Introduction to pip

> pip is a package management system used to install and manage software packages written in Python. Many open source packages can be downloaded using pip.

Usage:

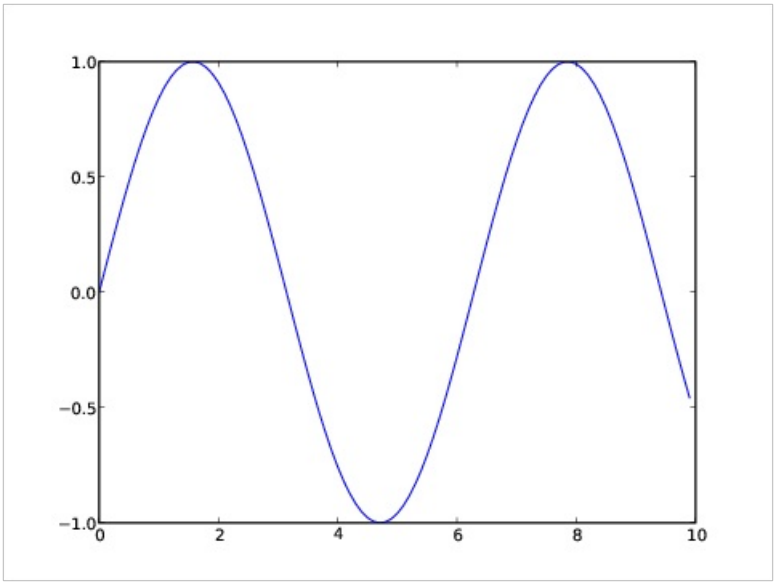
```
pip instsall <package name>
pip uninstall <package name>
pip install <package name>==<version number>
```

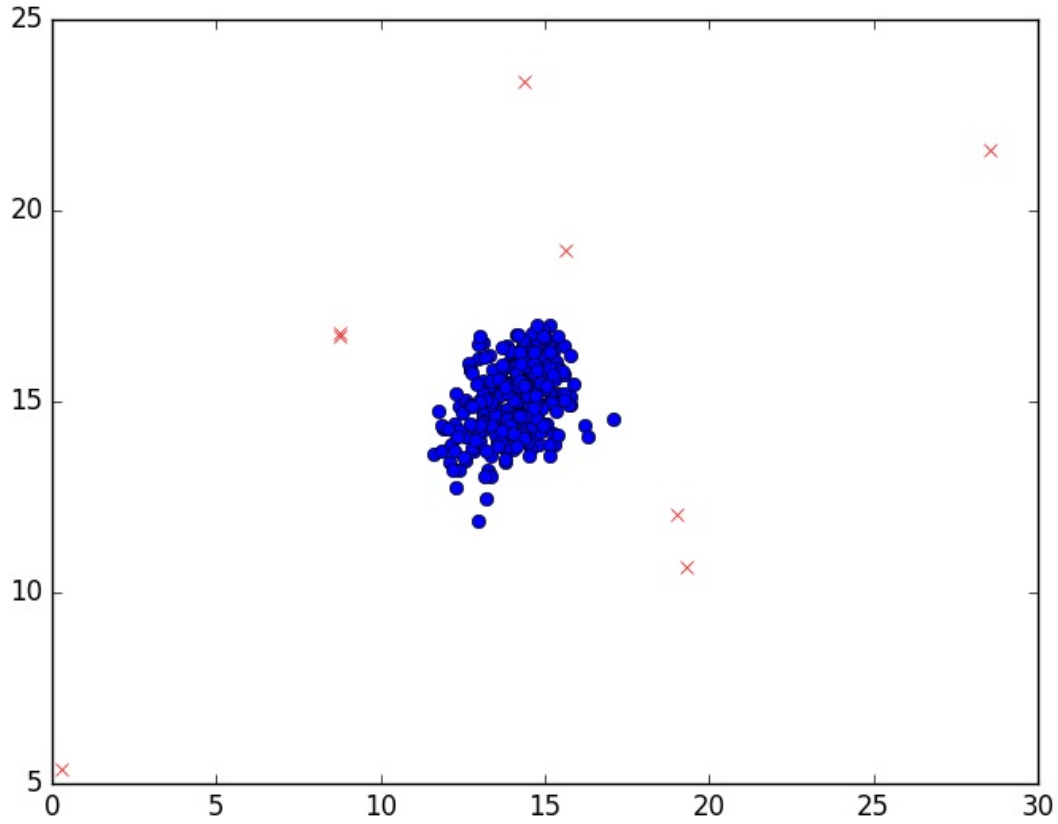
Introduction to numpy

> NumPy is the fundamental package for scientific computing with Python. It contains among other things:
>
> - a powerful N-dimensional array object
> - sophisticated (broadcasting) functions
> - tools for integrating C/C++ and Fortran code
> - useful linear algebra, Fourier transform, and random number capabilities

Introduction to matplotlib

> matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.





Install numpy and matplotlib

Type the following command in your command prompt/terminal to instsall these packages:

```
pip install numpy
pip install matplotlib
```

Follow the instructions at http://scipy.org/install.html and http://matplotlib.org/users/installing.html if you are not able to install them using the command above.

Gaussian distribution for anomaly detection

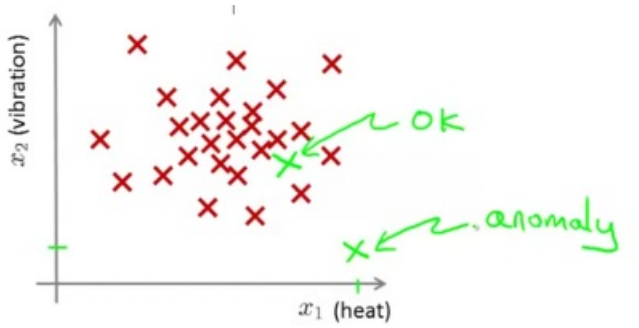
Reference: **Machine learning course on coursera by Andrew Ng**

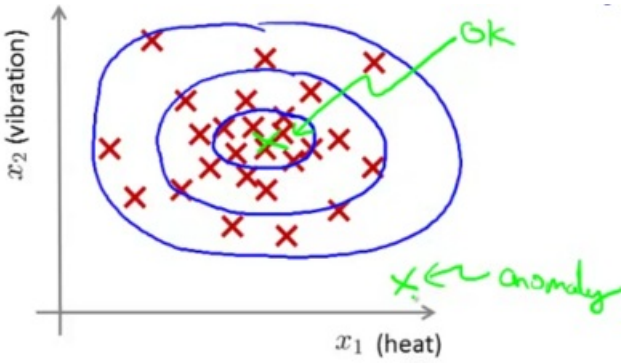
Example of anomaly detection: Anomaly detection in aircraft engines

Aircraft engine features are :

i. x1 = heat generated
ii. x2 = vibration intensity

Question: Is the new engine anomalous or should it receive further testing as the two possibilities in the following graph.





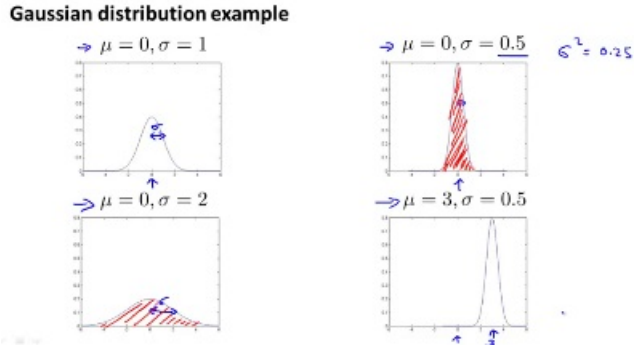
Anomaly detection can also be used to monitor computers in a data center. eg.

Features:
1. x1 = memory use
2. x2 = number of disk accesses / sec
3. x3 = CPU load
4. x4 = CPU load / network traffic etc.

Identify machines that are likely to fail and flag off for attention.

Gaussian Distribution

$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} exp\left(-\frac{(x-\sigma)^2}{2\sigma^2}\right)$$

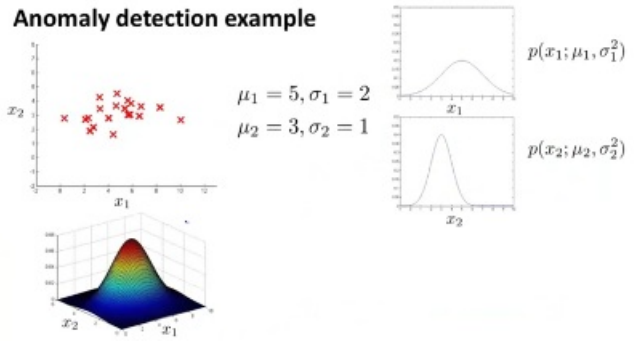


Standard Deviation: how far (on average) the data points are from the mean value of the data set

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Density estimation

- Lets say we have a training set of m examples, each example has n features.
- Assume that each feature is distributed as per gaussian probability distribution.
- The computed probability is thus:
  $p(x) = p(x_1; \mu_1, \sigma_1^2) * p(x_2; \mu_2, \sigma_2^2) * ... * p(x_n; \mu_n, \sigma_n^2)$



Classify a point as anomaly if the probability is lower than a selected threshold.

Exercise: anomaly detection

In this exercise, you will implement an anomaly detection algorithm to detect anomalous behavior in server computers. The features measure the throughput (mb/s) and latency (ms) of response of each server. While your servers were operating, you collected m = 307 examples of how they were behaving. You suspect that the vast majority of these examples are “normal” (non-anomalous) examples of the servers operating normally, but there might also be some examples of servers acting anomalously within this dataset.

Your tasks:

i. calculate the mean and std for the two columns (throughput (mb/s) and latency (ms))
ii. complete the predict method: for each row, change the last element to 1 (the last column indicates whether the points are anomalies or not, 1 means anomaly) if the gaussian probability is smaller than the threshold
iii. plot the detected anomalies and the rest points on the same graph using different colours
iv. adjust the threshold to separate the outliers and the rest