2. Outlier Detection

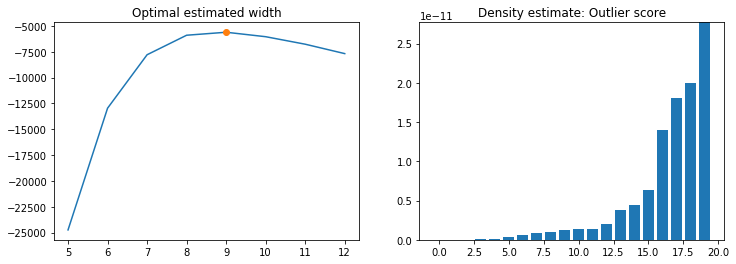
In this part, we used three different density estimators including Gaussian Kernel Density, KNN density and KNN average relative density to evaluate and rank the outlier score of our data set, and then we talk about whether there may be outliers in our data set based on three different scoring methods.

2.1 Gaussian Kernel Density Estimator

At first, we applied a gaussian kernel density Estimator on our data set which has 303 obersevations. A kernel density estimator is a deterministic approximation to the Gaussian Mixture Model (GMM) which tries to overcome some limitation of GMM, and in this report, we get the optimal kernel width by using leave-one-out cross validation which can calculate the maximum log of the likelihood for the whole data set. The following table is the values of the log of the likelihood in different :

|  |  |  |
| --- | --- | --- |
| Fold | Kernel Width | Log of the Likelihood |
| 1 | 0.000977 | -inf |
| 2 | 0.001953 | -inf |
| 3 | 0.003906 | -inf |
| 4 | 0.007813 | -inf |
| 5 | 0.015625 | -inf |
| 6 | 0.031250 | -24745.546978 |
| 7 | 0.062500 | -12963.483128 |
| 8 | 0.012500 | -7790.083359 |
| 9 | 0.250000 | -5901.747565 |
| 10 | 0.500000 | -5612.601211 |
| 11 | 1.000000 | -6044.790314 |
| 12 | 2.000000 | -6770.159939 |
| 13 | 4.000000 | -7681.069602 |

We selected 13 different , and we found that we got the highest log likelihood when kernel width , then we calculated the density using kernel width . The results of the outlier scores as follows:

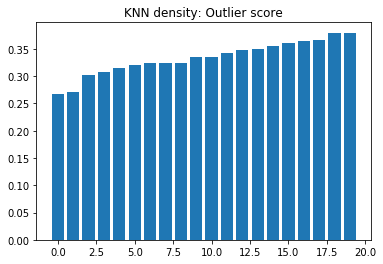


We show the 20 worst (lowest) scores in our data set in the right-hand pane of the Figure. It’s clear that there is a big difference between them, and the index of the first 5 observations are 291, 85, 223, 292 and 260.

2.2 K-Nearest Neighbor Density

Second, we used a KNN density estimator to measure the density of the data set. The KNN density tries to overcome the problem that KDE and GMM are unable to handle clusters of different densities well and it uses the K-nearest neighbor algorithm to calculate the average distance in Euclidian space. The core expression as follows:

In this report, we chose K=4 because our data set is relatively small. The results of the outlier scores are plotted below:



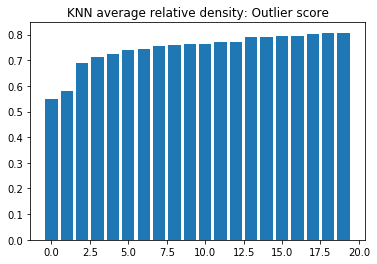
In this figure, we showed the 20 worst scores of all observations, but they don’t have an obvious difference base on this histogram. The indices of the worst 5 observations are 291, 85, 223, 204 and 292.

2.3 K-Nearest Neighbor Average Relative Density

Finally, we used KNN average relative density because we also wish to find some points where the density is lower than what it typically is for surrounding points. The KNN average relative density is one method which can be useful to define a notion of density that is relative to the neighborhood of the object. The expression as follows:

So in this way, we could get the results that the density of a given observation x relative to the average of its K nearest neighbor.

In this estimator, we used the same K as what we did in KNN density, and the results as follows:



This figure also plotted the 20 worst observations of the data set, the indices of the worst 5 observations are 48, 24, 195, 122 and 141, which are relatively different from the previous two estimators, but we also found that index of the 10th observation in this figure is 291.

2.4 Outlier Detection based on Scoring Methods

Based on previous experiments and results from three different scoring methods, we found that there are some observations which have relatively low scores based on density. There are 2 common definitions of the outlier:

1) *Hawkin’s definition*: An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism. 2) *Probabilistic definition*: An outlier is an object that has a low probability with respect to a probability distribution model of the data.

Based on the Probabilistic definition, we can say that there are some outliers in our data set because we used some scoring methods to estimate the density distribution of our data set, and the results showed there are some observations which low density, which means that these observations have a low probability compared with the probability distribution model of the whole data. However, when we checked the real data of these observations such as the 291st observation:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *55* | *0* | *0* | *128* | *205* | *0* | *2* | *130* | *1* | *2* | *1* | *0* |

We found that there was no outliers according to the Hawkin’s definition because the values of the all attributes are in a normal scope.

So, generally, there may be outliers based on the scoring methods (Probabilistic definition) and there is no outlier according to the real meaning and values of the observations and attributes (Hawkin’s definition).