**Machine Learning Model with k-Nearest Neighbor (kNN) Classification using Pyhton**

**Introduction**

knn is a non-parametric classification and instance nbased learning algorithm that can be used for either classification or regression, in either case the input parameters consists of the k closest training features in the data set and the output depends on whether kNN is used for classification or regression. The aim for classification is to predict a class label , which id a selection from a predefined list of possibilities , while regression is used to predict a continuous or floating point number.

Consider a training set N samples given by {𝑋1,𝑋2,…,𝑋N } with its corresponding outputs {y1, y2,….yN where each sample data Xi contains M features denoted as {xai, x2i,…xMN}, then the yq that gives the smallest Euclidean distance dq

Where *dq= ((xa1 -xq1)2*+(xb2-xq2)2+…(xMN-xqN)2)1/2

We can then generalize the above by considering k neighbors such that the target of a given feature {xai , xbi…. xMi} is the mode of state yi based on the k-th nearest neigbor and in the case of a tied number of states , the state woth the largest population in the training set is priotised. The number of nearest neighbors is coded as n\_neigbors in Python’s sklearn library and is the most important tuning parameter in a knn-based classifier.

**Illustration of k-Nearest Neigbour using sklearn to determine the Contraceptive Method Choice for Women**

This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of interview. The problem is to predict the current contraceptive method choice (no use, long-term methods, or short-term methods) of a woman based on her demographic and socio-economic characteristics.

Number of Instances: 1473

Number of Features: 9

* Wife's age (numerical)
* Wife's education (categorical) 1=low, 2, 3, 4=high
* Husband's education (categorical) 1=low, 2, 3, 4=high
* Number of children ever born (numerical)
* Wife's religion (binary) 0=Non-Islam, 1=Islam
* Wife's now working? (binary) 0=Yes, 1=No
* Husband's occupation (categorical) 1, 2, 3, 4
* Standard-of-living index (categorical) 1=low, 2, 3, 4=high
* Media exposure (binary) 0=Good, 1=Not good

Target: Contraceptive method used (class attribute) 1=No-use 2=Long-term 3=Short-term

The steps involved are outlined below.

1. Understanding the question. This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of interview. The problem is to predict the current contraceptive method choice (no use, long-term methods, or short-term methods) of a woman based on her demographic and socio-economic characteristics.
2. Prepare Data. There were three possible contraceptive methods, No-use, Long-term and Short term, which made the task a three-class classification problem. The possible methods are called classes in the classification problem, and the class of a contraceptive method is called its label. The dataset is contained in a .csv file which was imported into the Jupyter notebook, there are a total of 10 columns with the last column, referred to as the target column containing the output, while the rest are referred to as the features. The columns will be split using sklearn library into X and Y component, corresponding to features and target variables, into two groups of training and test data usually in 0.75 ratio. The split of the dataset into a training set, to build our model, and a test set, is to evaluate the behavior of the model with unseen data.
3. Step 3: Tuning the AI parameter. The k nearest neighbors classification algorithm, makes predictions for a new data point by considering its closest neighbor(s) in the training set. The algorithm is implemented in the *KNeighborsClassifier* class, which contains the algorithm to build and make predictions using the model. The class was initiated by setting the parameters. Then the model was built by calling the fit method, passing the training data X\_train and training outputs y\_train as parameters.
4. Step 4: Evaluate the model. It can be observed that the maximum accuracy on the test set is around 57%, observed over the range of (18-40) *n\_neighbour* values. Although, the classifier is considered successful given that the test accuracy > 1.25\*Pcc (~44%), the model only predicts the correct results ~58% of the time, thus isnt extremely accurate in predicting the contraceptive method used based on demographics and socio-economic characteristics.

Refrences

*\*\* Sources: \*\* (a) Origin: This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey (b) Creator: Tjen-Sien Lim (*[*limt@stat.wisc.edu*](mailto:limt@stat.wisc.edu)*) (c) Donor: Tjen-Sien Lim (*[*limt@stat.wisc.edu*](mailto:limt@stat.wisc.edu)*) (c) Date: June 7, 1997*