Machine Learning

3rd Assignment - Shahid Beheshti University

May 12, 2023

Due date: May 31

*** You are required to write a detailed report for implementation tasks. ***

1. What is the curse of dimensionality and how does it affect clustering?
2. In what cases would you use regular PCA, incremental PCA, randomized PCA, or random projection?
3. Does it make sense to chain two different dimensionality reduction algorithms?
4. What are the main assumptions and limitations of PCA?
5. How can clustering be used to improve the accuracy of the linear regression model?
6. How is entropy used as a clustering validation measure?
7. What is label propagation? Why would you implement it, and how? **(Extra Point)**

8. You are going to work on the [Supermarket dataset for predictive marketing](#). Your task is to use clustering to segment the customers into distinct groups based on their shopping behavior and demographics.
   - Explore and preprocess the dataset. This may involve handling categorical variables and normalizing or scaling numerical features and feature engineering.
   - Use K-means clustering to identify the optimal number of clusters. Experiment with different values of K and use metrics such as the elbow method and silhouette score to evaluate the performance of the clustering.
   - Visualize the clusters and analyze their characteristics. This may involve plotting the clusters in 2D or 3D using PCA or t-SNE.
   - Experiment with other clustering algorithms such as DBSCAN or hierarchical clustering, and compare their performance with K-means.
   - Try to reduce data dimensionality using PCA before training your model, use different numbers of components and report their effects. **(Extra Point)**
   - Interpret the results and provide insights to the store owners. What are the distinct customer segments that have been identified? How can the store owners use this information to improve their marketing strategy, product offerings, or customer experience? **(Extra Point)**