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?

Solution:

True Positive, TP = 1200

True Negative, TN = 600 + 1600 = 2200

False Positive, FP = 1000 + 200 = 1200

False Negative, FN = 400 + 400 = 800

Therefore,

Precision = TP / (TP + FP) = 0.5

Recall = TP / (TP + FN) = 0.6

Hence,

F_1 = 2 * (Precision * Recall)/ (Precision + recall) = 0.54 ~ 0.5

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?

Solution: The mean of the points is pulled somewhat upward from the centre of the compact cluster by point D.

(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?

Solution: No. This point would become a cluster by itself.

(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.

Solution: If absolute distances are important. For example, consider heart rate monitors for patients. If the heart rate goes above or below a specified range of values, then this has an physical meaning. It would be incorrect not to identify any patient outside that range as abnormal, even though there may be a group of patients that are relatively similar to each other and all have abnormal heart rates.

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

    Solution: The detection rate is simply 99%
    The false alarm rate = 0.99m*0.01/(0.99m*0.01+0.01*0.99)=50%

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.

    Solution: When new anomalies are to be detected, an unsupervised anomaly detection scheme must be used. However, supervised anomaly detection techniques are still important for detecting known types of anomalies. Thus, both supervised and unsupervised anomaly detection methods should be used. A good example of such a situation is network intrusion detection. Profiles or signatures can be created for well-known types of intrusions, but cannot detect new types of intrusions

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

    (a) Is noise ever interesting or desirable? Anomalies?

    Solution: No, by definition. Yes.

    (b) Can noise objects be outliers?

    Solution: Yes. Random distortion of the data is often responsible for outliers.

    (c) Are noise objects always outliers?

    Solution: No. Random distortion can result in an object or value much like a normal one.

    (d) Are outliers always noise objects?

Solution: No. Often outliers merely represent a class of objects that are different from normal objects.

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

Solution: Yes.

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?

Solution:
The graph whose nodes contain core and border points and whose edges connect directly-density-connected points will have less nodes and edges; as a result of that:
1. There will be more outliers
2. Some clusters that exist when using the first parameter setting will be deleted or split into several smaller sub-clusters

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

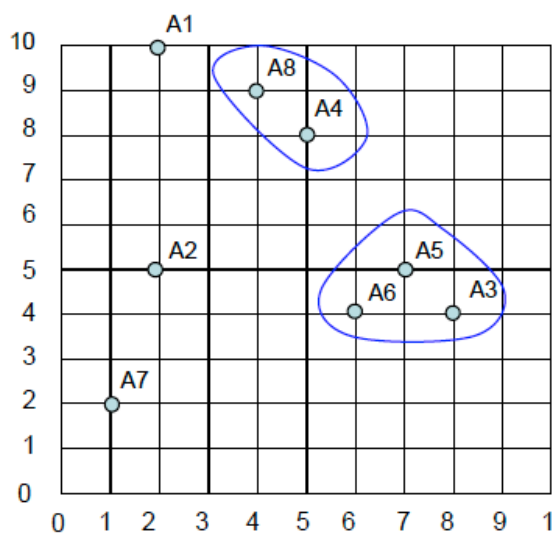
Draw the 10 by 10 space and illustrate the discovered clusters. What if Epsilon is increased to 10?

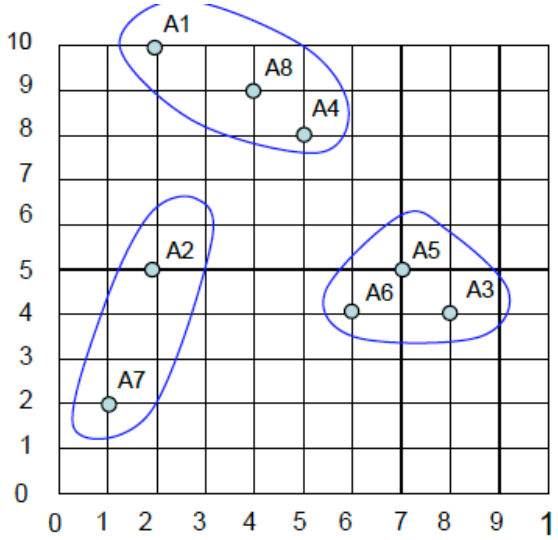
Solution:
What is the Epsilon neighbourhood of each point?
N2(A1)={}; N2(A2)={}; N2(A3)={A5, A6}; N2(A4)={A8}; N2(A5)={A3, A6}; N2(A6)={A3, A5}; N2(A7)={}; N2(A8)={A4}

So A1, A2, and A7 are anomalies, while we have two clusters C1={A4, A8} and C2={A3, A5, A6}

If Epsilon is 10 then the neighbourhood of some points will increase:

A1 would join the cluster C1 and A2 would joint with A7 to form cluster C3={A2, A7}.





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