Cairo University
Institute of Statistical Studies and Researches

Final Exam

Department of computer Science
Course Name: Intelligent Systems
Course Code: CS602
Instructor: Dr.Ammar Mohammed

Data: 31 December 2017
Duration: 3 Hours
Total Mark: 75

Note: Calculator is allowed.

Answer the following 5 Questions

Question: I (15 marks)

Choose the best answer (1.5 point for each)

1- What is true about K-Mean Clustering?

1. K-means is extremely sensitive to cluster center initialization
2. Bad initialization can lead to Poor convergence speed
3. Bad initialization can lead to bad overall clustering

The answer is:

A) 1 and 3 B) 1 and 2 C) 2 and 3 D)1, 2 and 3

2- For two runs of K-Mean clustering, is it expected to get same clustering results?

A. Yes B. No

3- The effectiveness of an SVM depends upon:

A) Selection of Kernel B) Kernel Parameters
C) Soft Margin Parameter C D) All of the previous

4- If you are using all features of a dataset and you achieved 100% accuracy on the training set, but 70% on validation set, what would expect from the model?

A) Underfitting B) Nothing, the model is perfect C) Overfitting

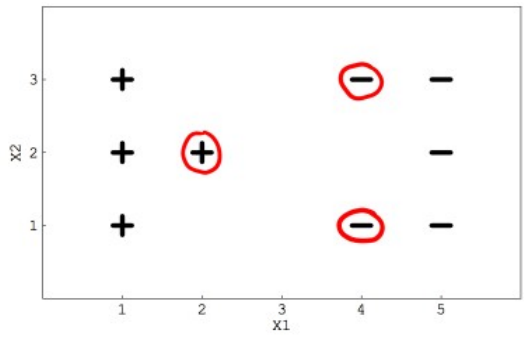
5- What is/are true about kernel in SVM?

1. Kernel function map low dimensional data to high dimensional space
2. It's a similarity function

The answer is:

A) 1 B) 2 C) 1 and 2 D) None of these

for the question 6 and 7:
Suppose you are using a Linear SVM classifier with 2 class classification problem. Now you have been given the following data in which some points are circled red that are representing support vectors.



6- If you remove the following red points from the data. Does the decision boundary will change?

A) Yes B) No

7- If you remove the non-red circled points from the data, the decision boundary will change?

A) True B) False

8- 5. What is the minimum number of variables/ features required to perform clustering?

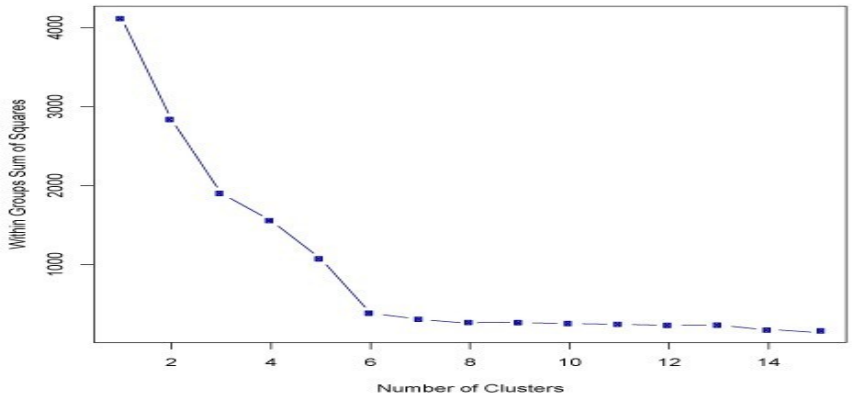
A) 0 B) 1 C) 2 D) more then 2

9- Which of the following method is used for finding optimal of cluster in K-Mean algorithm?

A) Elbow method B) Manhattan method

C) Euclidean method D) All of the above E) None of these

10- if you plot the run of number of clusters against the square error cost function, What should be the best choice for number of clusters based on the following results:



A. 5 B. 6 C. 14 D. Greater than 14

Question: II (15 marks)

1- Briefly describe, what does it mean n-cross-validation ? ***(3 marks)***

2- The following confusion matrix shows the results for classifying participants based on the gender into male and female. Answer the following:

	Actually male	Actually female
Predicted male	57	4
predicted female	6	32

A- How many mistakes are made when participants are predicated to be female ? **(2 marks)**

B- Calculate the Accuracy of the model ***(2 marks)***

C- Calculate the precision of the model***(2 marks)***

3- Briefly describe the main idea of SVM ? **(3 marks)**

4- given the following classification data

instance	**X**	**Y**	**class**
1	7	7	bad
2	7	4	bad
3	3	4	good
4	1	4	good

Use knn with k=3 and euclidean distance to classify the following instance (X=3 and Y=7) using euclidean distance. ***(3 marks)***

Question: III (15 marks)

1- Three binary nodes, N1 , N2 , and N3 , split examples into (0, 6), (1,5), and (3,3),respectively. For each node, calculate its entropy. ***(4 marks)***

2- Consider the training data (shown in the following Table) for a binary classification problem.

a1	**a2**	**Target class**
T	T	P
T	T	P
T	F	N
F	F	P
F	T	N
F	T	N

F	F	N
T	F	P
F	T	N

A- What is the entropy of this collection of training examples with respect to the positive class P? **(3 *marks)***

B- What are the information gains of ***a1*** and ***a2*** relative to these training examples? ***(6 marks)***

C- Which feature do you use as the starting (root) Node when applying a decision tree? ***(2 marks)***

Question: IV (15 marks)

1- What is the assumption of Naive Bayes Classifier? What is the problem that might appear when applying Naive Bayes method on certain data? Briefly describe how to solve this problem. ***(5marks)***

2- How does Naive Bayes classifier handle the conditional probability of a continuous feature? **(2 marks)**

3- Consider the following data set with features W, X, Y and a Boolean classification C

W	**X**	**Y**	**C**
T	T	T	T
T	F	T	F
T	F	F	F
F	T	T	F
F	F	F	T

And now you encounter a new tuple having the attributes W=F, X=T, Y=F. How should this example be classified using naive Bayes classifier ? Show your computation ***(8 marks)***

Question: V (15 marks)

1- Briefly describe the main idea of the K-Nearest Neighbor (KNN) Algorithm? ***(3 marks)***

2- What are the weakness of k-means **(3 *marks)***

4. Suppose you want to cluster the eight points shown below using k-means

	$Attr_1$	**$Attr_2$**
x1	1.0	1.0
x2	1.5	2.0
x3	3.0	4.0

x4	5.0	7.0
x5	3.5	5.0
x6	4.5	5.0
x7	3.5	4.5

Assume that k = 2 and the initial cluster means is chosen the individual x1, x4. Apply the k-means algorithm until convergence (i.e., until the clusters do not change), using the euclidean distance. ***(9 marks)***