UDACITY

Logo

Back to Machine Learning Engineer Nanodegree

Machine Learning Capstone Project

REVIEW | CODE REVIEW | HISTORY

Requires Changes

5 SPECIFICATIONS REQUIRE CHANGES

This is a very cool analysis and a great read to a very real world and practical problem. You have demonstrated a full understanding of the entire machine learning pipeline and your report definitely gets the readers attention with the results you have achieved. You just have to expand in a few of these sections, but will greatly improve your report. Please check out some of these other ideas presented here and we look forward in seeing your next submission!!

Definition

Student provides a high-level overview of the project in layman's terms. Background information such as the problem domain, the project origin, and related data sets or input data is given.

Nice work here with your opening section, as you have given good starting paragraphs to outline the project and have provided background information on the problem domain. Definitely a good unsupervised learning problem.

And you have provided good research to back your claims. It is always important to provide similiar research on such a topic.

The problem which needs to be solved is clearly defined. A strategy for solving the problem, including discussion of the expected solution, has been made.

> "The goal is get the list of providers who should be investigated for opioid prescription. "

Problem statement is clearly defined here, and glad that you mention that this is a clustering problem in this section.

And very nice job mentioning you machine learning pipeline here, as this gives the reader some ideas in what is to come in your report and how you plan on solving this important task.

Metrics used to measure performance of a model or result are clearly defined. Metrics are justified based on the characteristics of the problem.

Silhouette Coefficient is a great metric to use for such an unsupervised learning problem. Would suggest providing a bit more analysis in terms of why Silhouette Coefficient is chosen over other unsupervised learning metrics. But great analysis of how this computed.

Analysis

If a dataset is present, features and calculated statistics relevant to the problem have been reported and discussed, along with a sampling of the data. In lieu of a dataset, a thorough description of the input space or input data has been made. Abnormalities or characteristics about the data or input that need to be addressed have been identified.

Very nice job describing your dataset. Glad that you show some descriptive stats, show a sample of your data, go into a bit of detail in the features here and how you 'cleaned' up this dataset. As this allows the reader to get an understanding of the structure of the data you are working with. This is definitely a common thing we need to do in most datasets like this!

Maybe also look into computing the Kolmogorov-Smirnov test for goodness of fit. (https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.kstest.html)

Need any data transformations for the features?

A visualization has been provided that summarizes or extracts a relevant characteristic or feature about the dataset or input data with thorough discussion. Visual cues are clearly defined.

Nice visuals here. The scatter plots and histograms are fine ideas.

> "Applying feature scaling by apply natural log to the values we get the following description of the data"

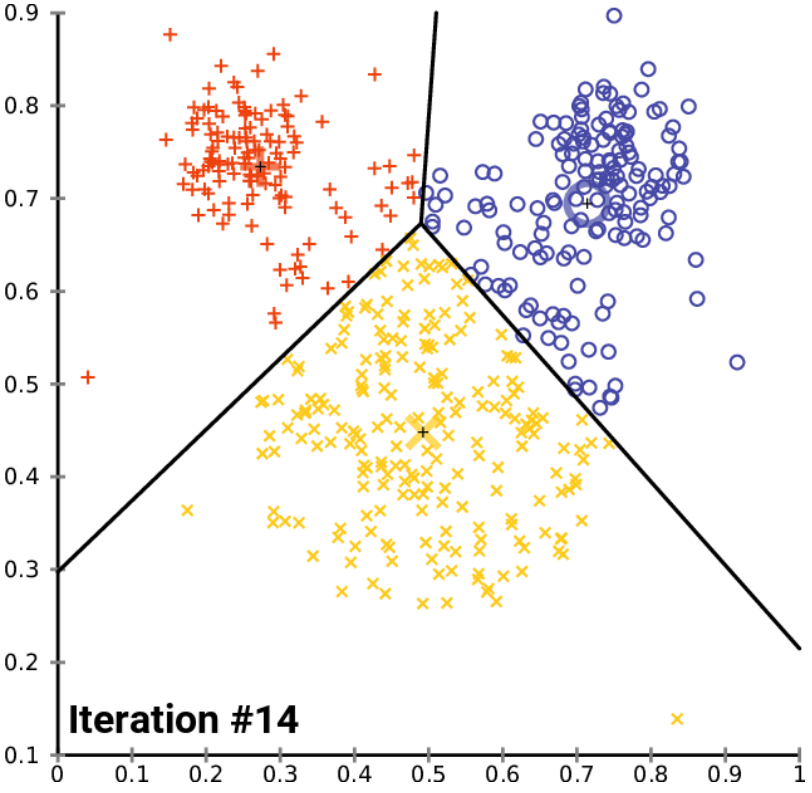
Could also check out the 'normality' of the features with a quantile-quantile (q-q) plot in terms of why applying the natural log would be a good idea

```
import scipy
import matplotlib.pyplot as plt
for feature in data.columns:
    scipy.stats.probplot(data[feature], plot=plt)
    plt.title(feature)
    plt.show()
```

Algorithms and techniques used in the project are thoroughly discussed and properly justified based on the characteristics of the problem.

In this Algorithms and Techniques section, you will need to thoroughly discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain.

Therefore please go into detail in how your algorithms of GaussianMixture, KMeans, MiniBatchKMeans and PCA work and why they are appropriate for solving such a problem (include that here instead of your Implementation section). Don't be afraid to use mathematical formulas or visuals to explain these, if desired.



Student clearly defines a benchmark result or threshold for comparing performances of solutions obtained.

> "There might be several of models that have been developed since I am using 2014 CMS data and since CMS itself came up with models in 2010 to identify the fraudulent providers. The office of inspector general of department of Health and Human Services has come up with list of excluded providers, this is called exclusion list. Now please keep in mind that this exclusion list is for all types of frauds/misdiagnosis/malpractices not just Opioid over prescription. Goal is identify the list of providers who shall be further investigated.
> OIG exclusion list, list of providers identified as fraudulent
> https://oig.hhs.gov/exclusions/exclusions_list.asp
> https://oig.hhs.gov/exclusions/downloadables/UPDATED.csv"

Remember that your benchmark model needs to be run on the exact same dataset that you are using, so your benchmark results are comparable to your 'final optimized' model's results. It would be better to run a simple model such as some type of hard-coded solution to segment these individuals.

Or even something like Hierarchical clustering

Methodology

All preprocessing steps have been clearly documented. Abnormalities or characteristics about the data or input that needed to be addressed have been corrected. If no data preprocessing is necessary, it has been clearly justified.

Nice work with your PCA plot and biplot. You can also visualize the percent of variance explained to get a very clear understanding of the drop off between dimension. Here is a some starter code, as np.cumsum acts like `+=` in python.

```
import matplotlib.pyplot as plt
x = np.arange(1, len(data.columns))
plt.plot(x, np.cumsum(pca.explained_variance_ratio_), '-o')
```

The process for which metrics, algorithms, and techniques were implemented with the given datasets or input data has been thoroughly documented. Complications that occurred during the coding process are discussed.

> "I will also go with Gaussian Mixture Model clustering"

What are the Silhouette Coefficient scores for the GMM clustering algorithm? You have done a great job of this for K-Means, but what about GMM?

Also for this section make sure you mention if any complications occurred during the coding process. Anything go wrong here? Was there any part of this process that was more difficult than the others? etc.. Stating any complications that occurred is always an important step in replicating your work.

The process of improving upon the algorithms and techniques used is clearly documented. Both the initial and final solutions are reported, along with intermediate solutions, if necessary.

> "I shall put this in a data frame and pick the record with max "total explained variance" for first two dimensions and max silhouette_score. This is what I would get"

This is an excellent idea here, however in this **Refinement** section, you need to also show some of the *other* results as well for comparison. Try showing more than just one row in your `SearchForMaxSilhouette` table. This is needed to give justification for why this is the optimal combination. And this will also show the beauty of your chart.

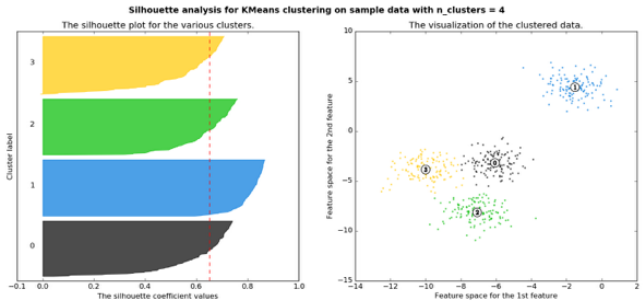
Results

The final model's qualities — such as parameters — are evaluated in detail. Some type of analysis is used to validate the robustness of the model's solution.

You have good analysis of your final models and you have provided some good analysis to validate the robustness of the model's solution with the different types of frauds, misdiagnosis and malpractices providers. Although, not ideal, it does give some good insight into the model.

Another cool interpretation method for Silhouette score is like this

(http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html)



The final results are compared to the benchmark result or threshold with some type of statistical analysis. Justification is made as to whether the final model and solution is significant enough to have adequately solved the problem.

Make sure you include a Justification section. In this section, your model's final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project.

Conclusion

A visualization has been provided that emphasizes an important quality about the project with thorough discussion. Visual cues are clearly defined.

A fine visual here. We can also add the median values from the data and very easily visualize the cluster centroids with a pandas bar plot

```
true_centers = true_centers.append(data.describe().ix['50%'])
true_centers.plot(kind = 'bar', figsize = (16, 4))
```

Student adequately summarizes the end-to-end problem solution and discusses one or two particular aspects of the project they found interesting or difficult.

You will need to summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work.

Discussion is made as to how one aspect of the implementation could be improved. Potential solutions resulting from these improvements are considered and compared/contrasted to the current solution.

> "We can definitely add more features such as income level to further narrow down the investigation list. "

For sure! Especially with an 'interpretable' dataset like this one, understanding your dataset, creating features, and important pre-processing steps are always needed!

Could also look into using the MeanShift algorithm to *discover* the optimal K.

Quality

Project report follows a well-organized structure and would be readily understood by its intended audience. Each section is written in a clear, concise and specific manner. Few grammatical and spelling mistakes are present. All resources used to complete the project are cited and referenced.

Your writing is very clean and it is very easy to understand what you are saying. I personally thank you as this report is very easy to read :)

Code is formatted neatly with comments that effectively explain complex implementations. Output produces similar results and solutions as to those discussed in the project.

Code does look good. Ipython notebooks are great to use.

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)

RETURN TO PATH

Student FAQ