School of Computing and Information Systems (CIS)
The University of Melbourne
COMP90073
Security Analytics
Tutorial exercises: Week 6

1. How the following measures guides us in anomaly detection problems? Give a scenario where each can be used.

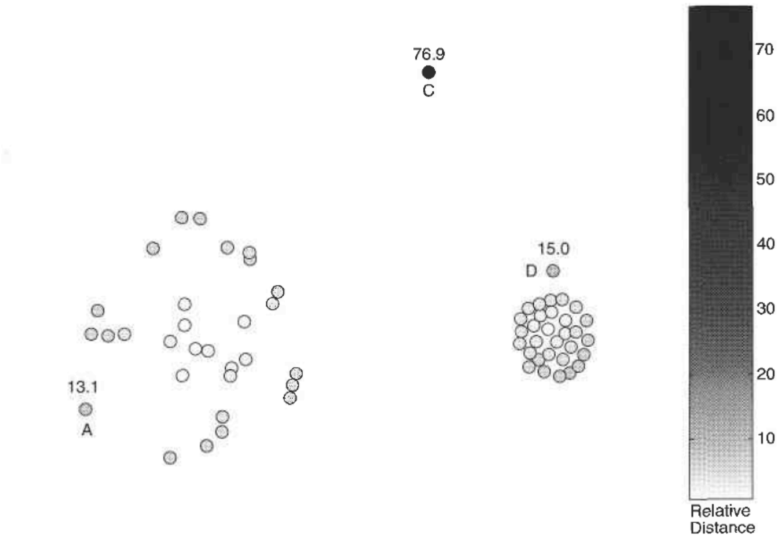
   a. Precision
   b. Recall
   c. F-score
   d. AUC

2. Following are the results observed for clustering 6000 data points into 3 clusters: A, B and C:

		Actual			
		A	B	C	SUM
Predicted	A	600	400	200	1200
	B	1000	1200	200	2400
	C	400	400	1600	2400
	SUM	2000	2000	2000	

What is the F_1-Score with respect to cluster B?

3. Consider the K-means scheme for outlier detection described in and the below figure.



a. The points at the bottom of the compact cluster shown in the above figure have a somewhat higher outlier score than those points at the top of the compact cluster. Why?

b. Suppose that we choose the number of clusters to be much larger, e.g., 10. Would the proposed technique still be effective in finding the most extreme outlier at the top of the figure? Why or why not?

c. The use of relative distance adjusts for differences in density. Give an example of where such an approach might lead to the wrong conclusion.

4. If the probability that a normal object is classified as an anomaly is 0.01 and the probability that an anomalous object is classified as anomalous is 0.99, then what is the false alarm rate and detection rate if 99% of the objects are normal? (Use the definitions given below.)

- Detection rate = number of anomalies detected/total number of anomalies
- False alarm rate = number of false anomalies/number of objects classified as anomalies

5. When a comprehensive training set is available, a supervised anomaly detection technique can typically outperform an unsupervised anomaly technique when performance is evaluated using measures such as the detection and false alarm rate. However, in some cases, such as fraud detection, new types of anomalies are always developing. Performance can be evaluated according to the detection and false alarm rates, because it is usually possible to determine, upon investigation, whether an object (transaction) is anomalous. Discuss the relative merits of supervised and unsupervised anomaly detection under such conditions.

6. Distinguish between noise and outliers. Be sure to consider the following questions.

a. Is noise ever interesting or desirable? Anomalies?

b. Can noise objects be outliers?

c. Are noise objects always outliers?

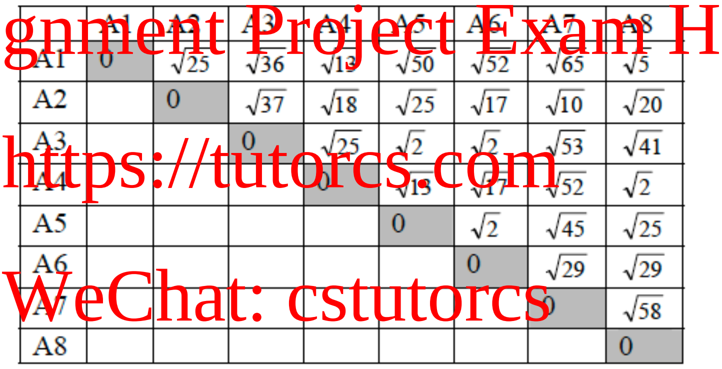
d. Are outliers always noise objects?

e. Can noise make a typical value into an unusual one, or vice versa?

7. Assume you run DBSCAN with MinPoints=6 and epsilon=0.1 for a dataset and obtain 4 clusters and 5% of the objects in the dataset are classified as outliers. Now you run DBSCAN with MinPoints=8 and epsilon=0.1. How do you expect the clustering results to change?

8. If Epsilon is 2 and minpoint is 2, what are the clusters that DBScan would discover with the following 8 examples: 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 matrix based on the Euclidean distance is given below:



	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Draw the 10 by 10 space and illustrate the discovered clusters. What if Epsilon is increased to 10?

9. You may use Python or Weka for the following exercises
Download the Ionosphere data set from the UCI Machine Learning Repository
https://archive.ics.uci.edu/ml/datasets/ionosphere
    a. Use of the LOF method and determine the ranking of the anomalies
    b. Rank the data points based on their k-nearest neighbour scores, for values of k ranging from 1 through 5.
    c. Normalize the data, so that the variance along each dimension is 1. Rank the data points based on their k-nearest neighbour scores, for values of k ranging from 1 through 5.
    d. How many data points are common among the top 5 ranked anomalies using different methods?

10. Repeat the above exercise with the network intrusion data set from the UCI Machine Learning Repository
https://archive.ics.uci.edu/ml/datasets/kdd+cup+1999+data