**Summary Report on Machine Learning**

As an assignment, two types of algorithms have been chosen from the list of algorithms provided by the instructor. First one from supervised and the other from unsupervised learning.

**SVM, Support Vector Machines** are set of supervised learning methods used for classification, regression and outlier detection. In this type of algorithm both the input data and the corresponding labeling are given. This means that the learn data has to be labeled beforehand. SVM work well for discrete labels even though it can be modified to do regression (Support Vector Regression).

A simple linear SVM classifier works by making a straight line between two classes. The data points on one side of the line represent a category and the data points on the other side of the line will be put into a different category.

The SVM can be used or applied to detect cancerous cells based on images or in the case of regression model they can be used for predicting future driving routes.

**K-nearest neighbors (KNN)-**Clustering- it is one type of unsupervised learning in which no labels are provided to the learning algorism. It classifies objects. There is no general way to define an optimal value for 'k'. This k value depends on the data. As a general rule we can say that increasing 'k' reduces the noise but on the other hand makes the boundaries less distinct.

The picture depicted below visually shows in a simple manner how K-nearest neighbor classifier works. (I thought visualization is also a good tool for presentation!).

To find out the unknown or which animal it might be we have to find the neighbors. The answer depends on the value of ‘K’. If k= 1 it would be cat or if k=4 it would also be a cat. Only the nearest neighbors are considered.



Application of K-nearest neighbors (KNN) is usually in character recognition and image analysis. For example, it can be used in linguistics to recognize misspelled word. Also, online systems use it for recommending products and movies.

Main steps of a kNN algorithm

1: Define *k*

2: **while** (stopping criterion is not met) **do**

3: Compute distances from other data points to point *i*

4: Sort the computed distances

5: Select the *k* points with smallest distances

6: Assign the test point to the class by the simple majority

7: Return the class

8: **end while**

Disadvantages- it is not suitable for big data sets. If *k* is too small, then it may lead to overfitting and more sensitive to noise in data, whereas large *k* values may lead to higher bias and lower accuracy because it may include samples that are not actual neighbors. For distances dominated by dimensions that have largest ranges, rescaling and normalization is needed.