MACHINE LEARNING CSE-6363-003

Programming Assignment – 1

Accuracy Table:

DATASETS	Euclidean Distance	Polynomial kernel	Radial basis kernel	Sigmoid kernel
iris.data	96.0%	96.0%	95.99%	33.33%
wdbc.data	92.74%	92.92%	62.65 %	58.76%
yeast.data	58.98%	58.58%	58.78%	58.51%

Comparisons and Observations :-

NOTE : The mean accuracies are fluctuating every time you run the program and calculating the mean accuracy(mean_accuracy()) of the fold accuracies(list).The values in the table are the most appeared value while running the program more than once.

Also , run the kernel function or Euclidean distance function in get_neighbors function to get the respective accuracy result.

Adjustable parameters

I tried different values of alpha , beta , p (polynomial factor) , sigma . These are all adjustable parameters, used in different kernel distance formulas .

Alpha = used (1/feature length) value or sometimes value between 0 and 1(for avoiding math domain error, ex: 0.90, doesn't affects the result).

Beta = value range (0 – 4). Mostly used → beta=0

p(polynomial factor) = value range (1-4). Mostly used → p=2

sigma = Values between 0 and 1. Mostly used → sigma=0.30

These are the adjustable parameters, adjusted to find the better accuracy of the model.

Values

Fold size = length of dataset / folds

Folds = 10 (folds value for cross validation)

K (neighbor range) = 3 (tried different values too, like, for yeast dataset used bigger value as the dataset is bigger)

Conclusion of the Observation : -

The K- Nearest Neighbor model with K fold cross validation using Euclidean distance is giving better accuracy on iris and Wisconsin Breast Cancer dataset than the accuracy on yeast dataset.

The K- Nearest Neighbor model with K fold cross validation(K-nn model) using **polynomial kernel** distance on iris and Wisconsin Breast Cancer dataset is giving better accuracy results than conventional K-nn model.

K-nn model using **Radial basis kernel** distance, doesn't giving any better result than the conventional model. By changing the adjustable parameters of kernel function you may get the better accuracy result, like in yeast dataset but the adjustable parameters plays major role here. Therefore, can't be sure of better result.

K-nn model using **Sigmoid kernel** distance, doesn't give better results and the accuracy fluctuation range is big enough to say that the kernel is not giving better and steady results for following datasets. For example, Sigmoid kernel function, for iris dataset gives bad accuracy result and for Wisconsin Breast Cancer dataset the accuracy fluctuates between 37 to 62 percent, which is not satisfiable enough.

Therefore, Like the conclusion of the referred research paper ,

(Kernel Nearest Neighbor)

Reference :

https://www.researchgate.net/profile/Kai_Yu7/publication/220578072_Kernel_Nearest_Neighbor_Algorithm/links/02e7e533620f1ac895000000/Kernel-Nearest-Neighbor-Algorithm.pdf

Three data sets were used for experimenting. It is inconvenient to adjust the parameters in sigmoid kernel because there are two parameters that may dissatisfy the Mercer condition. Thus, in our experiments, only polynomial kernel was used.

Mercer Condition: According to the Mercer condition, if K(x,y) is positive semi-definite, it can be a kernel. According to the Hilbert Schmidt theory. K(x ,y) can be an arbitrary symmetric function that satisfies the Mercer condition.

I observed that for datasets (like Iris, Wisconsin Breast Cancer) only **Polynomial kernel** gives **better** and **satisfiable** accuracy results. For Yeast dataset it gives accuracy quite like accuracy of K-nn model using Euclidean Distance.

References:

- https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/
- https://machinelearningmastery.com/implement-resampling-methods-scratch-python/
- https://www.researchgate.net/profile/Kai_Yu7/publication/220578072_Kernel_Nearest_Neighbor_Algorithm/links/02e7e533620f1ac895000000/Kernel-Nearest-Neighbor-Algorithm.pdf
- https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
- https://archive.ics.uci.edu/ml/datasets/Iris
- https://archive.ics.uci.edu/ml/datasets/Yeast