Datamining Exam

Christian Hohlmann Enevoldsen, MRB852

3. april 2014

1.2

The sample mean is comuted by taking the average of all y's in $SSFR\_Train\_Y$.
The result is: -10.911616939999998
bias is computed by the sum of every y, subtracted with the mean and power by two

The bias is: 0.8952439658310365

I followed the algorithm in the assignment to get this result.

1.3.1

To calculate the affine linear model, I used the algorithm 3 in the lecture notes

The 5 parameters is: $\begin{bmatrix} -0.79400153 \\ -1.2229592 \\ -0.32858475 \\ -0.78633056 \\ -8.14943349 \end{bmatrix}$

1.3.2

The mean square error of the training set is: 0.274754600685

The MSE is below the bias, which is good because it means that the model is more stable, as we are closer to the average for all entries in the data.

1.3.3

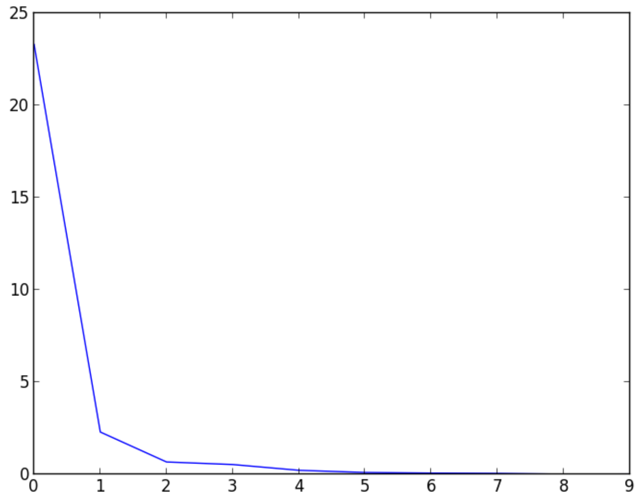
To compute the MSE I have used the algorithms described in the lecture notes. Point 3.4 on page 12. I have implemented it in Python, and used numpy as a helper tool.
The MSE for the TEST set is: 0.275179630629
The MSE for the TEST and TRAINING sets are very similar, which means that the model is good.

2.1

The accuracy is computed by using python, numpy and the algorithm 1 (1-NN) in lecture notes. the result is: 0.984666667 = 98.46666%
Note that the code is not optimized, so it will take a couple of minutes to compute.

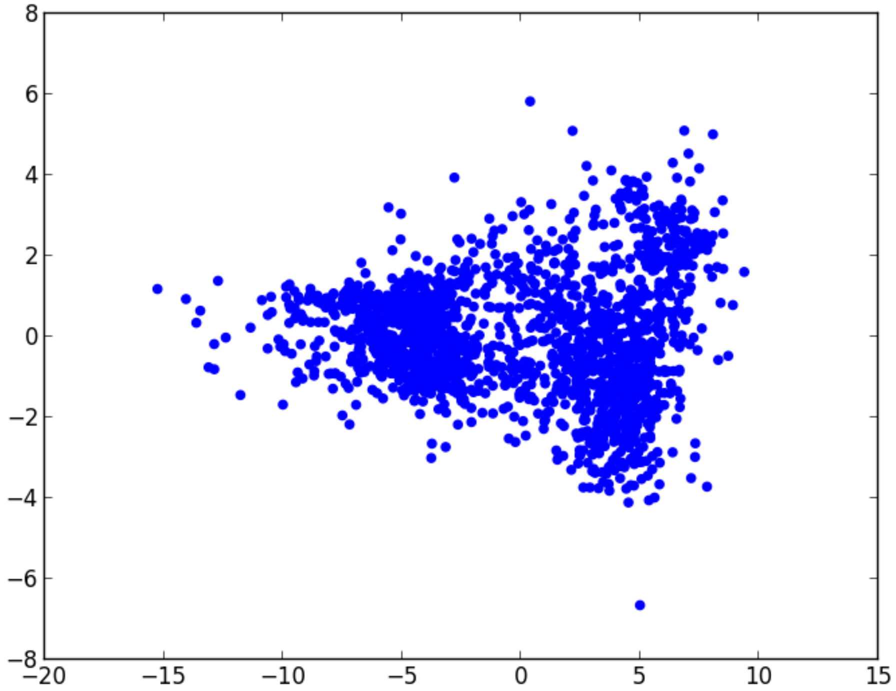
2.2

I have made use of the matplotlib.pyplot to plot and scatter the data.
The plot:



The scatter:

The scatter is the visual result of applying PCA with M = 2 to the dataset.



I calculated the the variance by taking the sum of the two first eigenvalues and divided it by the sum of all eigenvalues.
This tells that we need 2 components to explain 90 percent of the variances. It is also clear in the plot.
2.3

For every data point we are assigning them a centroid, and after every iteration we are trying to relocate them based on the mean of all the datapoints assigned to the centroid.
The code is not really working, because I stored everything in one matrix, and I have python problems.