亞洲大學___________考試試題

文件編號：AC-0054
機密等級：2 限閱

考試日期：　　年　　月　　日

考試科目	機器學習	**系所別**	人工智慧博士學位學程	命題教師	
*請同學每題作答，準備好 ppt，考試後繳交檔案予院辦。					

1. For each of the following tasks, identify which type of learning is involved (supervised, unsupervised, or reinforcement) and the training data to be used.

   Junk email recognition.

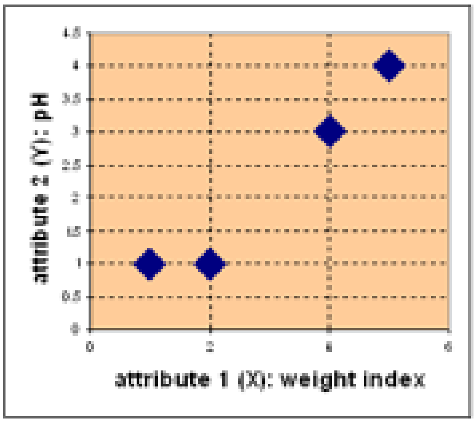
   Stock market forecast

   Medical image analysis

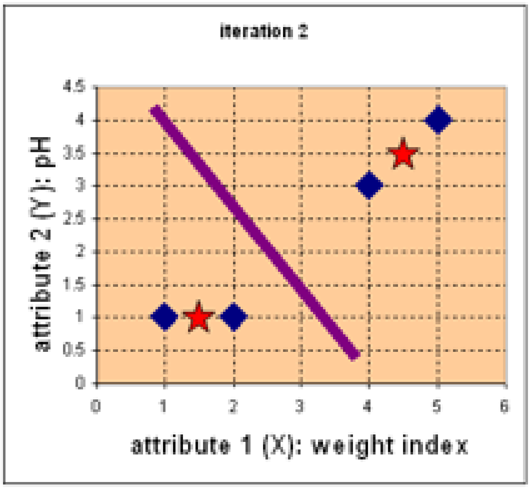
   Machine learning for self-driving car
2. What is the Perceptron? Why can we use the Perceptron to learn from data?
3. Describe the main idea of logistic regression, and use a practical example to illustrate your description.
4. Please explain the following keywords for evaluating machine learning model: confusion matrix, , Accuracy, Precision, Recall & F1 Score ROC curve, and AUC.
5. What is overfitting and when dose overfitting occur? How to deal with overfitting?
6. Suppose we have 4 types of medicines and each medicine has two attributes or features as shown in table below. Our goal is to group these objects into K=2 group of medicine based on the two features (pH and weight index).

	attribute 1 (X):	attribute 2 (Y)
	weight index	pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

Each medicine represents one point with two attributes (X, Y) that we can represent it as coordinate in an attribute space as shown in the following figure.



The result of clustering is shown in the following figure.



For any data point $\mathbf{x}_i$, let $d\,(\mathbf{x}_i\,, D_j\,)$ denote the average distance of $\mathbf{x}_i$ to the data points in cluster D_j , and let $j\,(i\,)$ denote the index of the cluster that $\mathbf{x}_i$ belongs to. Let $a(\mathbf{x}_i)=d(\mathbf{x}_i,D_{j(i)})$ be the average **Manhattan** distance of $\mathbf{x}_i$ to the points in its own cluster $D_{j(i)}$, and let $b(\mathbf{x}_i)=\min_{k\neq j(i)} d(\mathbf{x}_i,D_k)$ be the average **Manhattan** distance to the points in its neighboring cluster. Compute the Silhouettes scores for each medicine.

Medicine	Cluster	a($\mathbf{x}_i$)	b($\mathbf{x}_i$)	Max(a($\mathbf{x}_i$), b($\mathbf{x}_i$))	Silhouettes scores
A	1				
B	1				
C	2				
D	2				

Assume using **Manhattan** distance, compute the linkage functions for clusters 1 and 2. First, determine the centroid for clusters 1 and 2. Centroid of cluster 1 is (______ , ______) and centroid of cluster 2 is (______ , ______).Linkage functions [again assuming the *given distance metric* is Manhattan metric]

$L_{\text{single}}(1,2)$	$L_{\text{complete}}(1,2)$	$L_{\text{average}}(1,2)$	$L_{\text{cnetroid}}(1,2)$

7. Consider a set of 5 experiments and a gene g_1 that has an expression that can be represented as $g_1 = (1, 2, 3, 4, 5)$. Let us also consider the genes $g_2 = (100, 200, 300, 400, 500)$ and $g_3 = (5, 4, 3, 2, 1)$. Complete the following table. Based on the distance value, classify the genes accordingly.

Minkowski distance is defined by, $d_{Minkowski}(A,B) = [\sum_{i=1}^{n}(|a_i - b_i|^p)]^{1/p}$, where n denotes the number of experiments, and A and B denote the measurements. Note that the exponent is $1/p$.

	g_1 and g_2	g_1 and g_3	g_2 and g_3
Pearson Correlation coefficient			
Are the two genes belong to the same cluster? (Y or N)			
Dot product (cosθ)			
Same cluster? (Y or N) assuming distance within the same cluster is < 0.2			
Minkowski distance, p=0.5 Note that the exponent is 1/p.			
Same cluster ? (Y or N) assuming distance within the same cluster is < 50			
Canberra metric $Canb(X,Y) = \sum_{i=1}^{Q} \frac{\lvert x_i - y_i \rvert}{\lvert x_i \rvert + \lvert y_i \rvert}$ where *Q* denotes the number of experiments, and x and y denote the measurements.			
Same cluster ?(Y or N) assuming distance within the cluster is < 1.5			
Mahattan distance			
Same cluster? (Y or N) assuming data within the cluster is < 15			

Does the clustering results derive from the **Pearson** correlation coefficient calculation consistent with the **Dot product** results? Yes, they are consistent. No, they do not consistent.