COMP7990 (Section 1 and Section 2)

Quiz

Write all your answers in your answer books

Question 1 (30 Marks)

1.1 **(10 Marks)** Suppose that we have the following data: [200, 300, 400, 600, 1000], normalize the data by min-max normalization by setting new min to 0 and new max to 1.

Answer: since $x_i' = \frac{(x_i - min)}{max - min}(max_{new} - min_{new}) + min_{new} = \frac{(x_i - 200)}{1000 - 200}1 + 0$, *the answer is [0, 1/8, 1/4, 1/2, 1] (i.e., 0, 0.125, 0.25, 0.5, 1).*

1.2 **(10 Marks)** Suppose we have the following values for prices (already sorted in increasing order): [4, 10, 15, 21, 21, 22, 25, 28, 30, 31, 31, 32]. Use smoothing by bin means to smooth the above data using equal-width binning with 4 bins. Illustrate your steps.

Answer:
1. Compute the width of each bin: (32-4)/4 = 7.
2. Partition data into bins (equal-width):
Bin 1 ([4, 11)): [4, 10]
Bin 2 ([11, 18)): [15]
Bin 3 ([18, 25)): [21, 21, 22]
Bin 4 ([25, 32]): [25, 28, 30, 31, 31, 32]
3. Computing the mean of each bin:
Bin1: 7
Bin2: 15
Bin 3: 21.33
Bin 4: 29.5
4. Smoothing by mean
[7, 7, 15, 21.33, 21.33, 21.33, 29.5, 29.5, 29.5, 29.5, 29.5]

1.3 **(10 Marks)** In real-world data analytic problems, data with noisy values for some attributes is often occurred. Why noisy values are common in real-world application? Describe various methods for handling noisy data.

Answer: Noisy data may due to: (1) Errors in data collection devices; (2) Wrong input; and (3) Technology limitation. Methods for handling noisy data: (1) Binning; (2) Regression; and (3) Clustering.

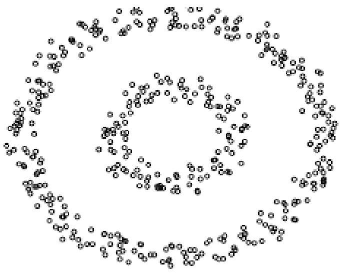
Question 2 (40 Marks)

2.1 (**10 Marks**) Describe the classification process of *K*-nearest neighbor algorithm.

Answer: The classification processes of knn contains the following three steps:

(1) Compute the distance between a test sample and each training samples
(2) Sort by distances and get the k nearest neighbor of the test sample
(3) Use majority vote to predict the class label of the test sample.

2.2 **(10 Marks)** Given the following dataset, will *k*-means work well on it? Explain your answer in detail.



*Answer: k-means will **not work** well on this dataset. K-means works only if the clusters are round shaped. This data contains nonlinear structure that k means algorithm cannot capture.*

2.3 (**10 Marks**) Explain the reason why the nonlinear activation function is required in a neural network model.

Answer: Without non linear activation function, the neural network will be reduced to a linear model.

2.4 (**10 Marks**) Discuss the differences between perceptron algorithm and support vector machines.

Answer: SVM aims to find the separating hyperplane with maximum margin whereas perceptron does not incorporate this maximum margin principle. Perceptron algorithm will not converge on non-separable data where SVM can. SVM can solve nonlinear classification problem whereas the perceptron algorithm can not.

Question 3 (30 Marks)

3.1 Suppose that we want to cluster eight data points as shown in the following table (x_1 and x_2 are the two features) into three clusters.

ID	x_1	x_2
A1	2	10
A2	2	5
A3	8	4
A4	5	8
A5	7	5
A6	6	4
A7	1	2
A8	4	9

The distance function is Euclidean distance. Suppose initially we assign A1, A4, A7 as the center of each cluster. Run the *k*-means algorithm for 1 iteration, at the end of the iteration show

(a) **(10 marks)** The cluster assignments (i.e., which samples belong to which clusters)

Answer:

	Cluster1 center: A1	Cluster2 center: A4	Cluster3 center: A1
A1	**0**	$\sqrt{13}$	$\sqrt{65}$
A2	$\sqrt{25}$	$\sqrt{18}$	**$\sqrt{10}$**
A3	$\sqrt{72}$	**$\sqrt{25}$**	$\sqrt{53}$
A4	$\sqrt{13}$	**0**	$\sqrt{52}$
A5	$\sqrt{50}$	**$\sqrt{13}$**	$\sqrt{45}$
A6	$\sqrt{52}$	**$\sqrt{17}$**	$\sqrt{29}$
A7	$\sqrt{65}$	$\sqrt{52}$	**0**
A8	$\sqrt{5}$	**$\sqrt{2}$**	$\sqrt{58}$

cluster 1: {A1}; cluster 2:{A3, A4, A5, A6, A8}; cluster 3: {A2, A7}

(b) **(10 marks)** The center of the new clusters

Answer: new cluster centers {2, 10}, {6, 6}, {1.5, 3.5}

3.2 **(10 marks)** Describe at least two limitations of *k*-means algorithm. And how to address these two limitations?

Answer: (1) K-means is extremely sensitive to cluster center initialization. We can try multiple initialization and choose the best result to address this limitation; (2) Works only if the cluster are round shaped and of equal size/density cluster. Probabilistic cluster methods can address this limitation; (3) Can not capture non-linear structure. Kernel k-means can address this limitation; (4) Makes hard assignments of points to clusters. Probabilistic cluster methods can address this limitation.