar to the following:



(The three datapoints I am choosing are 3, 50, 218)

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

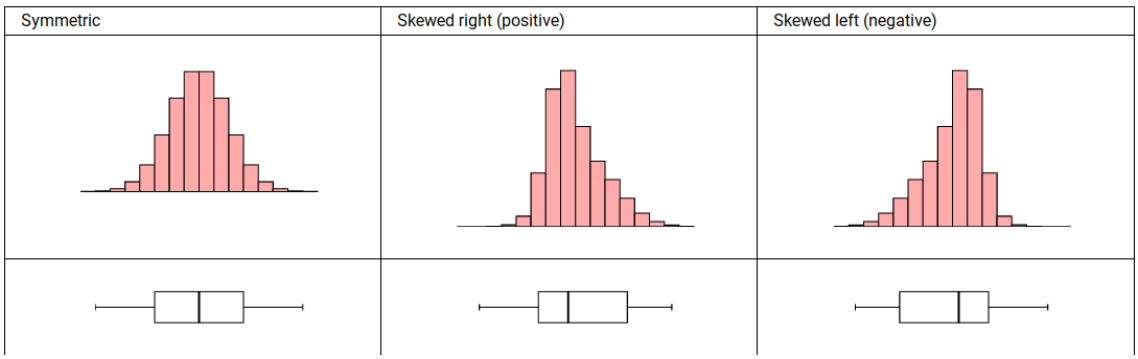
Since R2 is the coefficient of determination, the score is a negative score, thus this means that the feature could not be predicted by other features. The feature is necessary for identifying customers' spending habits if other correlated features are presented.

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

- Grocery <-> Detergents_Paper are correlated.
- Milk <-> Detergents_Paper, and Milk <-> Grocery are also correlated but not to the same degree.

DATA DISTRIBUTION:

- It is noticeable that there are outliers in the dataset, which would lead to the following:
- The dataset is highly skewed;
- Large number of data points are near 0. Moreover, it is noticed that the median falls below the mean. Please look at the following diagram regarding to common data distribution you would see:



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 may consider to make use of the following code snippet to get the repeated outliers:

```
repeated_outliers=[]
for el in counter.elements():
    if counter[el]>1:
        repeated_outliers.append(el)
print "Following records are outliers for more than one feature:", list(se
t(repeated_outliers))
```

Here we find 42 outliers, in 440 samples, nearly 10 percentage of the total number.If we remove all the outlier,we would lose a lots of information.
In this case, we should find some samples that are considered outliers for more than one feature. By use python list and set,we found 5 samples, So we remove those 5 samples instead of remove all the outlier to keep the information in the samples and make sure we have enough samples to analyse.

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.

- As the interpretation of the principal components is based on finding which feature are most strongly correlated with each component, i.e., which of these numbers are large in magnitude for feature weights, the farthest from zero in either positive (positively correlated) or negative (negatively correlated) direction.
- For more information, please look at here for the interpretation of PCA.

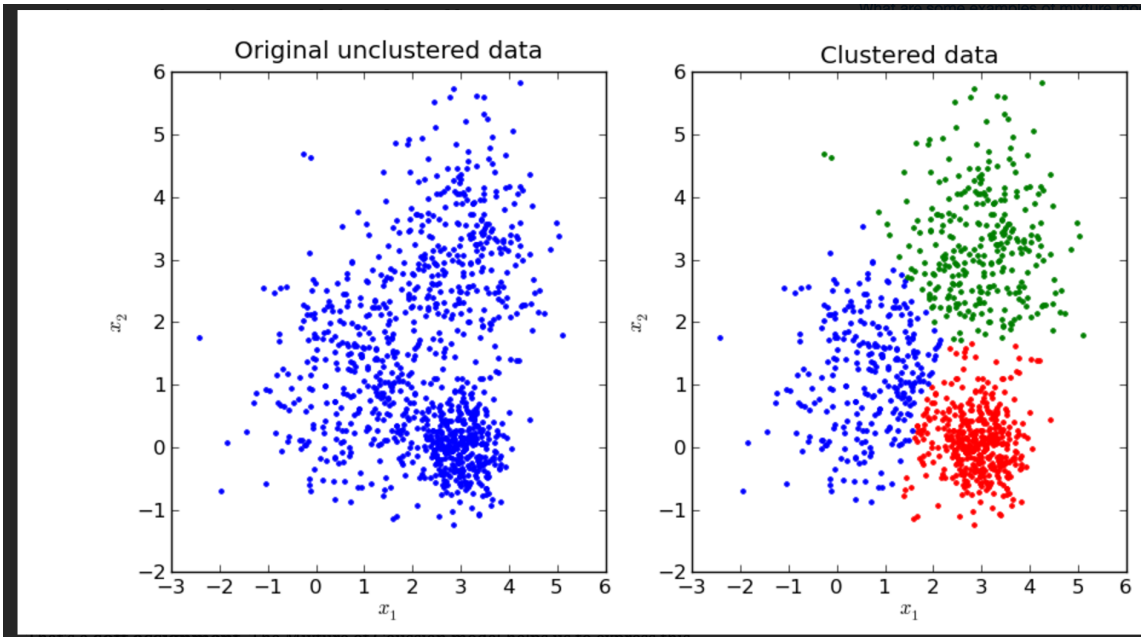
PCA has been properly implemented and applied to both the scaled data and scaled sample data for the two-dimensional case in code.

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.

To be more elaborated, please look at the following example for the difference between GMM and K-Means:

1. Please look at the following diagram:



2. The difference between KMeans and GMM are about hard or soft assignment:

- Let's says we are aiming to break them into three clusters, as above. K means will start with the assumption that a given data point belongs to one cluster.
- Choose a data point. At a given point in the algorithm, we are certain that a point belongs to a red cluster. In the next iteration, we might revise that belief, and be certain that it belongs to the green cluster.
- However, remember, in each iteration, we are absolutely certain as to which cluster the point belongs to. This is the "hard assignment".
- What if we are uncertain? What if we think, well, I can't be sure, but there is 70% chance it belongs to the red cluster, but also 10% chance its in green, 20% chance it might be blue. That's a soft assignment.
- The Mixture of Gaussian model helps us to express this uncertainty. It starts with some prior belief about how certain we are about each point's cluster assignments.
- As it goes on, it revises those beliefs. But it incorporates the degree of uncertainty we have about our assignment.

WHICH ONE SHOULD WE CHOOSE:

- KMeans and GMM should be both working here.
- However, please note that K Means works well when the data clusters are relatively simple in shape, well separated and distinct - so that the clustering is accurate as K-Means does the hard assignment, we are certain each point belongs to which cluster.This algorithm requires the number of clusters to be specified.
- Given enough time, K-means will always converge, however this may be to a local minimum. This is highly dependent on the initialization of the centroids.
- GMM less presume on the shape of the data, which means that GMM would be more general applicable.

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 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.

As an example:

- Choose the n-most central customers for each segment, and assign them to the A group (3-day a week delivery) for a trial period. Outputs for the test could be customer feedback, number of complaints, loss of business, etc.
- These outputs would be measured against the 'B' (5-day a week delivery) group within each cluster and versus the population as a whole.

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.

DOWNLOAD PROJECT

RETURN TO PATH