Anomaly detection algorithm

Objective: To build an algorithm to detect failing servers (anomalous behaviour) in a network of computers

Approach: Fit a Gaussian model to data, and detect data points with low probabilities as anomalies.

Methodology:

If the probability value of a data point is less than a threshold value, it is determined as an anomaly. The threshold is selected by iterating through a set of values and selecting the one that yields the best F1 score on the cross-validation set (labeled examples).

The input data constitutes two features: throughput (Mb/s) and latency of response(ms) for 307 unlabeled examples.

Calculate mean and variance of each feature. Estimate probability distribution of all data points in the training data set.

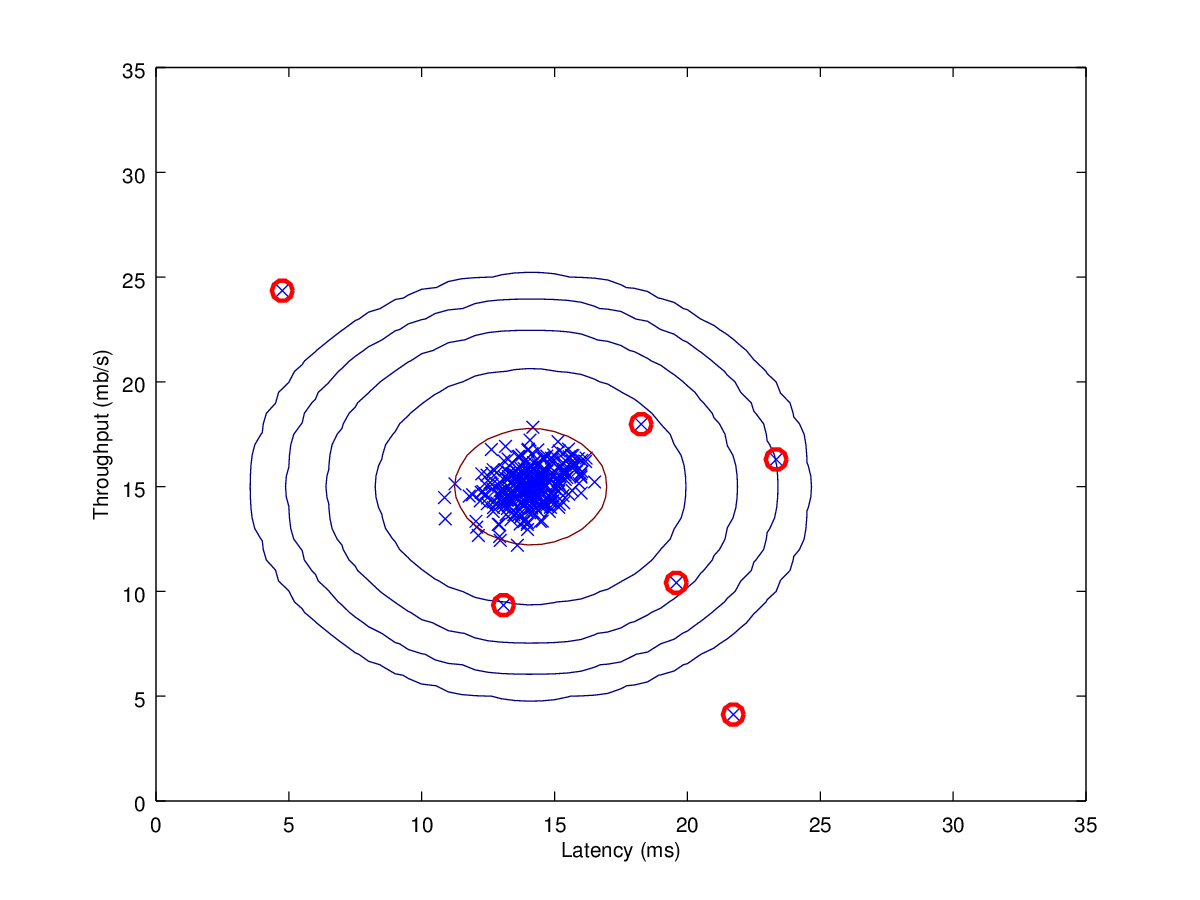
CALCULATE GAUSSIAN PARAMATERS

SELECT THRESHOLD

STUDYING INPUT DATA

Result:

A model that can detect anomalous behavior in a network of computers has been built. This model is not constrained by the dimension of the data. Hence the same model can be trained on datasets with additional features and can be used to detect anomalies in high dimensional dataset. The detected anomalies are marked in red as shown below.



Appendix:

This model was built in Octave 4.0.3. It can be run an MATLAB, as well.

Note: 30 percent of attached program files were provided by “Coursera”

Attached files:

ex8 – Anomaly detection model

estimateGaussian-Function to calculate mean, variance, and estimate Gaussian distribution

SelectThreshold-Function to select threshold value

Formulas: