Pattern Recognition

# Optimal Bayes Classification

**Wojciech Korzeniowski**

## Task 1

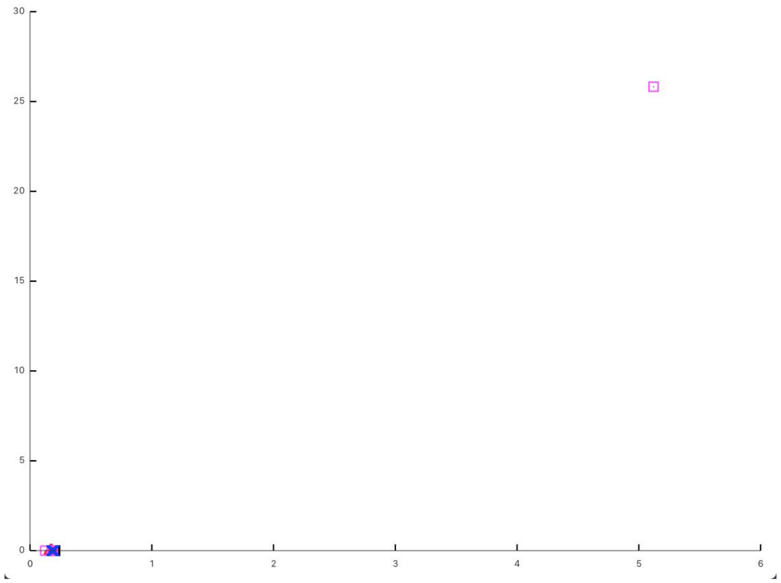
Original data have following mean and median values.

| | Feature 1 | Feature 2 | Feature 3 | Feature 4 | Feature 5 | Feature 6 | Feature 7 |
|---|---|---|---|---|---|---|---|
| mean | 1.8679e-01 | 1.4839e-02 | 2.1045e-01 | 2.0882e-01 | 7.9658e+01 | 1.0604e+00 | 9.0846e-03 |
| median | 1.8259e-01 | 1.4785e-04 | 1.7434e-04 | 1.9996e-06 | -8.9358e-11 | 1.3626e-10 | -1.8427e-14 |

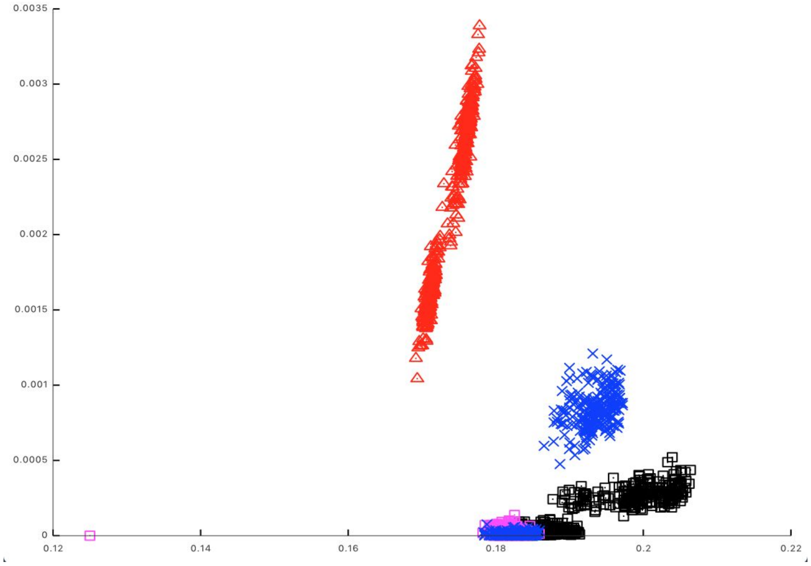
Mean and median values after outliers removal. Changes in mean values are huge, it would have huge impact on classification results.

| | Feature 1 | Feature 2 | Feature 3 | Feature 4 | Feature 5 | Feature 6 | Feature 7 |
|---|---|---|---|---|---|---|---|
| mean | 1.8412e-01 | 6.8258e-04 | 9.4867e-04 | 1.4729e-05 | -6.9796e-10 | 2.9713e-07 | -9.4668e-11 |
| median | 1.8259e-01 | 1.4785e-04 | 1.7434e-04 | 1.9996e-06 | -9.0827e-11 | 1.3626e-10 | -1.8742e-14 |

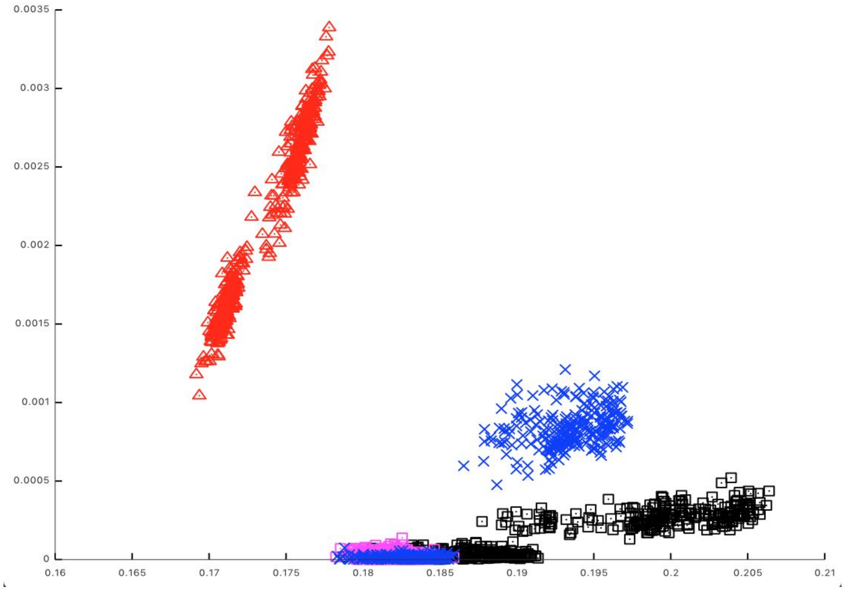
I used `plot2features` to check how distribution of two selected features (3 and 5) looks like.



I decided to delete sample with the highest values of features



Now I see that there is another outlier, this time with min value.



At last, train set without outliers.

## Task 2

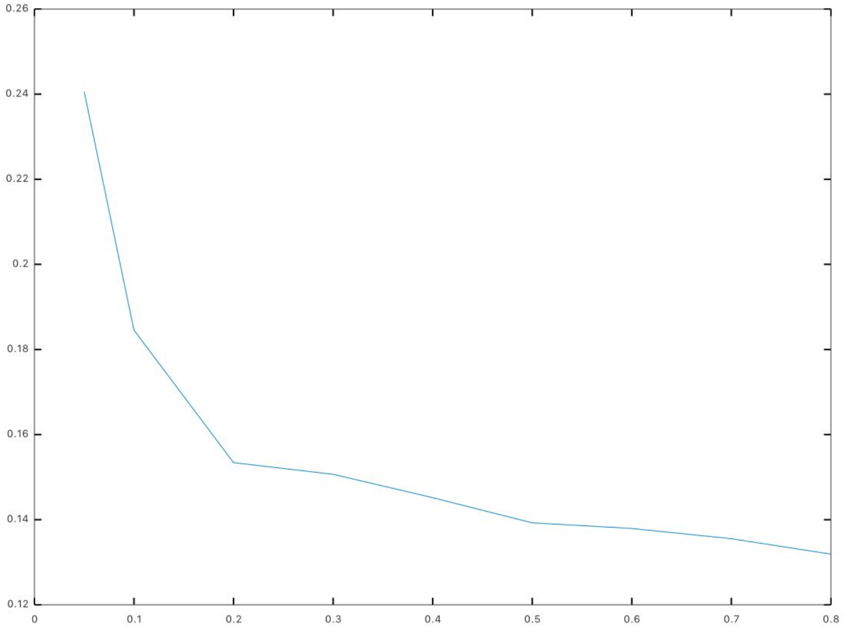
For features that I selected (2 and 3) I got following classification results.

| Pdf estimation | Independent features | multidimensional | Parzen window |
|---|---|---|---|
| Error | 33.114% | 25.932% | 12.719% |

## Task 3

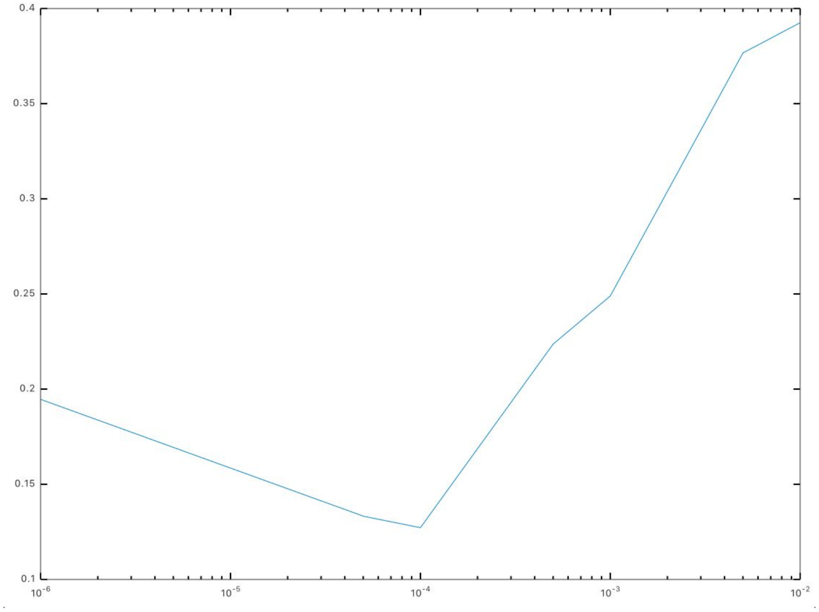
Number of samples in training sets influences the classification quality. Especially when Parzen Window is used to approximate pdf function. But as we can see, event if we use just 5% samples from 1822 examples of training set, which gives use about 91 samples in training set, so it means that we use 22 samples per class we  gets much better results compared to others method of pdf estimation.

| Part of set | 5% | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% |
|---|---|---|---|---|---|---|---|---|---|
| error | 24.06% | 18.45% | 15.34% | 15.07% | 14.52% | 13.93% | 13.79% | 13.55% | 13.19% |



## Task 4

Parzel windows need to be adjusted to given set, we see that too small and too big values gives us worst classification results. For this set best parzen windows size value is around $10^{-4}$ .



## Task 5

When we change *apriori* probability we should should Test Set.
Test Set should represents real data so proportions between number of samples of each class should be corresponding to given *apriori* probability.

Error without change in proportions: **13.32%**
Error after proportion adjustment: **10.52%**

## Task 6

I've checked how 1NN classification works with raw data and with normalised ones.

| Normalisation | none | $x' = \frac{x - \mathrm{mean}(x)}{\max(x) - \min(x)}$ | $x' = \frac{x - \bar{x}}{\sigma}$ |
|---|---|---|---|
| Error rate | 16.67% | 3.45% | 3.72% |

As we can see have that normalization of data have huge impact on classification results.