Meets Specifications

Perfect submission! 

I reviewed your previous submission as well. You have beautifully implemented the suggestions from the previous review! Therefore, there was not much to add in terms of comments.

Good luck for the next project! 

**Data Exploration**

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

**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.**

Nice work elaborating on the PCA dimensions and interpreting them as a representation of customer spending. However, strictly speaking, any PCA dimension, in itself, does not represent a particular type of customer, but a high/low value along the PCA dimension can help differentiate between different types of customers.

**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.**

**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.**

Excellent! You have correctly identified the key point here which is to conduct the A/B test on each segment independently. However, you intend to implement the change to half the population of each segment, which is not at all a good idea. As mentioned in the previous review, A/B testing is actually an experiment performed on small samples from the population, just large enough to get statistically significant results.

**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.**

tive explained variance for the first two and four dimensions. You could also use the following code to compute these values:

print pca\_results['Explained Variance'].cumsum()

**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.**

Good job comparing GMM and KMeans!  
From a practical standpoint, the main criteria for deciding between these two algorithms are the speed v/s second order information (confidence levels) desired and the underlying structure of our data.

**Regarding your choice of algorithm:**

Both the algorithms will do fine here, although considering the fact that there are no visually separable clusters in the biplot, one might, indeed, prefer the soft-clustering approach of GMM, particularly since the dataset is quite small and scalability is not an issue.

For large datasets, an alternative strategy could be to go with the faster KMeans for preliminary analysis, and if you later think that the results could be significantly improved, use GMM in the next step while using the cluster assignments and centres obtained from KMeans as the initialisation for GMM. In fact, many implementations of GMM automatically perform this preliminary step for initialisation.

I provide below some citations which might prove useful, if you would like to go deeper into the dynamics of these algorithms:  
<http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/mixture.html>  
<http://www.nickgillian.com/wiki/pmwiki.php/GRT/GMMClassifier>  
<http://playwidtech.blogspot.hk/2013/02/k-means-clustering-advantages-and.html>  
<http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/K-Means_Clustering_Overview.htm>  
<http://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means>  
<http://www.r-bloggers.com/k-means-clustering-is-not-a-free-lunch/>  
<http://www.r-bloggers.com/pca-and-k-means-clustering-of-delta-aircraft/>  
<https://shapeofdata.wordpress.com/2013/07/30/k-means/>

**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.**

Indeed, number of clusters = 2 gives the best silhouette score among the many considered!

**Important remark regarding the choice of outliers:**

This is one place where your choice of outliers plays a huge role. For example, repeat the analysis without removing any outlier. What is the optimal number of clusters that you get?

**Miscellaneous remarks:**

* If you want to get more support for your results obtained using silhouette analysis, one way is to check how ***balanced*** are the clusters obtained from different values of number of clusters, using the code given at this [link](http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py).  
  Remark that in certain cases, you can even choose a value for number of clusters which gives a sub-optimal score. For example, in the link provided, 2 is not considered optimal, despite having a better Silhouette score, because it doesn't result in ***balanced*** clusters, while 4 does.
* From [sklearn documentation](http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_samples.html), the Silhouette Coefficient is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample. Therefore, it makes sense to use the same distance metric here as the one used in the clustering algorithm. This is Euclidean for KMeans (default metric for Silhouette score) and Mahalanobis for general GMM.
* For GMM, [BIC](http://scikit-learn.org/stable/auto_examples/mixture/plot_gmm_selection.html) could sometimes be a better criterion for deciding on the optimal number of clusters, since it takes into account the probability information provided by GMM. I leave you to experiment with this.

**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.**

**Required:**

As specified in the statement of Q8, please justify your prediction of the establishments represented by the two clusters by comparing *explicitly* to the statistical measures of the dataset.

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

One interesting point to note from the cluster\_visualization plot is that the two clusters are essentially separated by a value on the first PCA dimension, which we saw earlier is predominantly a combination of Detergents\_Paper, Grocery and Milk. The rest of the features, which figure prominently only in the second PCA dimension, don't really matter!

**Conclusion**

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

**Required:**

You are on the right track with your intuition that the proposed change will impact the customers in different segments differently, but could you also briefly describe the implementation of A/B testing to verify your hypothesis?  
In particular, please state precisely how many A/B tests that you would need to run, and identify the experimental and control groups for each test.

To recall, the principle behind A/B testing can be stated as follows:

*A/B testing is an experiment performed on small samples from the population, just large enough to get statistically significant results. In A/B testing, everything besides the testing parameter should remain as similar as possible for both the experiment (A) and the control (B) groups, so that we can study the change in behavior caused by the testing parameter.*

Following links might be of interest here. In particular, the last link discusses A/B testing in the context of clustering:  
<https://www.quora.com/When-should-A-B-testing-not-be-trusted-to-make-decisions/answer/Edwin-Chen-1>  
<http://techblog.netflix.com/2016/04/its-all-about-testing-netflix.html>  
<https://vwo.com/ab-testing/>  
<http://stats.stackexchange.com/questions/192752/clustering-and-a-b-testing>

**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.**

Nice summary of your clustering analysis. And good choice of using GMM, as the clusters do have a fair amount of overlap in reality. Although a perfect classification is not possible to achieve, soft clustering gives us confidence levels in our predictions, which would understandably be low at the boundary between two clusters.

**Code tip:**

You can calculate the accuracy score for clustering using the following code:

channel\_labels = pd.read\_csv("customers.csv")["Channel"]

channel\_labels = channel\_labels.drop(channel\_labels.index[outliers]).reset\_index(drop = True) - 1

# channel\_labels = abs(channel\_labels -1)

from sklearn.metrics import accuracy\_score

accuracy = accuracy\_score(channel\_labels,preds)

Note that I've subtracted 1 from channel\_labels, because the given channel\_labels are 1 and 2, while our cluster-labels are 0 and 1.  
Also, note that the assignment of labels - 0 and 1 - in the clustering algorithm is completely arbitrary. Therefore, you might have to keep or remove channel\_labels = abs(channel\_labels -1) in the above code, to ensure that the cluster and channel labels are "compatible".