Assignment2

[Ondertitel van document]

Meneer Doos

[Jaar]

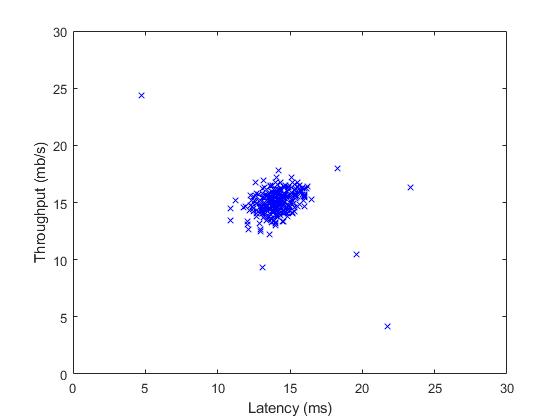
Assignment 2

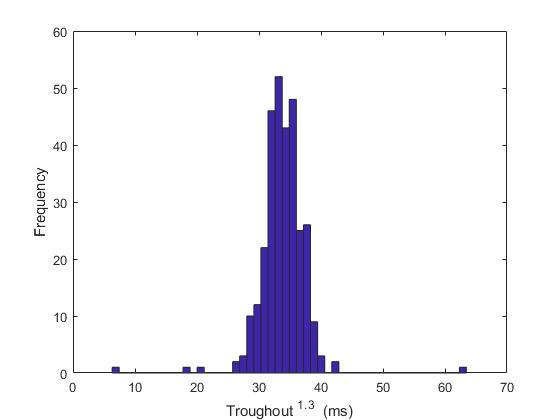
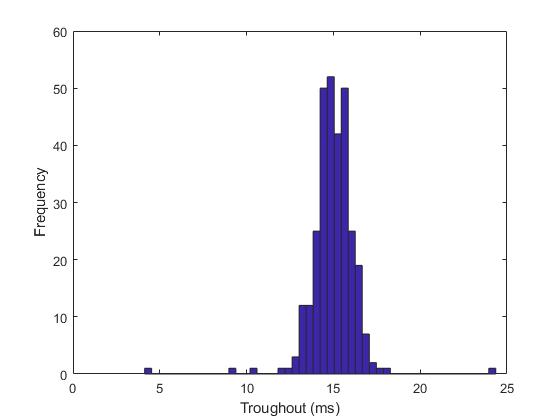
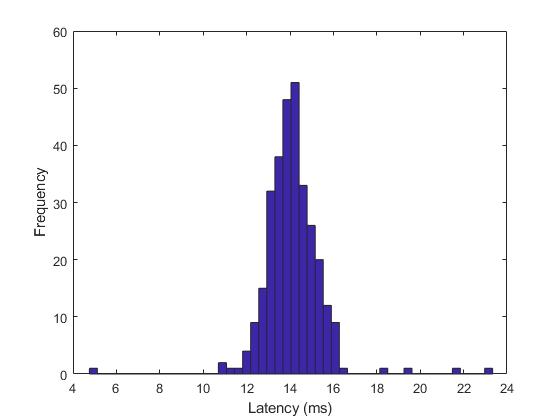
Example = een set (x1,x2)

# Anomaly detection

In the first part of the assignment, we will be implementing an anomaly detection algorithm. The algorithm will be used to detect anomalous behavior in server computers. We have an unlabeled dataset {x(1), x(2)}, n=307 .

## Part one

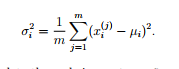




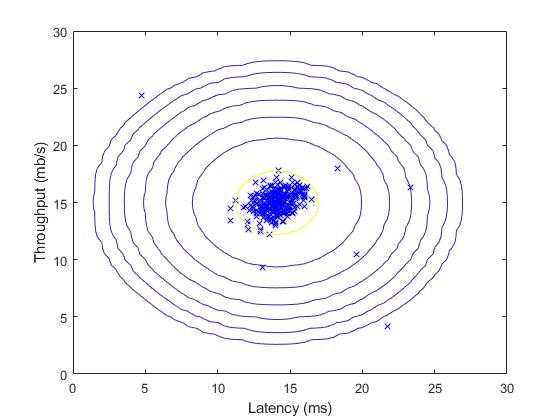
Plotting the dataset shows there are only a few outliers. We have a large amount of positive data, and a small amount of negative data. Using supervised learning would be hard, because the algorithm has a hard time learning from the examples. Instead we rather calculate the probability, where a low probability would be a negative example. Plotting a histogram shows the gaussian distribution of the dataset. Figure . and . show the non-manipulated data. By manipulating the data with ^(1.3) , the gaussian distribution can be slightly improved on … data2… . However, the non-manipulated data will be used in the next parts of the assignment.

## Part two

To perform the anomaly detection, u and o” are estimated for our dataset. To complete the code in estimateGaussian.m the formulas below are implemented with a for loop to go trough the dataset.

Next, the probability is calculated. This can be done with formula below. When working with multivariate Gaussian distribution, the probability for each feature is added resulting in the probability density. Using the density together with matlab contour function, the contours of the fitted Gaussian distribution can be plotted. MEER UITLEG 

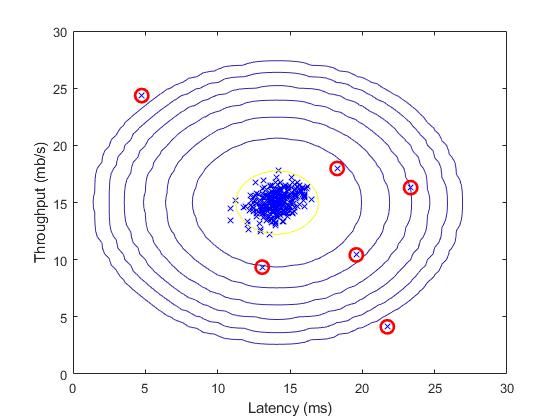


## Part three

We look for best threshold to differentiate the outliers, this means when the probability density is below the selected value of epsilon it is considered as an outlier.   
By looping trough the different values epsilon ( starting with the lowest value for the probability density till the highest value for the probability density, moving in steps of (range of the probability density/1000) ). For each of these values the F1 score is calculated, incase of a better F1 score, the epsilon will be set as best epsilon. This loop will eventually reach the best value for epsilon.

Iets van die speciale sum vermelden ?

As mentioned before, the probability density then is compared to this epsilon value. If the probability density is lower, the example is seen as an outlier. These outliers are circled in red.



## Part four

The code developed in previous parts of the assignment is now applied to a harder problem. In this dataset, the examples are described by 11 different features.

# Collaborative filtering: Recommender System