CS 4395 Intro to Machine Learning

Dr. Karen Mazidi

Students:

- Leo Nguyen - ldn190002
- Simon Kim - sxk190106
- Abed Ahmed - asa190005
- Dylan Kapustka - dlk190000

Searching for Similarity

For Classification:

KNN

In order to vote for the most popular label (in the case of classification) or average the labels, KNN first calculates the distances between a query and all the instances in the data. Then it chooses the K examples closest to the query (in the case of regression). By locating the nearest neighbor class, the K-nearest neighbor classifier algorithm's basic version predicts the target label. Euclidean distance is used to determine which class is closest to the point that has to be classed.

Decision Trees

A decision tree is a graphical depiction of every option for making a choice depending on certain circumstances.We attempt to create a condition on the features for each step or node of a classification decision tree in order to fully separate all of the labels or classes present in the dataset.In decision trees, we begin at the tree's root when anticipating a record's class label. We contrast the root attribute's values with that of the attribute on the record. We follow the branch that corresponds to that value and go on to the next node based on the comparison.

- **For Regression:**

- Even though, both kNN and Decision tree can work both on classification and regression. It does not work well with regression as both of them are non-parametric models. They can not provide the coefficients on the result. However, they can provide a very good visualization of the data. In general, kNN regression is built by calculating the distance between points. The number of points will be decided by k value. As you can see, that is a lot of computation, so the kNN does not work well with large data sets and high numbers of dimensions. It is also sensitive with outliers and missing values as it requires number type data to work with. The decision tree works by splitting the observations into smaller partitions until all the observations in a group are similar. However, it will not use any attribute to split. The model

will compute the entropy, information gain, and Gini Index first to select the information attributes as the checking to split the observation. The other will be omitted or abundant.

Clustering:

K-Means Clustering: K-Means is a clustering technique that divides input data points into k groupings. The learning algorithm's training phase is represented by this grouping procedure. The outcome would be a model that, in response to the training that the model underwent, accepts a data sample as input and returns the cluster to which the new data point belongs. How does this help? Well, in a very basic way, that's how content promotion and suggestion generally function. Websites may decide to group users into bubbles or clusters based on qualities or actions they have in common. By doing this, the suggested material will be relatively relevant since individuals who have previously engaged in similar activities are more likely to be interested in related information.

Hierarchical Clustering: Hierarchical clustering, commonly referred to as hierarchical cluster analysis, divides items into clusters based on how similar they are. The result is a collection of clusters, each of which differs from the others while having things that are generally similar to one another.

Each observation is first treated as a separate cluster in hierarchical clustering. Then, it continually completes the next two actions: First, determine the two clusters that are most similar to one another, and then combine those two clusters. This iterative procedure carries on until all of the clusters are combined.

Model Based Clustering: A statistical method for grouping data is known as model-based clustering. It is assumed that the observed (multivariate) data was produced by a limited combination of component models. A probability distribution, often a parametric multivariate distribution, characterizes each component model.

Each component, for instance, is a multivariate Gaussian distribution in a multivariate Gaussian mixture model. The cluster to which an observation belongs is determined by the component that generated that observation

Model-based clustering is an attempt to improve the fit between a mathematical model and the supplied data. It is based on the theory that data is created by combining elements of a simple probability distribution.

1. **How PCA and LDA work. And why they might be useful techniques for machine Learning**

PCA reduces data dimension by projecting data on the axis with a straight line in the direction with the most significant data variance. LDA has the same concept of "projecting data" but uses a method that maximizes variance between classes to find axes that maximize class separation and minimizes internal class variance to find axes.

Both PCA and LDA are used to reduce the dimension of the data. The dimension of data means an attribute, and the degree of dispersion between data increases as the attribute increases. Therefore, it reduces the difficulty of analyzing data by lowering the dimension quickly to judge the tendency of distant data, such as creating a planar interstellar map to see the stars of the three-dimensional universe at a glance. The low difficulty of data analysis is helpful because it affects execution time.