AISHWARYA VAIDYANATHAN
1001255978

ASSIGNMENT 2

Abstract:

In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms. In this assignment, we have implemented k-NN algorithm on three different datasets whose data are first normalized, 10 FCV is performed and then provided as input to the k-NN. Analysis of the output is also being performed.

Introduction:

k-NN has been used in statistical estimation and pattern recognition already as a non-parametric technique. k-NN is a non-parametric lazy learning algorithm, when you say this, it means that it does not make any assumptions on the underlying data distribution. This is useful, as in the real world, most of the practical data does not obey the typical theoretical assumptions made.

It is also a lazy algorithm which means that it does not use the training data points to do any generalization. In other words, there is no explicit training phase or it is very minimal. This means the training phase is fast. Lack of generalization means that k-NN keeps all the training data. More exactly, all the training data is needed during the testing phase. This contrasts with other techniques like SVM where you can discard all non-support vectors without any problem. Most of the lazy algorithms especially k-NN makes decision based on the entire training data set (in the best case a subset of them).

The dichotomy is obvious here, there is a non-existent or minimal training phase but a costly testing phase. The cost is in terms of both time and memory. More time might be needed as in the worst case; all data points might take point in decision. More memory is needed as we need to store all training data.

The k-Nearest Neighbors:

k-nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If k = 1, then the case is simply assigned to the class of its nearest neighbor.

In this assignment, we have implemented three distance functions based on which we are classifying the data. Euclidian, Polynomial Kernel and Radial Basis Function.

ASSIGNMENT 2

- Euclidian distance:

$$\mathrm{d(x,y)} = \|x-y\|$$

- Polynomial Kernel:

$$K(x,y) = (1+<x,y>)^{p}$$

- Radial Basis Function:

$$K(x,y) = e^{\{-(\frac{\|x-y\|^2}{\sigma^2})\}}$$

$$d^2(\varphi(x),\varphi(y)) = K(x,x) - 2K(x,y) + K(y,y)$$ (Hilbert’s Function)

Datasets:

Sr#	Source	# Instances	# Attributes	Type of Attributes
1	Ecoli	336	8	1. mcg: McGeoch's method for signal sequence recognition. 2. gvh: von Heijne's method for signal sequence recognition. 3. lip: von Heijne's Signal Peptidase II consensus sequence score. Binary attribute. 4. chg: Presence of charge on N-terminus of predicted lipoproteins. Binary attribute. 5. aac: score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins. 6. alm1: score of the ALOM membrane spanning region prediction program. 7. alm2: score of ALOM program after excluding putative cleavable signal regions from the sequence. 8. locsite (class attribute)
2	Glass	214	10	1. RI: refractive index 2. Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10) 3. Mg: Magnesium 4. Al: Aluminum 5. Si: Silicon 6. K: Potassium 7. Ca: Calcium 8. Ba: Barium 9. Fe: Iron

				10. Type of glass: (class attribute) • building_windows_float_processed • building_windows_non_float_processed • vehicle_windows_float_processed • vehicle_windows_non_float_processed (none in this database) • containers • tableware • headlamps
3	Yeast	1484	9	1. mcg: McGeoch's method for signal sequence recognition. 2. gvh: von Heijne's method for signal sequence recognition. 3. alm: Score of the ALOM membrane spanning region prediction program. 4. mit: Score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins. 5. erl: Presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen). Binary attribute. 6. pox: Peroxisomal targeting signal in the C-terminus. 7. vac: Score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins. 8. nuc: Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins. 9. class (class attribute)

Note: Removed sequence number attribute while feeding data to the K-NN algorithm as it’s a nominal data which does not contribute while classifying instances.

ASSIGNMENT 2

Results and Analysis:

K-NN results for Ecoli Data:

K value	Euclidian distance	Polynomial Kernel	RBF Kernel
3	84.0326340326	83.7296037296	84.0326340326
5	85.5944055944	85.8974358974	85.5944055944
7	85.8974358974	86.8065268065	85.8974358974
9	87.4125874126	86.8065268065	87.4125874126
11	85.2913752914	85.5944055944	85.2913752914

- In Euclidian Distance, we found that 10FCV k-NN with k value accuracy is increasing up to k=9 then again decreasing
- Polynomial Kernel and RBF Kernel also demonstrates the same phenomenon of accuracy value is increasing with k value up to k=9 and then decreasing.
- In RBF Kernel, maximum accuracy was found for k = 9

K-NN results for Glass Data:

K value	Euclidian distance	Polynomial Kernel	RBF Kernel
3	52.4952380952	49.7904761905	48.2857142857
5	50.0380952381	45.3523809524	44.8761904762
7	49.7904761905	46.8571428571	45.4285714286
9	49.3142857143	46.4571428571	45.9047619048
11	49.7904761905	48.3619047619	46.8571428571

- Compared to the Ecoli data results in the previous section, k-NN algorithm was showing less accuracy while classifying instances of Glass data.
- One of the reason is Glass data has less number of instances available and they are not representing population distribution.

K-NN results for Yeast Data:

K value	Euclidian distance	Polynomial Kernel	RBF Kernel
3	53.3019203414	53.7820056899	53.3019203414
5	55.9263869132	55.4640825036	55.9263869132
7	57.8129445235	57.4146514936	57.8129445235
9	57.2012802276	57.469772404	57.2012802276
11	57.3364153627	56.9345661451	57.3364153627

ASSIGNMENT 2

- Both Euclidian and RBF Kernel are displaying approximately the same accuracy for a given k value. The highest accuracy is achieved for k=13.
- Yeast data has less accuracy compared to Ecoli data. The reason following this analysis is non-relevant attributes are dominating over relevant attributes in classification of instances of Yeast data.

Conclusion:

- Euclidian and RBF Kernel from the above analysis provide better results than Polynomial Kernel with K-NN.
- Knowledge regarding the domain of the data is important for attribute selection process.

References:

- https://elearn.uta.edu/bbcswebdav/pid-5156036-dt-content-rid-47332888_2/courses/2168-MACHINE-LEARNING-88317-002/Kernel_Nearest_Neighbor_Algorithm.pdf
- http://www.saedsayad.com/k_nearest_neighbors.htm
- http://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/