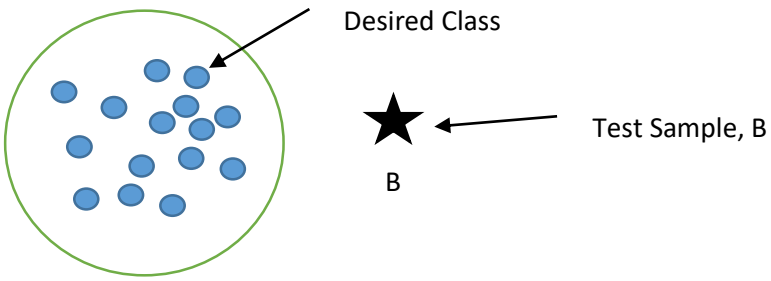
Question 1

The algorithm



The data set is trained with only one class e.g. the class setosa in the iris data set. The output expected should be the class label of the test sample B.

- Select the data set to be used as the training with only one class classification.
- Choose the threshold for rejecting or accepting the class label, say 2.
- Select k and find the nearest neighbor of **B,** say **C,** calculate the distance between the two, say ***d_b*.** The distance calculated here we used the Euclidean distance.
- Find the nearest neighbor of **C,** say **D,** and calculate the distance between the two, say ***d_c*.**
- Calculate the distance ratio as $d_r = d_b/d\_c$
- If $d_r >$ threshold, then reject class label (label unknown) else accept the class label.

Step 1

Verify the algorithm using the iris data set with 10 train instances and 5 test instances of class ***setosa.***

The following set of steps were followed and the validation of the algorithm done on excel.

- Select 15 instances, the first 15 instances of the iris data set are of class setosa so we selected the first 15 instances.
- Choose the first 10 of the data instances to be the training data set (with the class label) and the remaining last 5 instance to be the test data set (without the label).
- Select a threshold (limit) to verify rejection or acceptance of a class label. We set our threshold to be **2** given the kind of the range of the values of the data set
- Find the nearest neighbor of each of 5 points $b_1, b_2, \dots, b_5$ and calculate the distance between the nearest point and the data point to be classified as $d_{b_1}, d_{b_2}, \dots, d_{b_5}$.
- The nearest points of the respective points, $c_1, c_2, \dots, c_5$, find their nearest neighbor and calculate their distance as done the point above, $d_{c_1}, d_{c_2}, \dots, d_{c_5}$.
- Calculate the distance ratio $d_r = d_{b_i}/d_{c_i}$ for $i = 1, 2, \dots, 5$
- If $d_r >$ threshold, then reject class label (label unknown) else accept the class label.

The algorithm was able to tell which data set belongs to the class label already known (setosa) when random data set was used and it was able to tell the data set that does not belong to the class label

(class versicolor). The only problem is if we didn't know the class label was class versicolor, we will just get the result of rejected or does not belong to the class label.

Please see the excel file for the second part

Question 2

QUESTION 2) [CENG-420: 70 Points] [ELEC-569A: 50 Points]

The database below is from Movies Night dataset. Each row has a collection of movies watched by a group of users. What association rules can be found in this set if the target minimum support (i.e. coverage) is 60% and the target minimum confidence (i.e. accuracy) is 80%?

T1: King Arthur, American Pie, Daredevil, Batman vs Superman

T2: Cinderella, American Pie, Batman vs Superman, Enchanted,

T3: Daredevil, American Pie, Cinderella, Enchanted, Batman vs Superman

T4: Batman vs Superman, American Pie, Daredevil

Answer:

Itemset Support

Itemset	**Support**
American Pie	**4**
King Arthur	**1**
Daredevil	**3**
Batman Vs Superman	**4**
Cinderella	**2**
Enchanted	**2**

Itemset	Support
American Pie	4
Batman vs Superman	4
Daredevil	3

Itemset	Support
American Pie, Batman vs Superman	4
American Pie, Daredevil	3
Batman vs Superman, Daredevil	3

Itemset	Support
American Pie, Batman vs Superman, Daredevil	3

Frequent Itemset {American Pie, Batman Vs Superman, Daredevil} = {American Pie, Batman Vs Superman}, {American Pie, Daredevil}, {Batman Vs Superman, Daredevil}, {American Pie}, {Batman Vs Superman}. {Daredevil}

1. American Pie --> Batman vs Superman 4/4
2. Batman vs Superman --> American Pie 4/4
3. American Pie --> Daredevil 3/4
4. Daredevil --> American Pie 3/3
5. Batman vs Superman --> Daredevil 3/4
6. Daredevil --> Batman vs Superman 3/3
7. American Pie --> Batman vs Superman AND Daredevil 3/4
8. Batman vs Superman AND Daredevil --> American Pie 3/3
9. American Pie AND Daredevil --> Batman vs Superman 3/3
10. Batman Vs Superman → American Pie AND Daredevil 3/4
11. Daredevil --> American Pie AND Batman Vs Superman 3/3
12. American Pie AND Batman Vs Superman --> Daredevil 3 /4

Final rules

Confidence >=80%

1. American Pie --> Batman vs Superman 4/4
2. Batman vs Superman --> American Pie 4/4
3. Daredevil --> American Pie 3/3
4. Daredevil --> Batman vs Superman 3/3
5. Batman vs Superman AND Daredevil --> American Pie 3/3
6. American Pie AND Daredevil --> Batman vs Superman 3/3
7. Daredevil --> American Pie AND Batman Vs Superman 3/3

Question 3

How is kNN different from kmeans clustering?

Answer:

kmeans algorithm partitions a data set into clusters such that a cluster formed is homogeneous and the points in each cluster are close to each other. The algorithm tries to maintain enough separability between these clusters. Due to unsupervised nature, the clusters have no labels.

kNN algorithm tries to classify an unlabeled observation based on its k (can be any number) surrounding neighbors. It is also known as lazy learner because it involves minimal training of model. Hence, it doesn't use training data to make generalization on unseen data set.

Which of the following is true for neural networks?

(i) The training time depends on the size of the network.

(ii) Neural networks can be simulated on a conventional computer.

(iii)Artificial neurons are identical in operation to biological ones.

(a) all of them are true.

(b) (ii) is true.

(c) (i) and (ii) are true.

Answer:

The answer is (c).

The training time depends on the size of the network; the number of neuron is greater and therefore the number of possible 'states' is increased. Neural networks can be simulated on a conventional computer but the main advantage of neural networks - parallel execution - is lost. Artificial neurons are not identical in operation to the biological ones.