Started on	Friday, 4 May 2018, 1:13 PM
State	Finished
Completed on	Friday, 4 May 2018, 1:21 PM
Time taken	8 mins 13 secs
Grade	**10.0** out of 10.0 (**100**%)
Feedback	Well done!

Question 1

Correct

Mark 1.0 out of 1.0

What is an outlier? Choose the best definition.

Select one:

- a. A technique used for finding fraud.
- b. Something new that has not been seen in the data before.
- c. A data point that is very distant from other data points.
- d. A data point that significantly deviates from the other data points in a set, to the extent that it would have been generated by some abnormal phenomenon. ✓
  Yes, this is what we are looking for. Unfortunately it does not necessarily mean that is what we can find.
- e. A subset of data objects that deviate from the rest of the data.

Your answer is correct.

The correct answer is: A data point that significantly deviates from the other data points in a set, to the extent that it would have been generated by some abnormal phenomenon.

Correct

Marks for this submission: 1.0/1.0.

Question 2

Correct

Mark 4.0 out of 4.0

There are three kinds of outliers that require targeted techniques to search for each.

Contextual ✓ outliers can be transformed into a conventional approach by

partitioning some attributes for grouping first and using the others for prediction with ✓ . This is used, for example, to find outlying people that do not conform to usual patterns of behaviour within their demographic group.

Collective ✓ outliers can be transformed into a conventional approach by

identifying structural units of related objects and mapping these to single objects re ✓ . This is used, for example, for

mining social networks that model relations between people in a graph ✓ .

For Global ✓ outliers the conventional approaches are used, including

building a one-class classifier and identifying outliers outside the class ✓ ,

fitting a probability distribution and identifying points of low probability ✓ ,

using the local outlier factor to detect objects that are even more distant than others ✓ and

using the nested loop method for identifying outliers that are most distant from their ✓ .

Instruction: For the last 4 blanks above you will need to choose the answers in alphabetical order to get the marks awarded for the right answer.

Your answer is correct.

The correct answer is:
There are three kinds of outliers that require targeted techniques to search for each.

[Contextual] outliers can be transformed into a conventional approach by [partitioning some attributes for grouping first and using the others for prediction within the group]. This is used, for example, to find outlying people that do not conform to usual patterns of behaviour within their demographic group.

[Collective] outliers can be transformed into a conventional approach by [identifying structural units of related objects and mapping these to single objects represented by conventional features]. This is used, for example, for [mining social networks that model relations between people in a graph].

For [Global] outliers the conventional approaches are used, including [building a one-class classifier and identifying outliers outside the class], [fitting a probability distribution and identifying points of low probability], [using the local outlier factor to detect objects that are even more distant than others in their local neighbourhood] and [using the nested loop method for identifying outliers that are most distant from their nearest neighbours].

Instruction: For the last 4 blanks above you will need to choose the answers in alphabetical order to get the marks awarded for the right answer.

Correct

Marks for this submission: 4.0/4.0.

Question 3

Correct

Mark 1.0 out of 1.0

Non-parametric statistical methods do not fit a statistical distribution to the data, but instead fit a user-controllable histogram or smoothing function and look for objects in regions with unusually low frequency.

Select one:

- True ✓
- False

The correct answer is 'True'.

Correct

Marks for this submission: 1.0/1.0.

Question 4

Correct

Mark 1.0 out of 1.0

A statistical multivariate outlier detection problem can be solved as a univariate problem by

Select one or more:

- a. Selecting the most important variable and using that for a univariate technique.
- b. Using the chi-square to compute the outlier-ness of a multidimensional object by aggregating the difference from the mean along each dimension at a time. ✓
- c. Doing nothing. There are no commonly accepted or proven ways of doing this.
- d. Computing a distance function along one dimension as an aggregate function of multidimensional distances using, for example, Mahalaobis distance. ✓

Your answer is correct.

The correct answers are: Using the chi-square to compute the outlier-ness of a multidimensional object by aggregating the difference from the mean along each dimension at a time., Computing a distance function along one dimension as an aggregate function of multidimensional distances using, for example, Mahalaobis distance.

Correct

Marks for this submission: 1.0/1.0.

Question 5

Correct

Mark 1.0 out of 1.0

The nested loop algorithm for distance-based outlier detection computes, for as many of every pair of objects in the data it needs, the distance between each pair. The distance is compared to a threshold value and an object is considered an outlier if the proportion of other objects exceeding the threshold distance away is relatively small.

Select one:

- True
- False ✓

This is almost right, but the directionality is wrong.

In "*considered an outlier if the proportion of other objects exceeding the threshold distan*ce " the word "*exceeding*" should be more like "*inside*" or "*at a distance less than or equal to*"

The correct answer is 'False'.

Correct

Marks for this submission: 1.0/1.0.

Question 6

Correct

Mark 1.0 out of 1.0

In proximity-based outlier detection, the local reachability detection is defined as

$$lrdk(o) = \frac{\|N_k(o)\|}{\sum_{o' \in N_k(o)} reachdist(o' \leftarrow o)}$$

Here, the *k-distance neighbourhood* of *o,* $N_k(o)$ is

Select one:

- a. The set of objects that are k-nearest neighbours of *o*. ✓
- b. The neighbours of *o* that are further away than the distance threshold.
- c. The set of local outliers with respect to *o.*
- d. The set of neighbours that are within $reachdist(o' \leftarrow o)$.

Your answer is correct.

The correct answer is: The set of objects that are k-nearest neighbours of *o*.

Correct

Marks for this submission: 1.0/1.0.

Question 7

Correct

Mark 1.0 out of 1.0

Match the clustering based method for outlier detection to the kinds of outliers it is aimed at detecting:

Outliers that are relatively distant from the clusters in which they occur ✓	Use k-means clustering and look for outliers unusually far from the cluster centroid
Outliers that do not belong to any cluster ✓	Use DBSCAN to build clusters and identify outliers
Outliers belong to a small cluster and are unlike others in large nearby clusters, or else they are not very similar to the others in their own cluster ✓	Use FindCBLOF

Your answer is correct.

Be sure you understand what clusteriing based methods are.

The correct answer is: Outliers that are relatively distant from the clusters in which they occur → Use k-means clustering and look for outliers unusually far from the cluster centroid, Outliers that do not belong to any cluster → Use DBSCAN to build clusters and identify outliers, Outliers belong to a small cluster and are unlike others in large nearby clusters, or else they are not very similar to the others in their own cluster → Use FindCBLOF

Correct

Marks for this submission: 1.0/1.0.