This module covers…

- Plot Diagrams of low dimensional data sets like Box Plot, Run Chart, Scatter Plot and Histograms
- Draw conclusions out of the diagrams you've plotted
- Reduce dimensions of your data set
- Understand the concept of Phase Diagrams

In this Video you will learn…

Dimensionality Reduction

Principal Component Analysis

- Take a n - dimensional, euclidian vector space R^n
- Span a new k - dimensional, euclidian vector space R^k such that
  - k < n
  - $\frac{dist(p_a^n, p_b^n)}{dist(p_c^n, p_d^n)} = \frac{dist(p_a^k, p_b^k)}{dist(p_c^k, p_d^k)}, p_a^n, p_b^n, p_c^n, p_d^n \in R^n, p_a^k, p_b^k, p_c^k, p_d^k \in R^k$
  - k dimensions “explain” most of the variation
- Idea:
  - If k=3 plottable and distance ratios are preserved

information loss

- PCA is loose full compression
- the lower k is, the higher the loss
- loss measured by reconstruction error
  - $loss = sse(D, PCA^{-1}(PCA(D)))$
  - $sse(X, Y) = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}, X = X_1, X_2, .., X_N, X_n = x_n^1, x_n^2, .., x_n^d$

Summary

- dimensions of high dimensional data can be reduced
  - preserving key properties
  - making it feasible to plot
- information gets lost
- amount of loss can be quantified

The next video covers…

PCA in ApacheSpark / MLLib