INFO 1998 Project D Deliverable

Hierarchical Clustering, K-Means Clustering, PCA

Guideline and Rubric

Release Date: November 8th
Due Date: November 19th
Submit Through: CMS

Overview

For this project, we will further expand on our repertoire of machine learning algorithms. Specifically, we will focus on clustering and PCA. You will be using the same train.csv from the housing dataset for this project as well. Remember that clustering is an unsupervised machine learning algorithm, so there should be no labels used to train the models. Also recall that we use PCA as a mathematically motivated way to "extract" the most important features from a large dataset. It may be beneficial to use this reduced feature space to train unsupervised models, though this is not true in all cases.

Because you are doing unsupervised learning, your goal is for the clustering models you create to learn latent variables. You should also produce one or more visualizations for each model. There is no graded baseline or accuracy measure since this is unsupervised learning, so your paragraph to explain your thinking is important this project. Some things to consider when trying to show that the clustering models extract meaningful results are checking whether the clusters are ones you would expect and if they make sense to you. Additionally, try different hyperparameters to see how these affected your model results.

Don't forget to submit a **brief** paragraph on each of your models. This paragraph should discuss what kind of conclusions you can draw from your dataset, as well as the motivation for your procedure. Though there is no baseline for unsupervised learning, we expect that your paragraphs will demonstrate that the clusters formed are meaningful in some way, and what PCA tells you about the dataset.

Models

For the **hierarchical clustering model** you are aiming to find groupings/labels to place your data into. So you can perform your clustering on the whole dataset.

For the **K-means clustering model** you can decide how you would like to perform your clustering. You can try to cluster on the entire dataset or drop columns and see how well your algorithm clusters the data. Take a look at the iris example in the notes for more details.

For the **PCA** you are aiming to find the most relevant features from your dataset. In this sense, you are aiming to reduce the dimension of the feature space.

What to Submit:

A Jupyter Notebook containing:

- Code for the hierarchical clustering
- Appropriate dendrogram visualization
- One paragraph on your hierarchical clustering algorithm
- Code for the K-Means Clustering
- Appropriate clustering visualization
- One paragraph on your K-Means clustering algorithm
- Code for the PCA
- Appropriate PCA visualization
- One paragraph on your PCA algorithm

Grading Rubric

Criteria	Points
Dendrogram (Hierarchical Clustering)	
Correct algorithm used	3
Dendrogram is accurate and interpretable - Parameter selection - Appropriate visualization(s) produced	12
Paragraph explanation - Should include why you did whatever you did and what insights you gained from the process/results	5
K-Means Clustering	
Correct algorithm used	3
Clustering is accurate and interpretable - Appropriate variable chosen - Parameter selection - Appropriate visualization(s) produced	12
Paragraph explanation - Should include why you did whatever you did and what insights you gained from the process/results	5
PCA	
Correct algorithm used	3
PCA is accurate and interpretable - Parameter selection - Appropriate visualization(s) produced	12
Paragraph explanation - Should include why you did whatever you did and what insights you gained from the process/results	5
Total	**60**