Week 1

Clustering

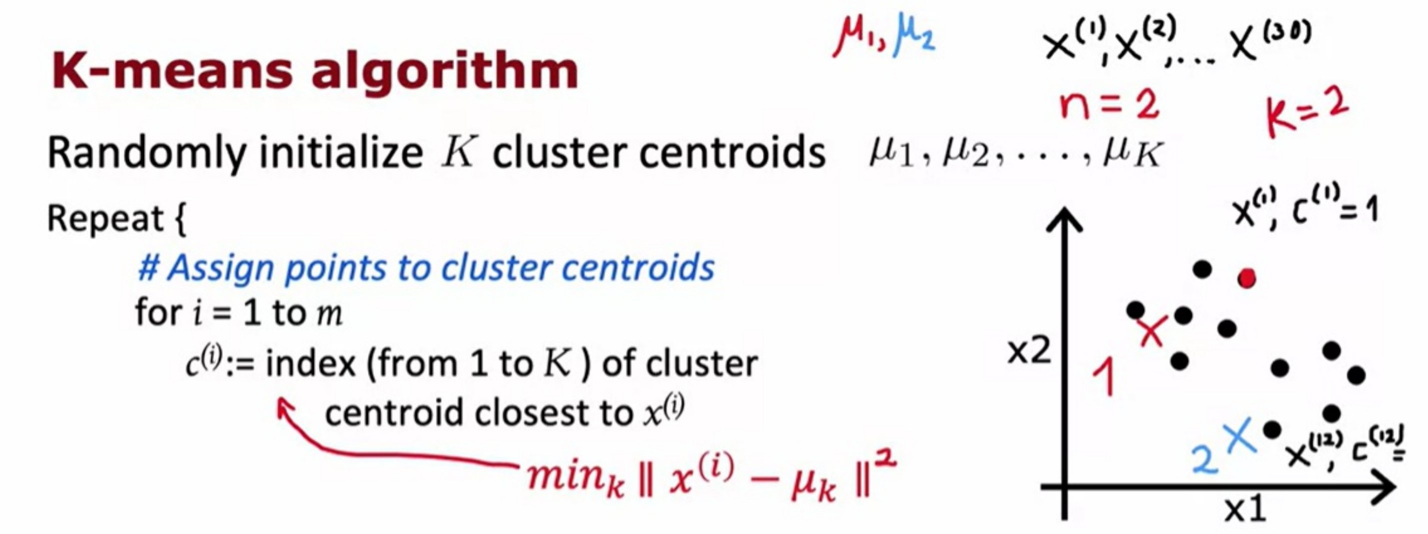
K-means Intuition

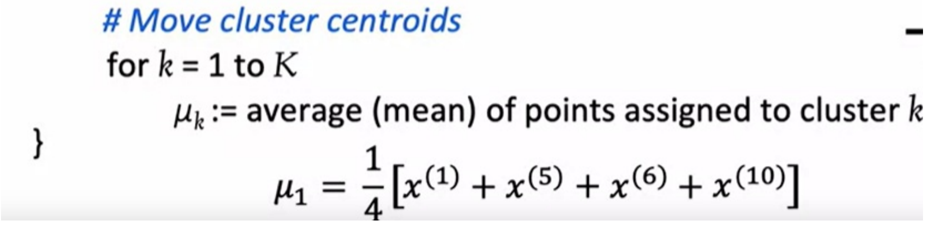
Randomly initialize centroids

1. Assign each points to its closest centroid
2. Recompute the avg location of the points and move the centroid
3. Repeat from 1.

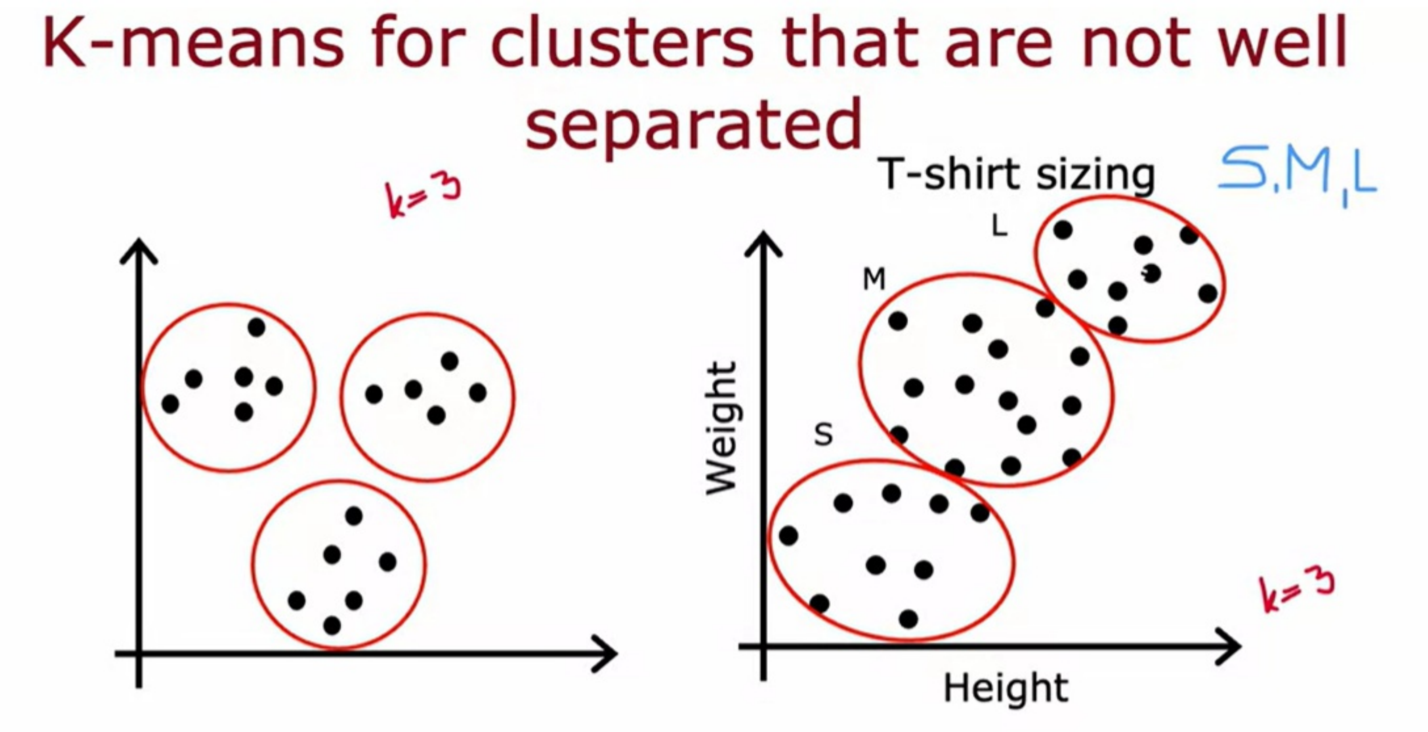
K-means algorithm

- Edge case: What if a cluster has no points?
  → Eliminate the cluster (more common)
  → Reinitialize the cluster centroid

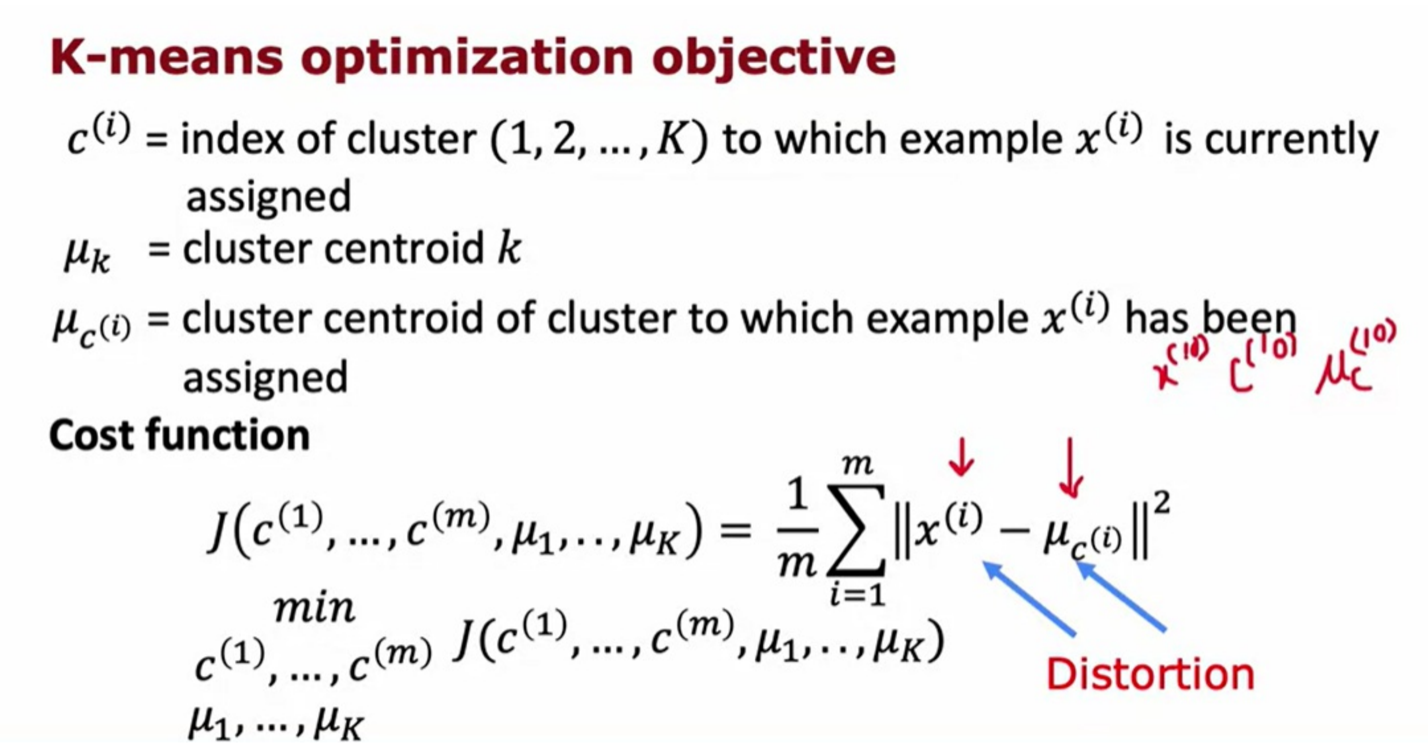




- K-means for clusters that are not well separated:



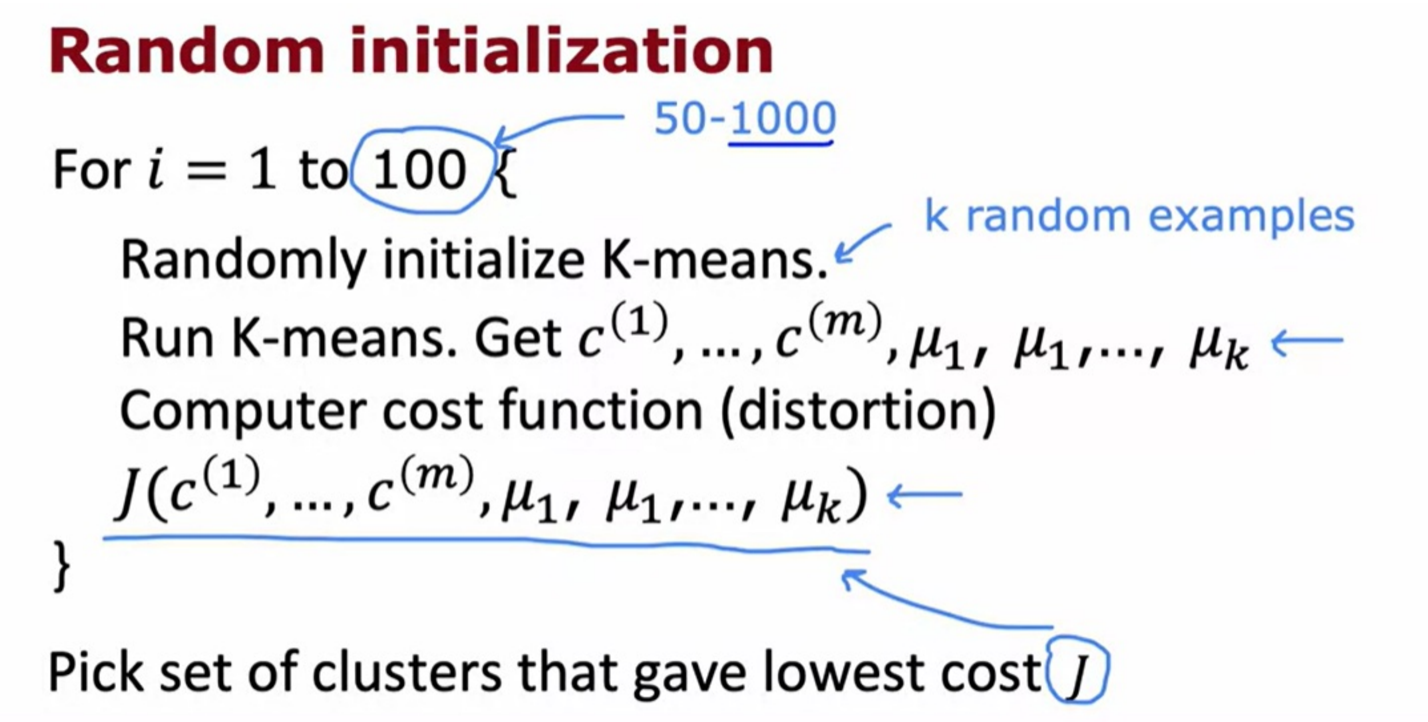
K-means Formula



- K-means will converge → if cost increase, there is something wrong
- If it remains the same, can stop running
- If it becomes slow after many iterations, you can stop as well

Initializing K-means

- Run K-means multiple times to find the best initialization
- Find one with the smallest J (Cost)



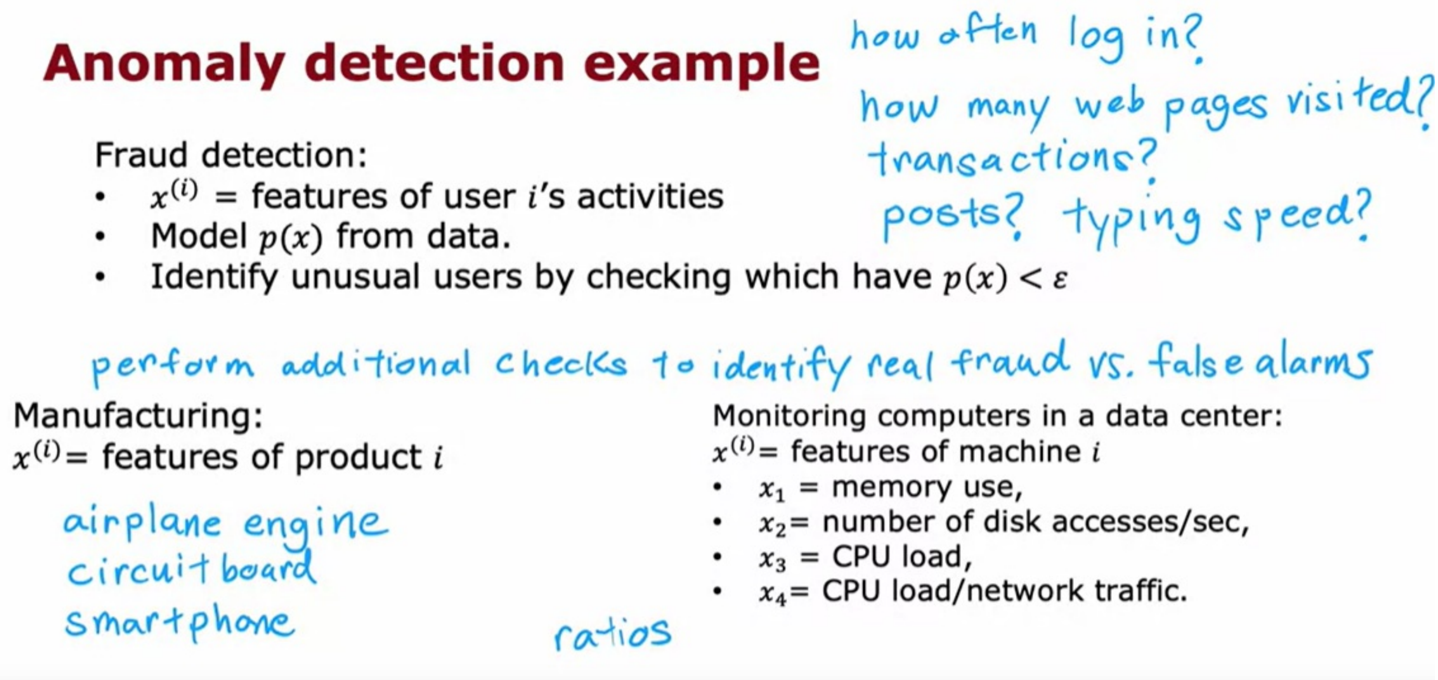
Choosing the number of clusters (k)

- Elbow method
  - Idea: Look at the point where the **rate** of decrease on cost is the fastest
  - Don't choose k just to minimize the cost function!!!

Anomaly Detection

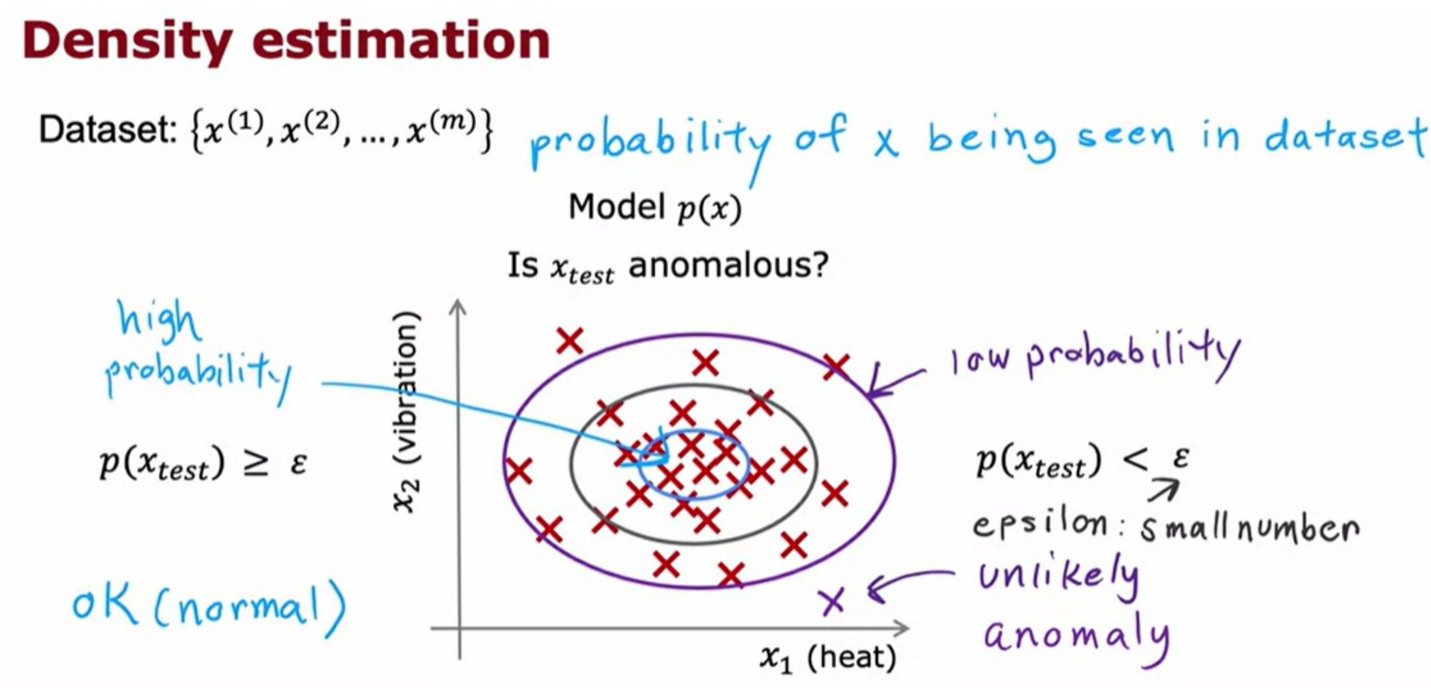
- Problem e.g: Installing airplane engine.
  → Could have an anomaly with high temperature and low vibration of engine
  → Needs to be addressed

Anomaly Detection Use Cases



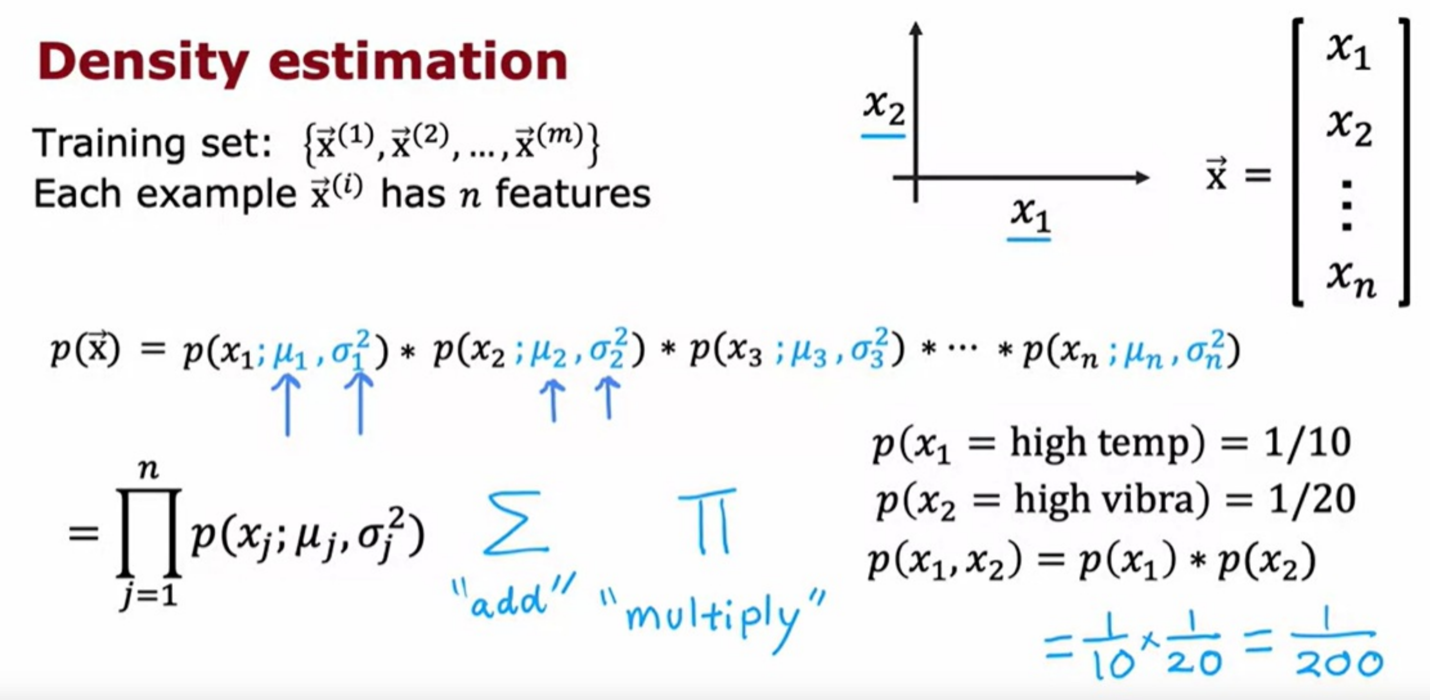
Density Estimation

- Probability of x being seen in the dataset
- Region with high probability and low probability
- If p(x_test) < **ε, data will be flagged as an anomaly**

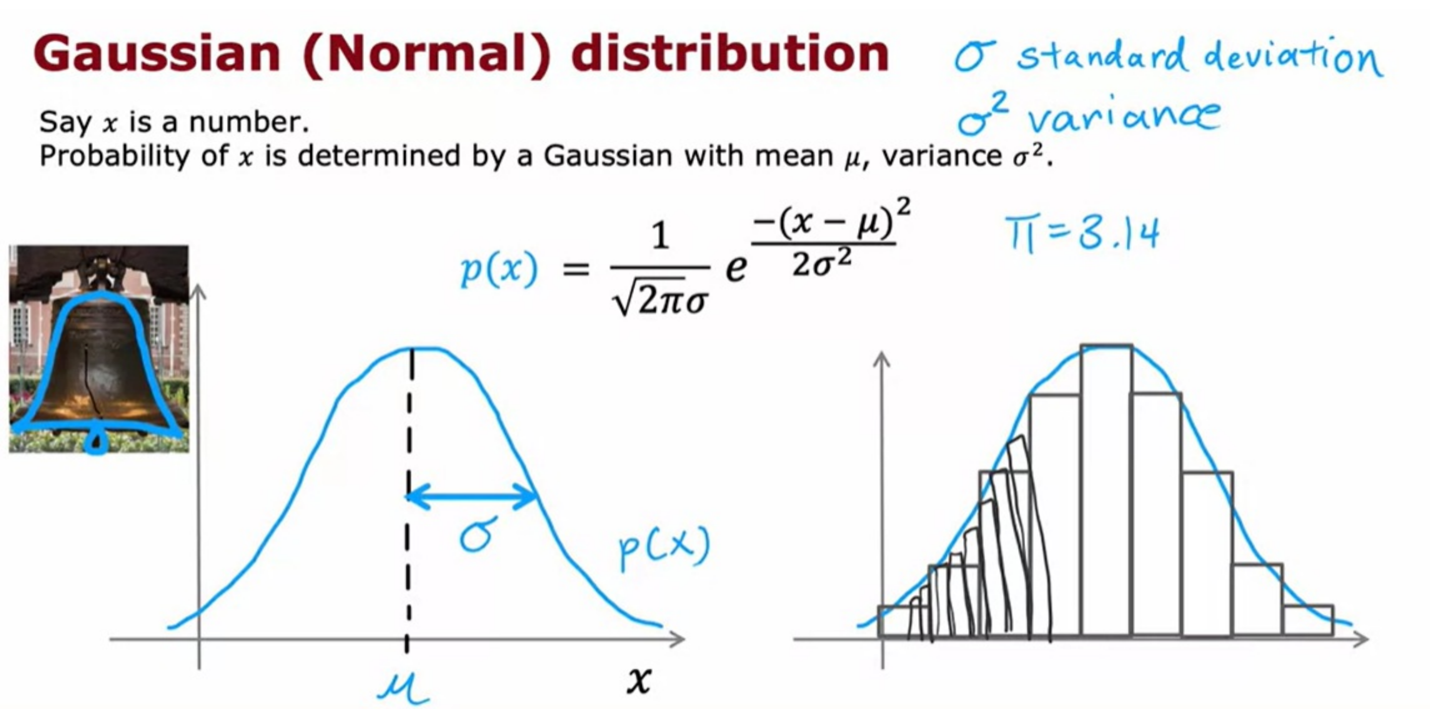


Density Estimation Formula

- Recall that x = heat, y = vibration

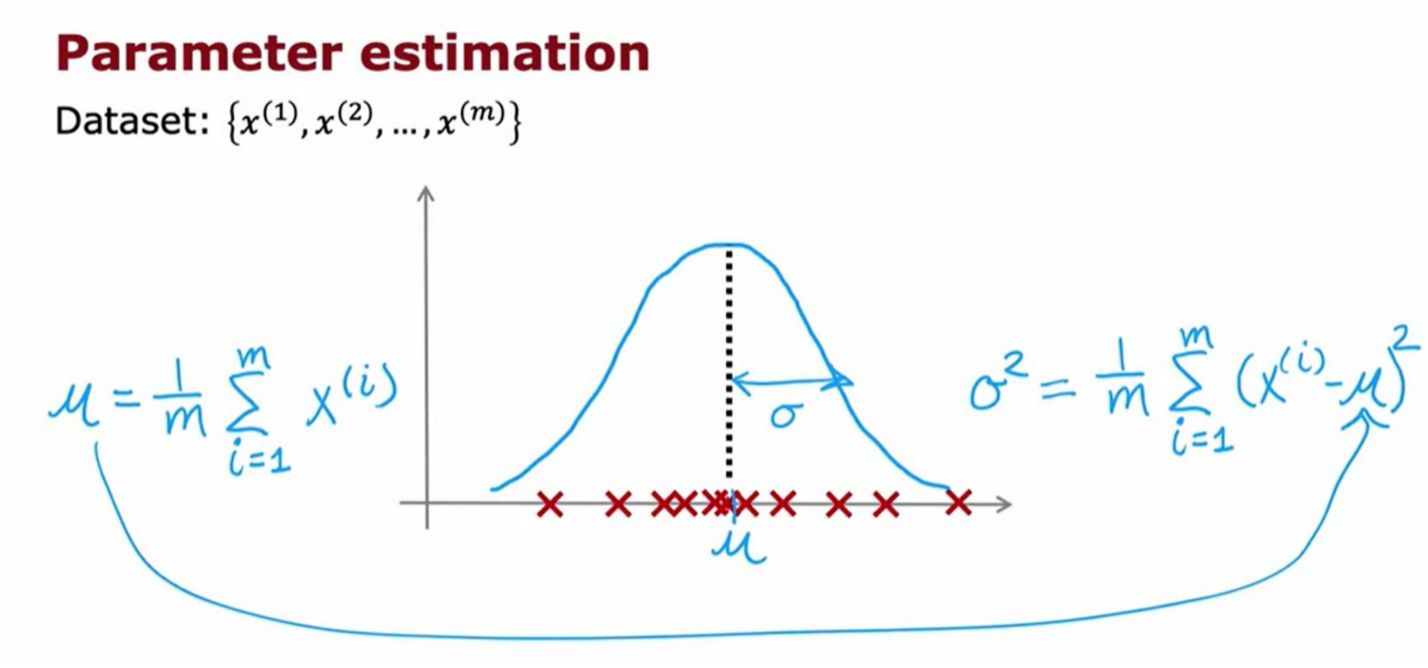


Gaussian Distribution



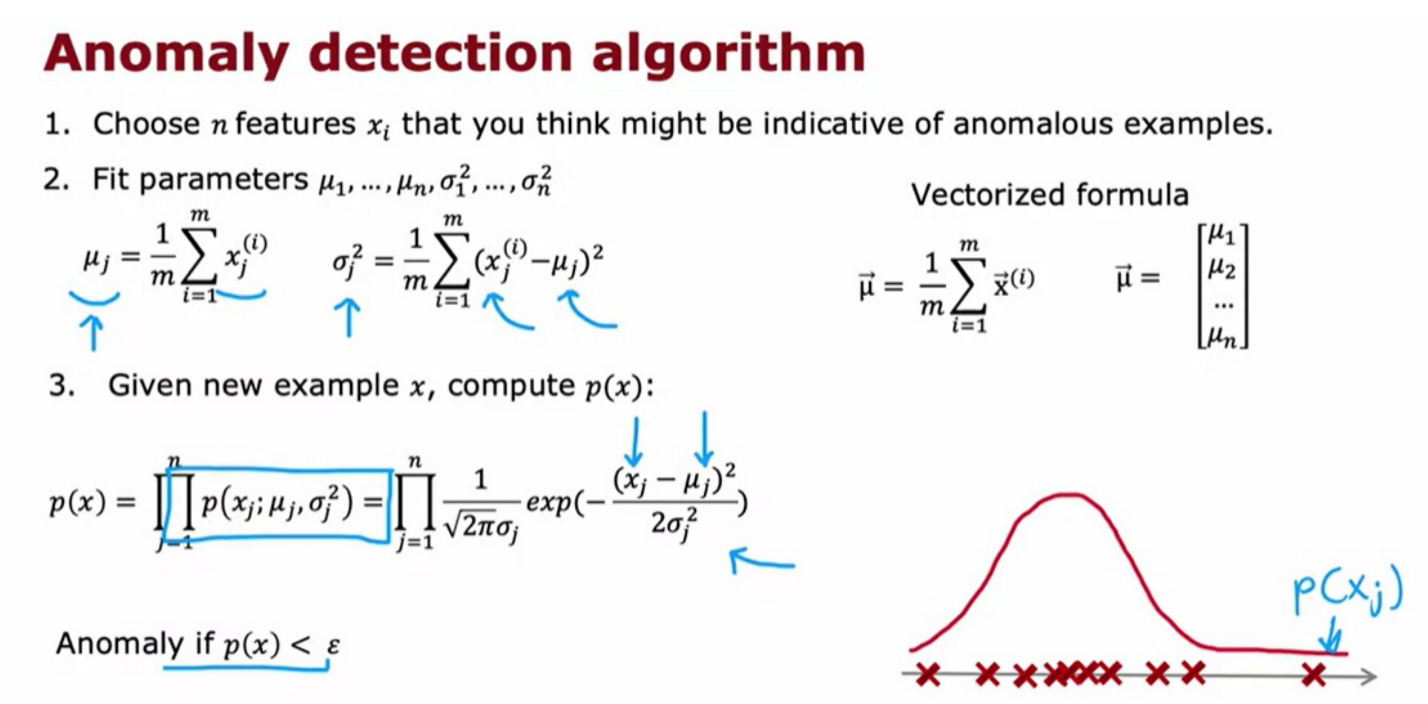
Parameter Estimation

- Maximum likelihood estimates formula
- Note: 1/m1/m1/m can be 1/m−11/m-11/m − 1 for variance

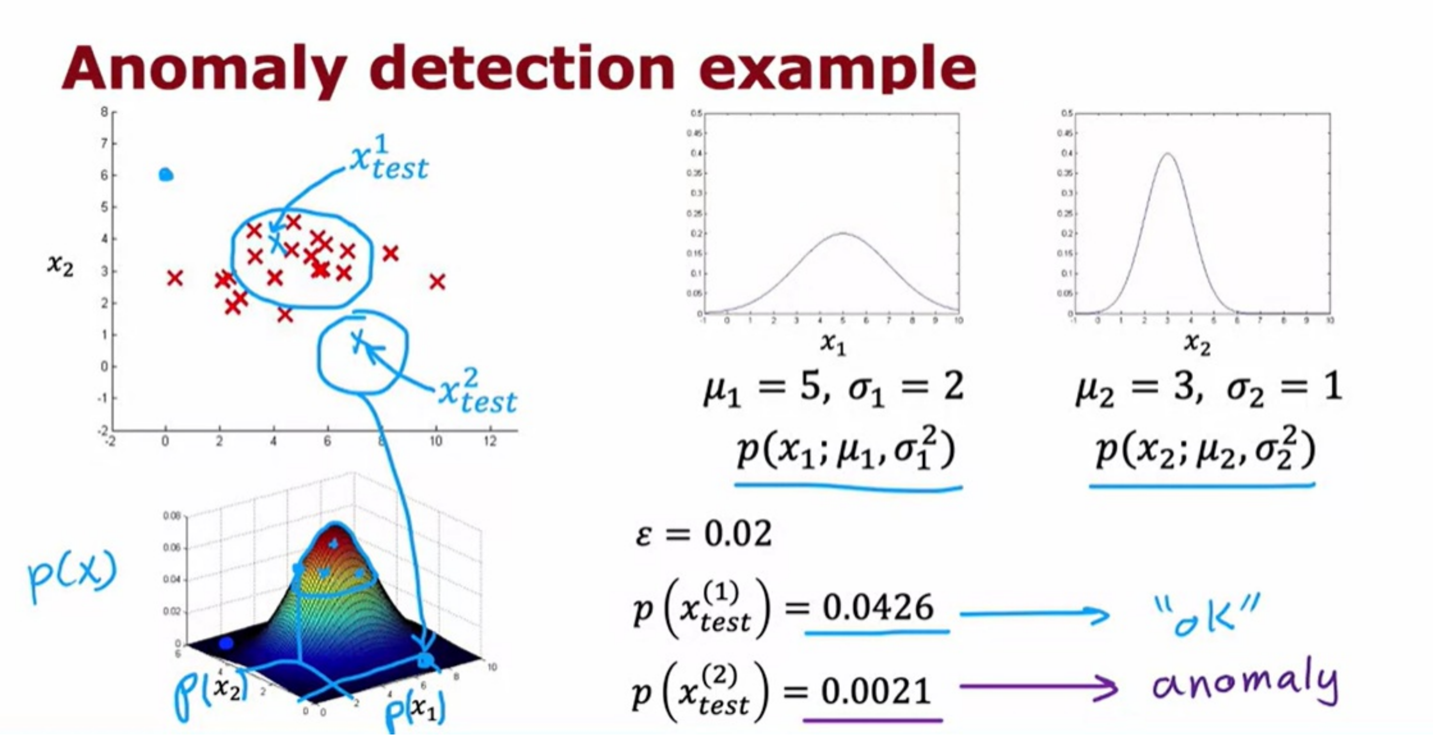


Anomaly Detection Algorithm

- Attempt to form a bell-shaped curve on the current datasets. Any data outside of the curve will have a low probability, and thus flagged as an anomaly

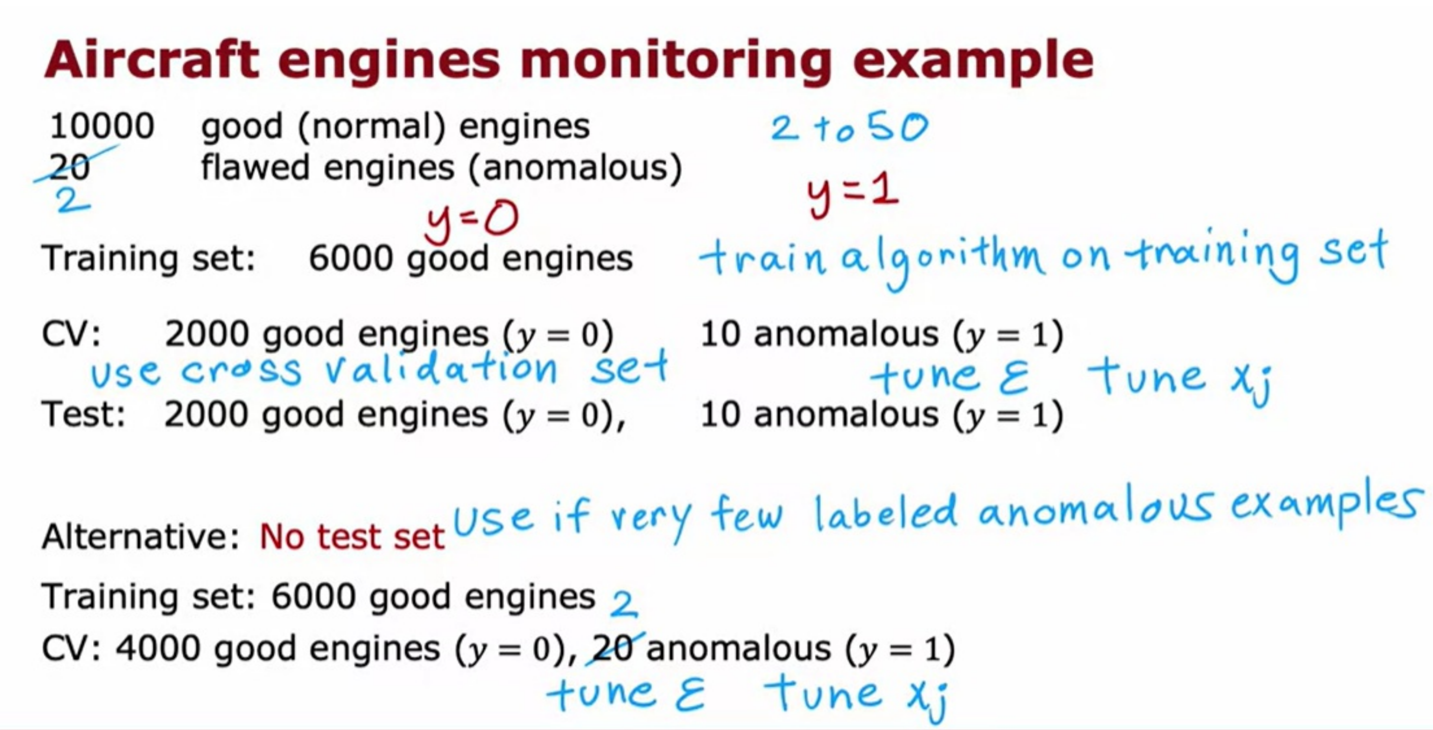


Anomaly Detection Algorithm Example

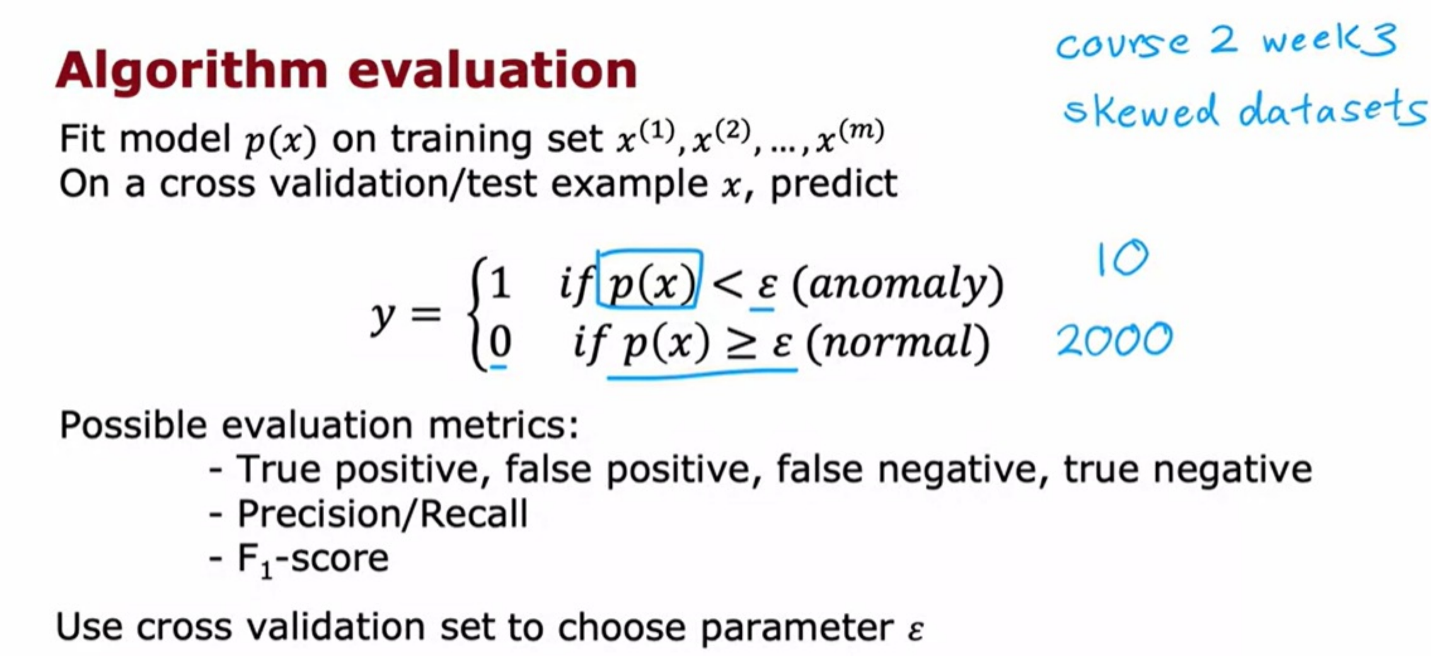


Developing and Evaluating Anomaly Detection System

- Training set, cross validation set, test set
- Test set can be omitted if very few labeled anomalous examples
- Caveat: Higher risk of overfitting, can't test the model in the future

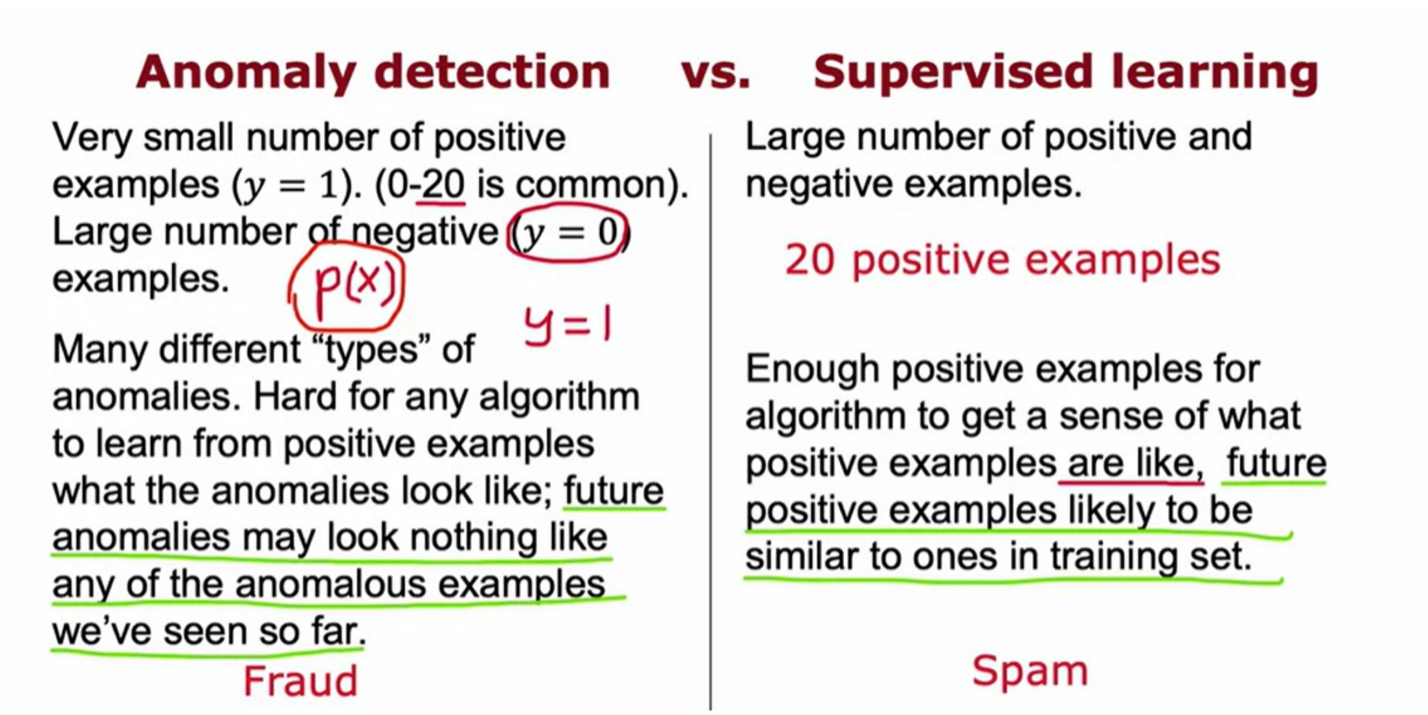


- Predicting an anomaly:

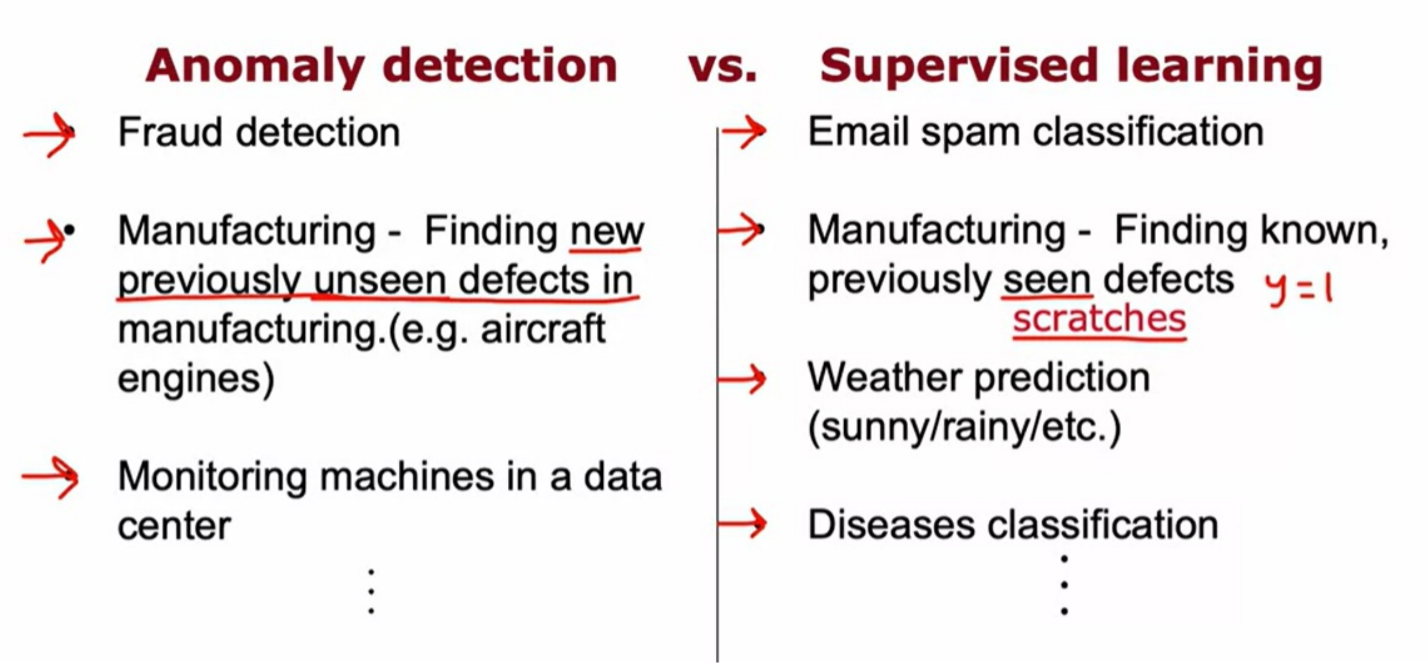


Anomaly Detection vs Supervised Learning

- How to decide between supervised learning and anomaly detection
- More variations → use anomaly detection
- Less types of data → supervised learning

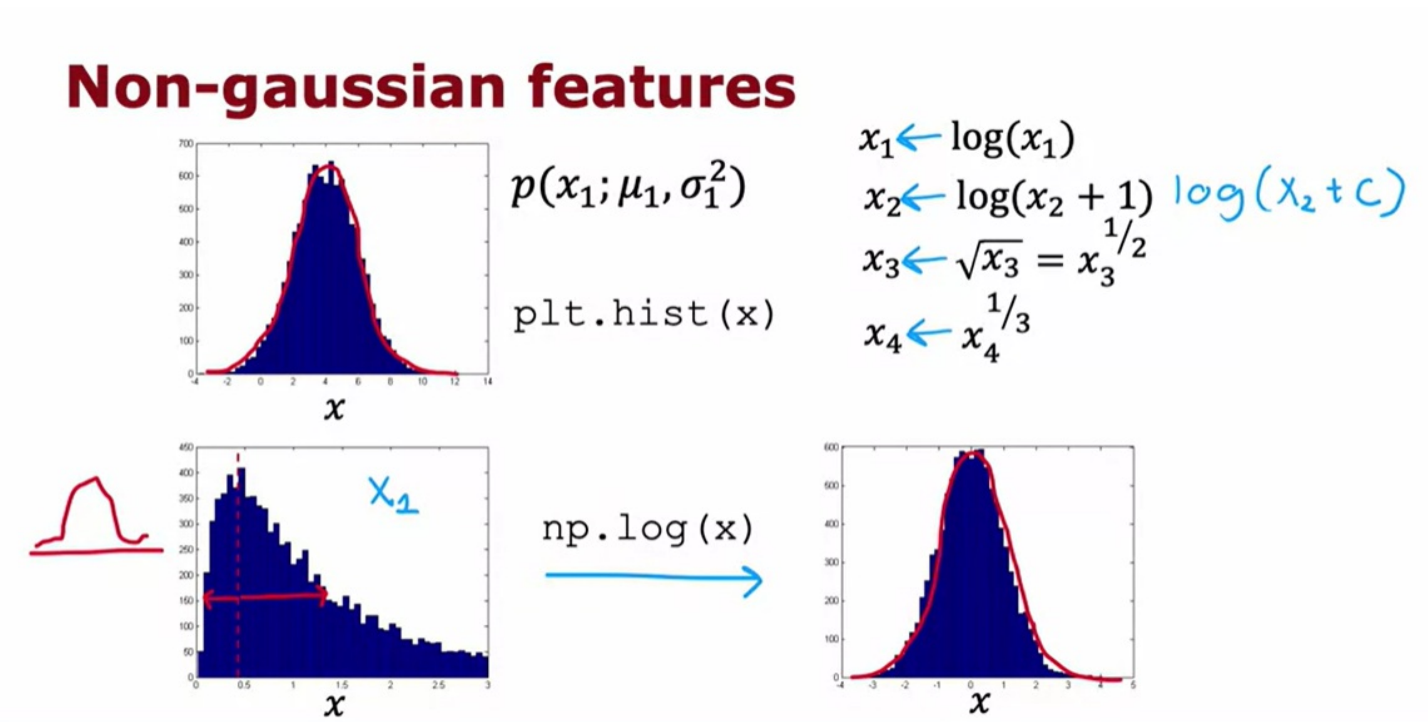


More Use Cases



Choosing Features for Anomaly Detection

- Select features that makes your data look Gaussian



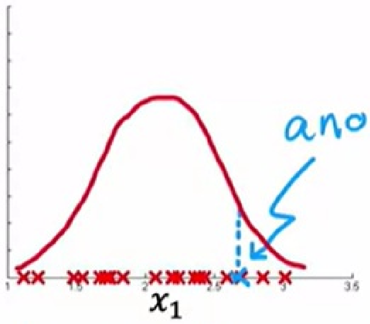
Error Analysis

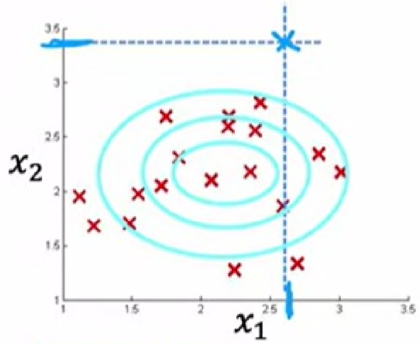
- If p(x)p(x)p(x) for both normal and anomalous is large, anomalous data can be missed
- Choose features that are likely to be more variable

Error analysis for anomaly detection

Want $p(x)$ $\geq \varepsilon$ large for normal examples x.
$p(x)$ $< \varepsilon$ small for anomalous examples x.

Most common problem:
$p(x)$ is comparable for normal and anomalous examples.
($p(x)$ is large for both)





X_1 num transactions X_2 typing speed