Machine Learning Engineer Nanodegree

Capstone Proposal

Nirupama PRV

Proposal

Domain Background

Education

PASSNYC is a not-for-profit organization that facilitates a collective impact that is dedicated to broadening educational opportunities for New York City's talented and underserved students. PASSNYC uses public data to identify students within New York City's under-performing school districts and, through consulting and collaboration with partners, aims to increase the diversity of students taking the Specialized High School Admissions Test (SHSAT).

The objective is to assess the needs of students by using this data to quantify the challenges students face in taking the SHSAT. By segmenting and clustering, we can identify the schools where minority and underserved students stand to gain the most from services like after school programs, test preparation, mentoring, or resources for parents.

My personal motivation in selecting this domain is two-fold. I work for a company matching high school athletes to sports programs, so I wish to understand clustering better as applicable to school programs. Also, as an online student and micro-volunteer with school programs, it is of personal interest too.

Problem Statement

In this project, the objective is to describe the variation in the different types of schools present in NYC. Based on demographic diversity or lack thereof other features of the school, the idea is to create categories of that represent subsets that together represent all the schools. Doing so will better equip these programs and how best to serve the schools.

We shall be using techniques such as Gaussian Mixture model, K-means clustering and PCA to get the best way to categorize the data.

Datasets and Inputs

The dataset(s) and/or input(s) being considered for the project is a public dataset from Kaggle and can be found at link below:
https://www.kaggle.com/passnyc/data-science-for-good/home
https://www.kaggle.com/new-york-city/nyc-school-district-breakdowns/

This is a dataset hosted by the City of New York and has demographic statistics broken down by school districts.

Solution Statement

Here, our solution is to determine the different segments present in the dataset. We will quantify the "goodness" of the clustering by calculating each data point's silhouette coefficient. The silhouette coefficient for a data point measures how similar it is to its assigned cluster from -1 (dissimilar) to 1 (similar). Calculating the mean silhouette coefficient will provide for a simple scoring method for our clustering and we shall take the best silhouette score of the different algorithms used.
Silhouette score can be computed using sklearn library.

Benchmark Model

For this project, the benchmark model is the kernel from the Kaggle dataset at:
https://www.kaggle.com/laiyipeng/target-schools-action-recommended-for-passnyc
It was one of the winners when the competition (using this dataset) originally ran. There are 3 clusters present in this analysis. I shall try to compare my model with this analysis.

Evaluation Metrics

For this project, the metric to be used for evaluation is silhouette scores. I will be using sklearn library to compute these values.

The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a, b). To clarify, b is the distance between a sample and the nearest cluster that the sample is not a part of. Note that Silhouette Coefficient is only defined if number of labels is 2 <= n_labels <= n_samples - 1.
This function returns the mean Silhouette Coefficient over all samples. To obtain the values for each sample, use silhouette_samples.
The best value is 1 and the worst value is -1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar.

Project Design

For this project my tentative workflow shall be as follows:

- Explore dataset.
- Conduct feature relevance and distributions.
- Apply different clustering algorithms, namely K means clustering, Gaussian Mixture model and PCA. And for benchmarked model.
- Select model with cluster numbers that delivers best results and similar or higher scores than the benchmark model.
- Cross validate using randomly selected data points.

Resources:
- http://scikit-learn.org/stable/modules/clustering.html
- http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
- http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

- https://en.wikipedia.org/wiki/K-means_clustering
- https://www.quora.com/What-are-the-advantages-of-K-Means-clustering
- http://scikit-learn.org/stable/modules/mixture.html
- https://www.quora.com/What-are-the-advantages-to-using-a-Gaussian-Mixture-Model-clustering-algorithm