John Mazon
DATA 622
Homework 4 - Final Project
5/19/22

2022 Spring Term (1) Machine Learning and Big Data DATA 622 001

Select one of the methodologies studied in weeks 1-10, and one methodology from weeks 11-15 to apply in the new dataset selected:
In the final homework I utilized K-Means clustering, PCA, Decision Trees and Classification

To complete this task:.
- describe the problem you are trying to solve:

The dataset selected contains information on all 802 Pokemon from all Seven Generations of Pokemon. The information contained in this dataset include Base Stats, Performance against Other Types, Height, Weight, Classification, Egg Steps, Experience Points, Abilities, etc. The information was scraped from http://serebii.net/
We desired to try to solve the problem of comparison between pkm that are legendary vs non-legendary. Comparison of attributes. Focus on capture rate to classify.

The K-means clustering algorithm is used to group unlabeled data set instances into clusters based on similar attributes. It has a number of advantages over other types of machine learning models, including the linear models, such as logistic regression and Naive Bayes. A lot of real-world data comes unlabeled, without any particular class. The benefit of using an algorithm like K-means clustering is that we often do not know how instances in a data set should be grouped. The meat of the K-means clustering algorithm is just two steps, the cluster assignment step and the move centroid step. If we're looking for an unsupervised learning algorithm that is easy to implement and can handle large data sets, K-means clustering is a good starting point. Most of the popular machine learning packages contain an implementation of K-means clustering. Based on my experience using K-means clustering, the algorithm does its work quickly, even for really big data sets. Elbow is one of the most famous methods by which you can select the right value of k and boost your model performance. We also perform the hyperparameter tuning to chose the best value of k. Let us see how this elbow method works. It is an empirical method to find out the best value of k. it picks up the range of values and takes the best among them. It calculates the sum of the square of the points and calculates the average distance.
PCA is primarily used for dimensionality reduction in domains like facial recognition, computer vision, image compression, and finding patterns in the field of finance, psychology, data mining, etc. PCA is used to extract the important information out of the dataset by combining the redundant features. These features are expressed in the form of new variables termed principal components. Since the visualization of the features in the dataset is limited, we can also use PCA to reduce the dimensionality of the dataset to 2 or 3 principal components and then visualize to get a better insight into it. It is important to perform data scaling before running PCA on the dataset. Because if we use data of different scales, then we end up missing leading

principle components. To do so, you need to perform mean normalization, and optionally you can also perform feature scaling. Overfitting mainly occurs when there are too many variables in the dataset. So, PCA helps in overcoming the overfitting issue by reducing the number of features. It is very hard to visualize and understand the data in high dimensions. PCA transforms high dimensional data to low dimensional data (2 dimension) so that it can be visualized easily. We can use 2D Scree Plot to see which Principal Components result in high variance and have more impact as compared to other Principal Components.

Decision Trees are useful supervised Machine learning algorithms that have the ability to perform both regression and classification tasks. It is characterized by nodes and branches, where the tests on each attribute are represented at the nodes, the outcome of this procedure is represented at the branches and the class labels are represented at the leaf nodes. Hence it uses a tree-like model based on various decisions that are used to compute their probable outcomes. These types of tree-based algorithms are one of the most widely used algorithms due to the fact that these algorithms are easy to interpret and use. Apart from this, the predictive models developed by this algorithm are found to have good stability and a decent accuracy due to which they are very popular. Classification tree methods (i.e., decision tree methods) are recommended when the data mining task contains classifications or predictions of outcomes, and the goal is to generate rules that can be easily explained and translated into SQL or a natural query language.A Classification tree labels, records, and assigns variables to discrete classes. A Classification tree can also provide a measure of confidence that the classification is correct. A Classification tree is built through a process known as binary recursive partitioning. This is an iterative process of splitting the data into partitions, and then splitting it up further on each of the branches.