CS 4375 Similarity & Ensemble

Group 1:
EFA190000
WXA20000
DXA190032
UXA180002
YFA190000

Part 5(a, b): Classification, Regression, and Clustering

Regression, Classification and Clustering are set of Machine learning algorithms that perform different roles. Regression and Classification are supervised machine learning algorithms that work by feeding a program a set of labeled data where input data is paired with a correct output. Through repetitions, these algorithms are trained to accurately predict future outputs based on some input. There are two types of Supervised learning algorithms. Classification is a technique where the goal is to predict the class or category of an input data point based on a set of features. Regression is a supervised learning technique used to predict continuous numerical values. The goal is to learn a relationship between the input features and a continuous output variable, such as predicting the price of a house based on its features or predicting the salary of an employee based on their experience and education. Clustering is a type of unsupervised machine learning algorithms learn from unlabeled examples, where the input data does not have any target variable. The goal is to discover patterns, structure, or relationships in the data. In this discussion, I will discuss two supervised learning algorithms namely KNN and Decisions tress as well as introduce a few examples of clustering algorithms.

kNN is a non-parametric algorithm used for classification and regression tasks. In the case of classification, the algorithm tries to predict the class of a new observation by looking at the k-nearest neighbors (kNN) of that observation in the training set. The kNN are the data points that are closest to the new observation in terms of their distance metric. The algorithm calculates the distance between the new observation and all the training set observations and selects the k observations that are closest to the new observation. Once the kNN are identified, the algorithm then assigns the new observation to the class that is most frequent among its kNN. In the case of regression, the algorithm calculates the average value of the k-nearest neighbors and assigns that as the predicted value for the new observation. The choice of the value of k is important in kNN algorithm. A smaller value of k will result in a more flexible model that fits the training data closely but may not generalize well on unseen data. A larger value of k will result in a smoother decision boundary and less overfitting but may not capture the fine details of the training data.

Decision Trees are tree-like structures used for classification and regression. The algorithm creates a decision tree by recursively splitting training data based on the values of the input features that result in the maximum information gain or minimum impurity. The impurity measures how well a feature separates the classes in the data. The split that results in the maximum information gain or minimum impurity is chosen at each step, and the process continues until a stopping criterion is met, such as reaching a maximum depth or minimum number of samples at a leaf node. After decision tree is constructed, the algorithm uses it to predict the class or value of a new observation by traversing the tree from the root node to a leaf node. At each node, the algorithm tests the value of a feature and moves to the left or right branch depending on the result of the test, until it reaches a leaf node that corresponds to a predicted class or value. Benefits of the decision tree includes its ability to handle both qualitative and quantitative data. Additionally, it is unaffected by missing values by treating them as a separate category or by imputing the missing value based on the majority class or average value at the node. Decision trees can also handle non-linear relationships between the input features and the output variable and can capture interactions between features.

Clustering is an unsupervised learning technique in machine learning where an algorithm group associated data into clusters. Some popular clustering algorithms include k-means, hierarchical clustering, and model-based clustering . K-means clustering partitions a given dataset into K clusters, where K is a pre-defined number of clusters. The algorithm assigning data points to its closest cluster center and then updates the enter based on the new data. The closer two data points are to one another, the more related they are. Hierarchical clustering creates a hierarchy by recursively partitioning data points into smaller and smaller clusters, until all the data points belong to their own individual clusters. Finally, model-based clustering uses predefined statistical models to estimate the parameters of the underlying data distribution, and then assigns each data point to the cluster that maximizes the likelihood of the data given the model.

Part 5(c): PCA and LDA

Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) dimensionality reduction algorithms in machine learning. They reduce the size of a data set while maintaining its characteristics as much as possible.

PCA identifies the underlying structure in the data by transforming the original features into a new set of uncorrelated features, called principal components. The algorithm steps are as follows:

- Standardization: Where data is standardized by subtracting the mean and dividing by the standard deviation.
- Covariance Matrix: A matrix that quantifies the relationships between the original features.
- Eigenvectors and Eigenvalues: Computed from the covariance matrix. Eigenvectors represent the maximum variance in the i=distribution. Eigenvalues represent the amount of variance explained by each eigenvector.
- Principal Components: The principal components are constructed by projecting the original data onto the eigenvectors.

LDA identifies the features that best separate the classes in data. Its goal is retaining as much of the class discrimination information as possible after reducing the data set. The algorithm steps are as follows:

- Standardization: The data is standardized by subtracting each item with the mean and dividing by the standard deviation.
- Between-Class Scatter Matrix: A matrix that quantifies the differences between the classes.
- Within-Class Scatter Matrix: A matrix that determines the variance in each class.
- Eigenvectors and Eigenvalues: A product of the inverse of the within-class scatter matrix and the between-class scatter matrix are computed.
- Linear Discriminants: The projection of the original data onto the eigenvectors.

In conclusion, PDA and LDA are excellent reduction algorithm with a few draw backs. PCA is limited to linear data sets is highly sensitive to outliers while LDA can be useful for feature selection, pattern recognition, and classification. However, it assumes that the underlying data is linear and that the classes have equal covariance matrices.