02450 Project 2

Report

by Karol Dzitkowski
Marco Becattini

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Introduction to Machine Learning and Data Mining
Tue Herlau
25th November 2014

Abstract

The objective of this second report is to apply the methods you have learned in the second section of the course on „Supervised learning: Classification and regression" in order to solve both a relevant classification and regression problem for your data.

Classification

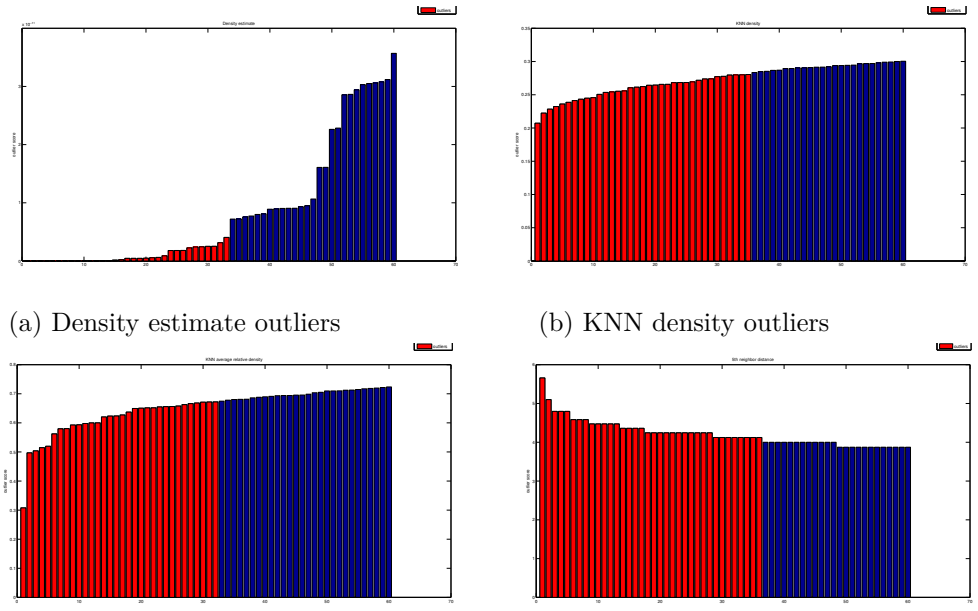
2.1 Problem

Outlier / Anomaly detection

In statistics, an outlier is an observation point that is distant from other observations. For this reason we apply outlier / anomaly detection for each class of our dataset in order to consider which data for each class could be excluded from the data set. Below is shown an example of outlier detection applied to the subset that correspond to the letter 'A'. We calculate for all the samples the leave-one-out Gaussian Kernel density, KNN density, KNN average relative density and distance to K-th nearest neighbor for with K=5.
For each of the metrics, we sort the outlier scores and inspect their distribution. We set an outlier threshold where there is a significant "jump" in the scores, in order to have around 35 outliers (about 5% of the total samples). We consider only the outliers which are detected by all four methods, and in this way we obtain 4 outliers (about 0.5% of the samples). In the bottom table are specified our results:

Method	**Selected outliers (indices)**
Gaussian Kernel density	[495 741 311 133 159 51 18 399 86 739 136 566 515 423 2 227 760 285 489 152 149 572 655 688 469 114 124 762 130 638 327 645 307]
KNN density	[18 311 495 741 133 159 515 86 399 51 739 136 489 566 285 116 688 760 227 152 423 500 469 114 655 2 149 657 504 696 238 604 572 439 124]
KNN average relative density	[197 156 102 215 356 31 399 217 127 573 701 617 124 355 530 86 36 304 450 184 133 452 321 183 746 728 714 724 638 627 285 149]
Distance to 5th nearest neighbor	[18 311 116 515 741 399 439 500 86 133 230 495 739 159 292 489 657 51 132 136 152 238 284 285 390 438 604 688 57 114 212 227 504 566 655 696]
Common to all method	[51 86 739 688]

In figure 3.1 we represent one bar plot for each metrics. Each column is a pattern and the y-axis is the outlier score density (or distance for for 5th nearest neighbor). Red columns are patterns with lower outlier score (or highest for 5th nearest neighbor) that we decided to set as otliers.



(a) Density estimate outliers

(b) KNN density outliers

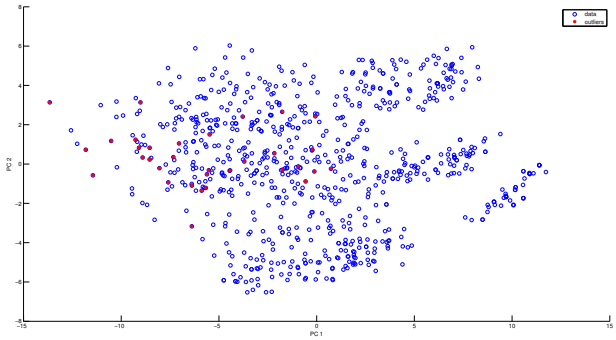
(c) KNN avg rel. density outliers

(d) 5th neighbor distance outliers

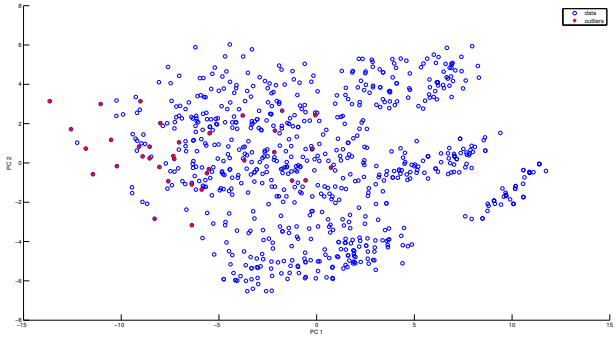
Figure 3.1: bar graph outliers

To further understand the outliers distribution, for each scoring method we plot the outliers in the PCA space. Figures 3.2 shows that is not really possible to see a pattern for outliers in the PC1/PC2 space (maybe it could be possible find only few of them). In figure 3.3 you can see the outliers detected by all four methods.

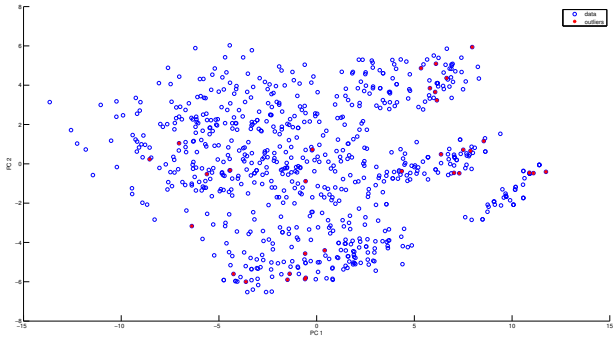
We applied outlier detection using different metrics, and by setting an appropriate threshold we identified some outliers, constituting about 0.5% of our dataset. Anyway we are never sure whether a certain pattern is an outlier but we can just suppose it especially because we have a wide dataset so is probable that some patterns are distant from other observations.



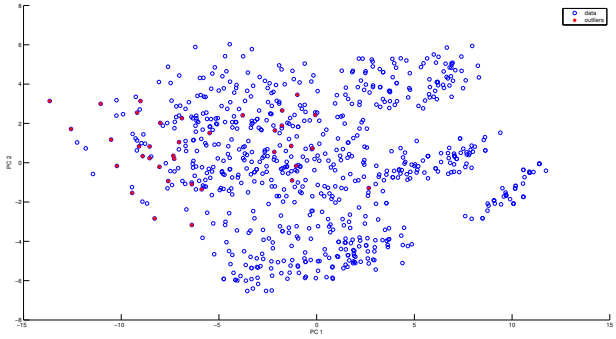
(a) Density estimate outliers



(b) KNN density outliers



(c) KNN average relative density outliers



(d) 5th neighbor distance outliers

Figure 3.2: Principal component plot with outliers

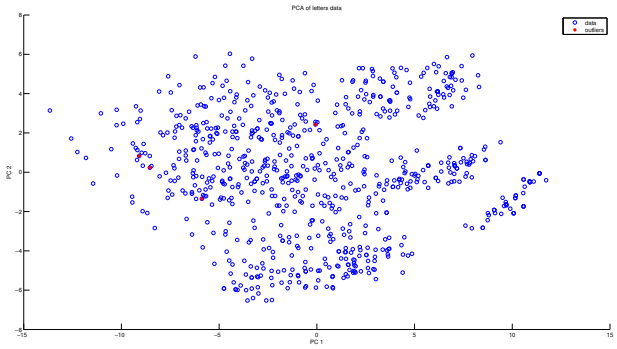


Figure 3.3: outliers detected by all four methods