Anomaly Detection

Assignment 3 Answers

(10 points)

1. CONTINUE WORKING ON YOUR PROJECT IF YOU ARE DOING IT!!!

2. Generate 1000 data points randomly in 100-dimensional space, where each dimension is generated from the uniform distribution in (0,1). Perform PCA with this data set. What is the dimensionality of the subspace required to represent (i) 80% of the variance, (ii) 95% of the variance, and (iii) 99% of the variance. (6 marks)
(This is based on exercise 8 of chapter3 of the textbook)
Ans: You can find python codes for generating 1000 random numbers from a 100 dimensional Uniform (0, 1) space in attached ipython notebook or html file.
After performing principal component analysis on this dataset, the dimensionality of the subspace required to represent:

(i) 80% of the variance is 69,
(ii) 95% of the variance is 90,
(iii) 99% of the variance is 97.

3. Sketch how you would construct a Variogram Cloud to analyze temperature differences across a random sample of cities in the US. What distance measure would you use for the X-axis? How would you reduce the computational effort? (2 marks)
Ans: A variogram is a scatterplot between the spatial distances on the X-axis, and the behavioral square deviations on the Y-axis.
The spatial distances considered on X-axis are the Euclidean distances between the temperature differences
The behavioral attribute deviation is the half the square distance between the behavioral attribute values.

Step by step we can follow the procedure as below:

1. Get the data for temperature in US cities along with latitude and longitude.
2. For each pair of cities find the Euclidean distances using latitude and longitude values.
3. Variogram cloud is obtained by plotting all possible squared differences in temperatures for each pair of cities against spatial distances.

The high computational complexity is one of the challenges in creating a variogram cloud. It is not always necessary to represent each and every pair of points on the plot. Each spatial dimension can be discretized into ranges, creating a 2-dimensional grid in the data. Distance is calculated between the pairwise spatial distances between objects within the grid and their pairwise behavioral attribute values.

4. Describe how you might find spikes in a time-series (2 marks).
Ans: A basic thing one can do is to plot the data and look for peaks.

The simplest method to detect the spikes in time series is to compute a moving average of your input values. If your series is $x_1, x_2, \ldots$, then you would compute a moving average after each observation as:

$M_k=(1-\alpha)M_{k-1}+\alpha x_k$

where α would determine how much weight give the latest value of x_k.
If your new value has moved too far away from the moving average,

for example $(x_k-M_k)/M_k > 15\%$

then you flag that observation as an outlier/ spike in the series.

Alternatively the simplest procedure can be stated as below:

1. Calculate the average and standard deviation of the time series.
2. The values which are more than multiple of standard deviation (2*sigma, 3*sigma, etc. according to the requirement of data) can be flagged as outliers.