JPL-Caltech Virtual Summer School

Big Data Analytics

September 2 – 12, 2014

Ciro Donalek (Caltech)

Unsupervised Learning

Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology

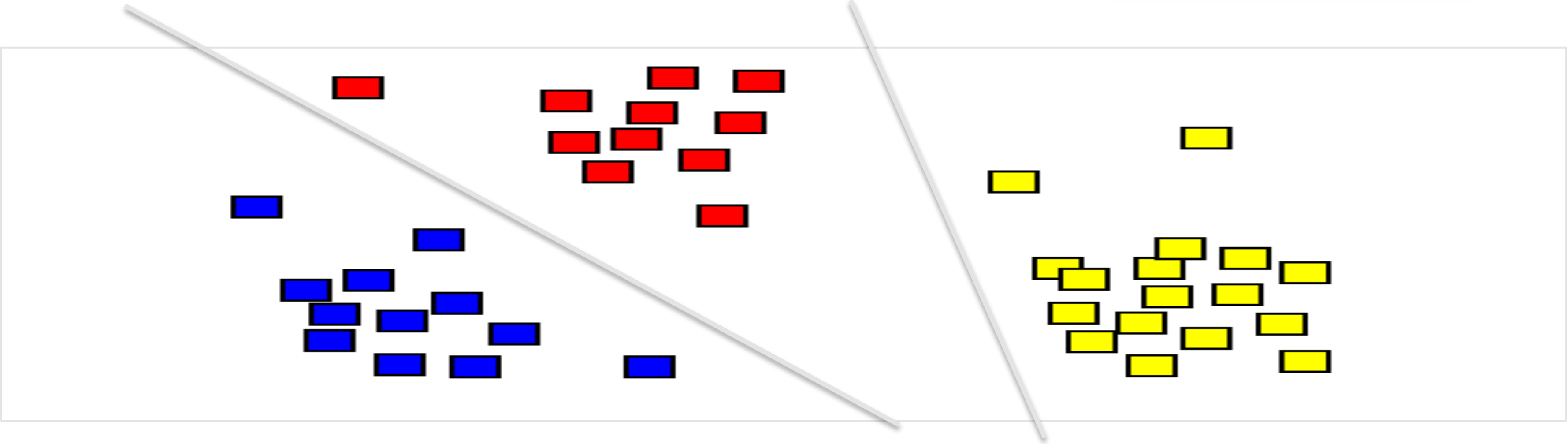
Outline

- Unsupervised Learning definition
- Clustering
- Core concepts: labeling, distances, measures
- Reinforcement Learning

Unsupervised Learning

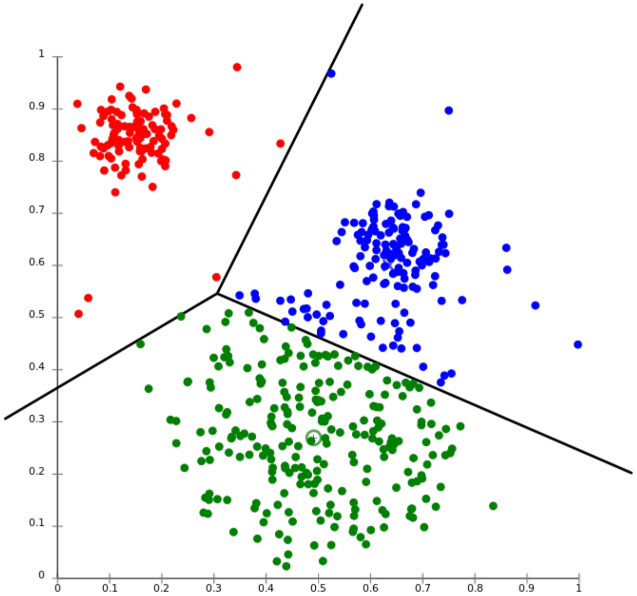
- Supervised: (input, correct output)
  Unsupervised Learning: (input)
- The model is **not** provided with the correct results during the training.
- Can be used to cluster the input data in classes **on the basis of their statistical properties only**.

Unsupervised Algorithms
K-Means
Self-Organizing Maps
RDF
Fuzzy Clustering
CURE
ROCK
Vector Quantization
Probabilistic Principal Surfaces
...



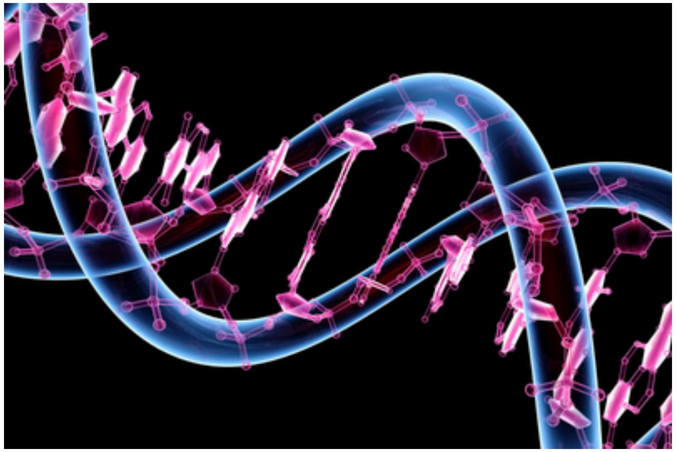
Goals of Unsupervised Learning

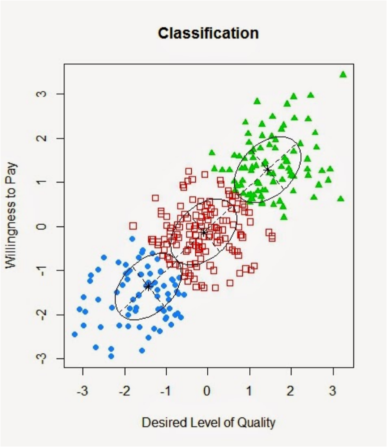
- Clustering: assigning a set of objects into groups so that the objects in the same cluster are more similar to each other than to those in other clusters.
- Search for outliers.
- Finding the hidden causes or sources of the data.
- Modeling the data density.
- Dimensionality reduction.



Some applications

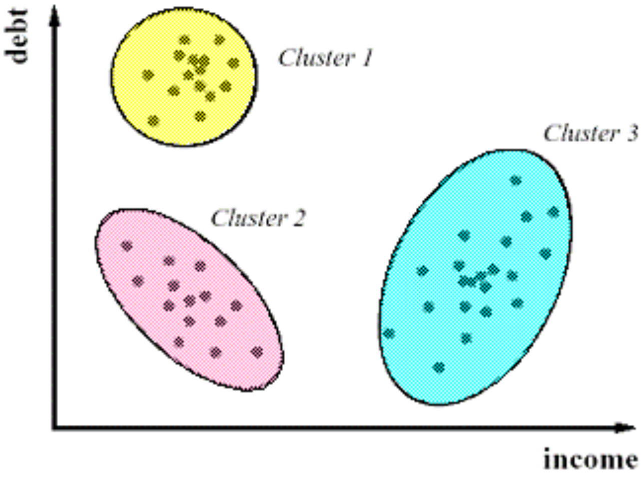
- Genomic: reduce the dimensionality of high dimensional data sets with hundreds of thousands of variables involved.
- Market research: discover distinct groups in their customer bases.
- Astronomical data analysis.
- Data compression, outlier detection, social network analysis, image segmentation, finance, etc.





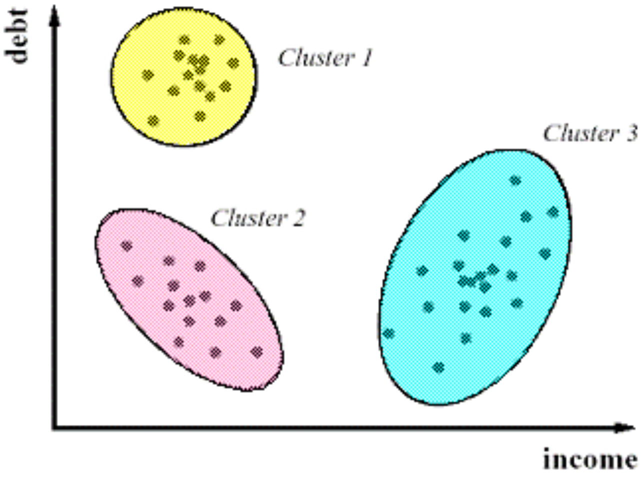
Types of Clustering

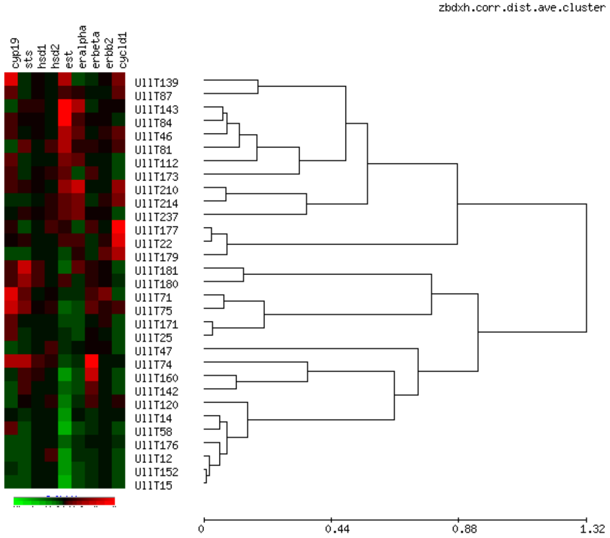
- PARTITIONING: construct various partitions and then evaluate them based on some criterion.



Types of Clustering

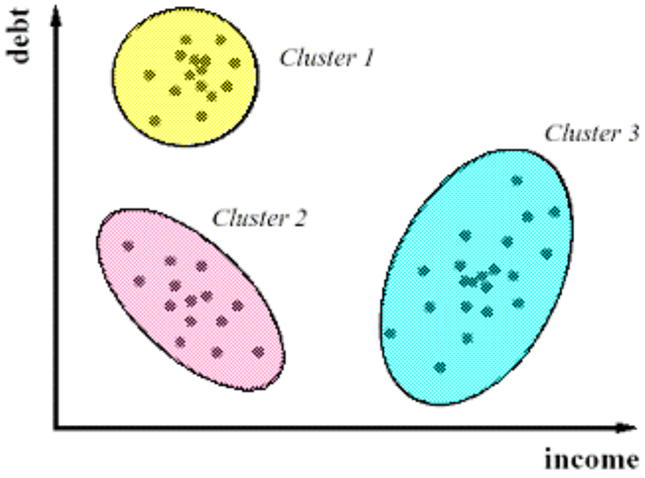
- PARTITIONING: construct various partitions and then evaluate them based on some criterion.
- HIERARCHICAL: finds successive clusters using previously established clusters (can be agglomerative or divisive).

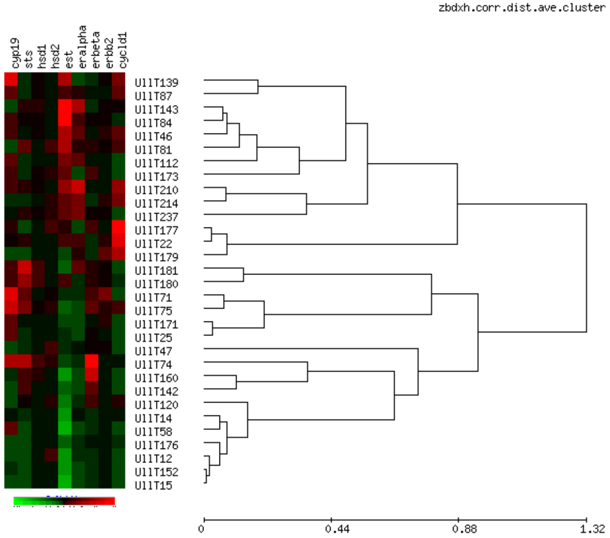


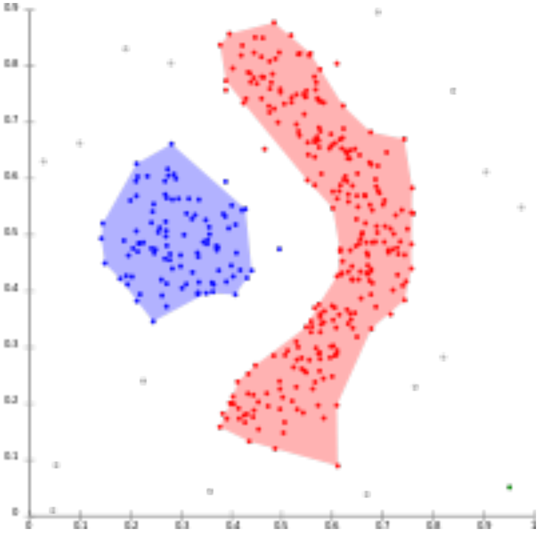


Types of Clustering

- PARTITIONING: construct various partitions and then evaluate them based on some criterion.
- HIERARCHICAL: finds successive clusters using previously established clusters (can be agglomerative or divisive).
- DENSITY-BASED: based on connectivity and density functions.

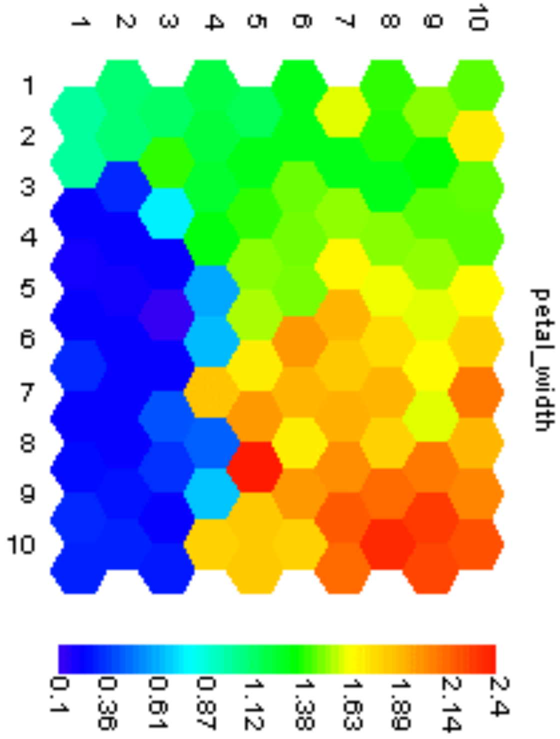






Labeling

- Assign a known label to an object.
- The labeling can be carried out even if the labels are only available for a small number of objects representative of the desired classes.
- Helps to identify clusters.

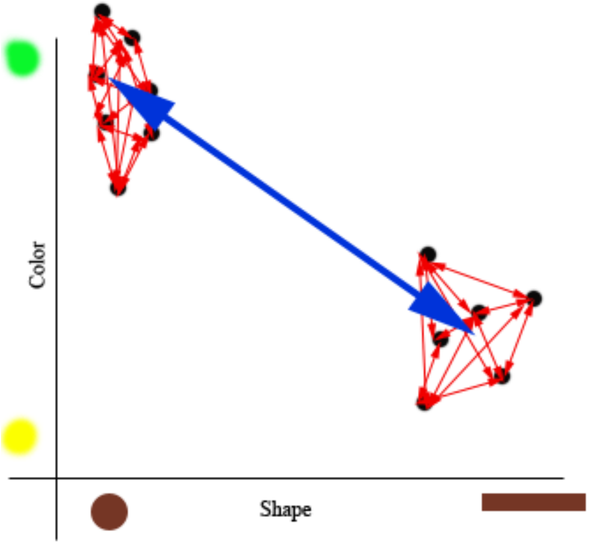


Good Clustering?

- Clusters can be evaluated with "internal" as well as "external" measures
  - Internal measures are related to the inter/intra cluster distance
  - External measures are related to how representative are the current clusters to "true" classes.

Good Clustering?

- Clusters can be evaluated with “internal” as well as “external” measures:
  - internal measures are related to the inter/intra cluster distance;
  - external measures are related to how representative are the current clusters to “true” classes.
- A good clustering is one where:
  - the intra-cluster distance is minimized: defined as the sum of distances between objects in the same clusters;
  - the inter-cluster distance is maximized: defined as the distances between different clusters.
- Define “distance”.



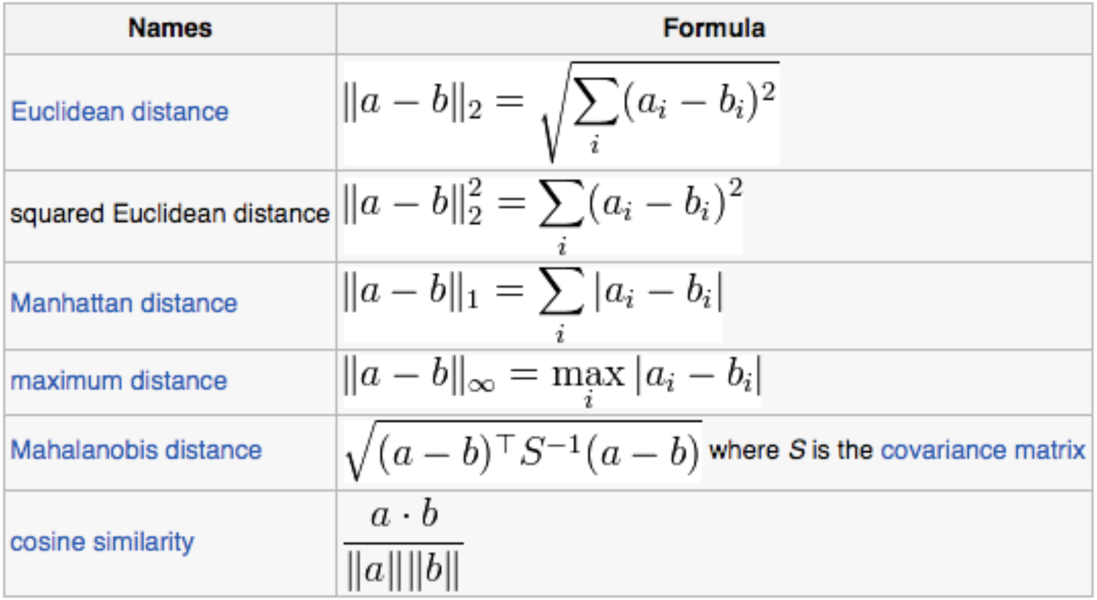
Distance between clusters

- Single link
  - smallest distance between an element in one cluster and an element in the other
    $dis(K_i, K_j) = min(t_{ip}, t_{jq})$
- Complete link
  - largest distance between an element in one cluster and an element in the other
    $dis(K_i, K_j) = max(t_{ip}, t_{jq})$
- Average
  - average distance between an element in one cluster and an element in the other
  - i.e., $dis(K_i, K_j) = avg(t_{ip}, t_{jq})$
- Centroid
  - distance between the centroids of two clusters
    $dis(K_i, K_j) = dis(C_i, C_j)$
- Medoid
  - distance between the medoids of two clusters
    $dis(K_i, K_j) = dis(M_i, M_j)$

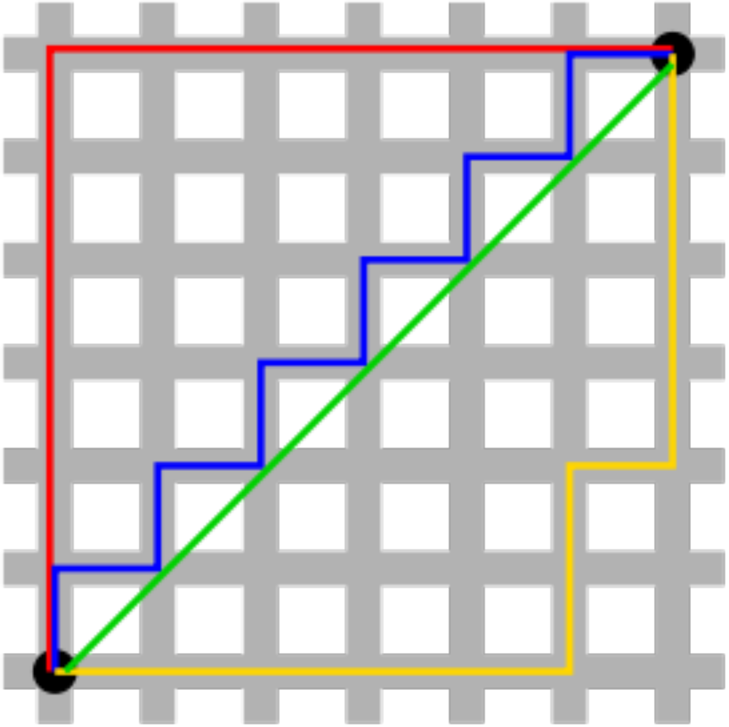
Credit: Data Mining: Concepts and Techniques, J. Han, M. Kamber

Distances

- Determine the similarity between two clusters and the shape of the clusters.



Names	Formula
Euclidean distance	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
squared Euclidean distance	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Manhattan distance	$\|a-b\|_1 = \sum_i \|a_i - b_i\|$
maximum distance	$\|a-b\|_\infty = \max_i \|a_i - b_i\|$
Mahalanobis distance	$\sqrt{(a-b)^\top S^{-1}(a-b)}$ where S is the covariance matrix
cosine similarity	$\frac{a \cdot b}{\|a\|\|b\|}$



In cases of Strings

- The **Hamming distance** between two strings of equal length is the number of positions at which the corresponding symbols are different.
  - measures the minimum number of *substitutions* required to change one string into the other

100**1001**
100**0100**
HD=3

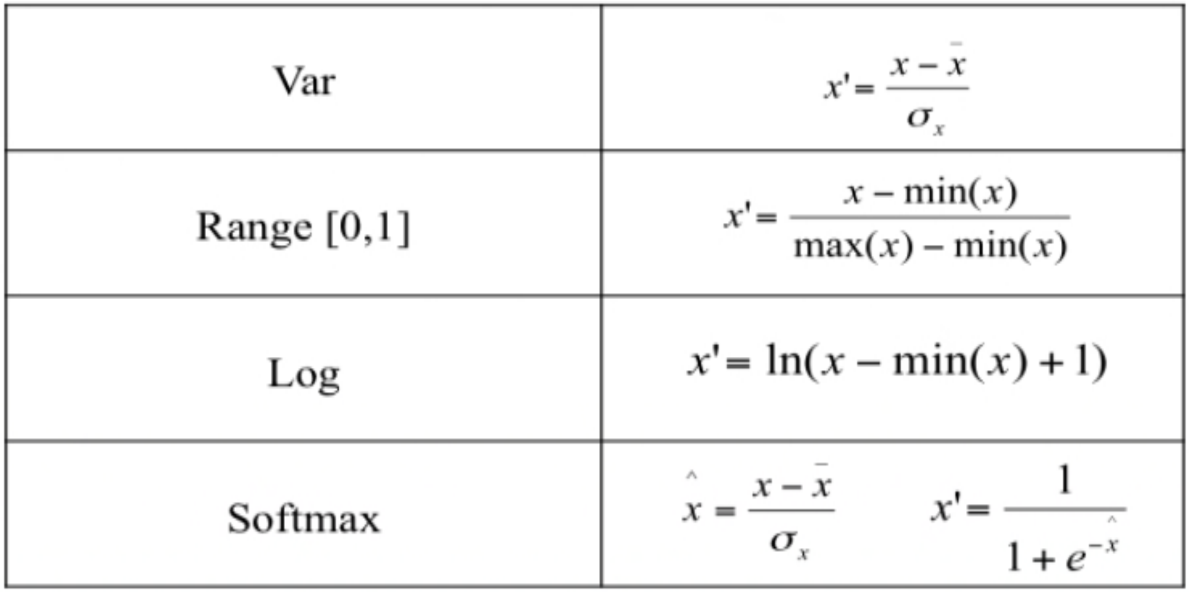
In cases of Strings

- The **Hamming distance** between two strings of equal length is the number of positions at which the corresponding symbols are different.
  - measures the minimum number of *substitutions* required to change one string into the other
- The **Levenshtein (edit) distance** is a metric for measuring the amount of differences between two sequences.
  - is defined as the minimum number of edits needed to transform one string into the other.

1001001
1000**100**
HD=3

LD(BIOLOGY, BIOLOGIA)=2
BIOLOG**Y** -> BIOLOG**I (substitution)**
BIOLOGI -> BIOLOGI**A (insertion)**

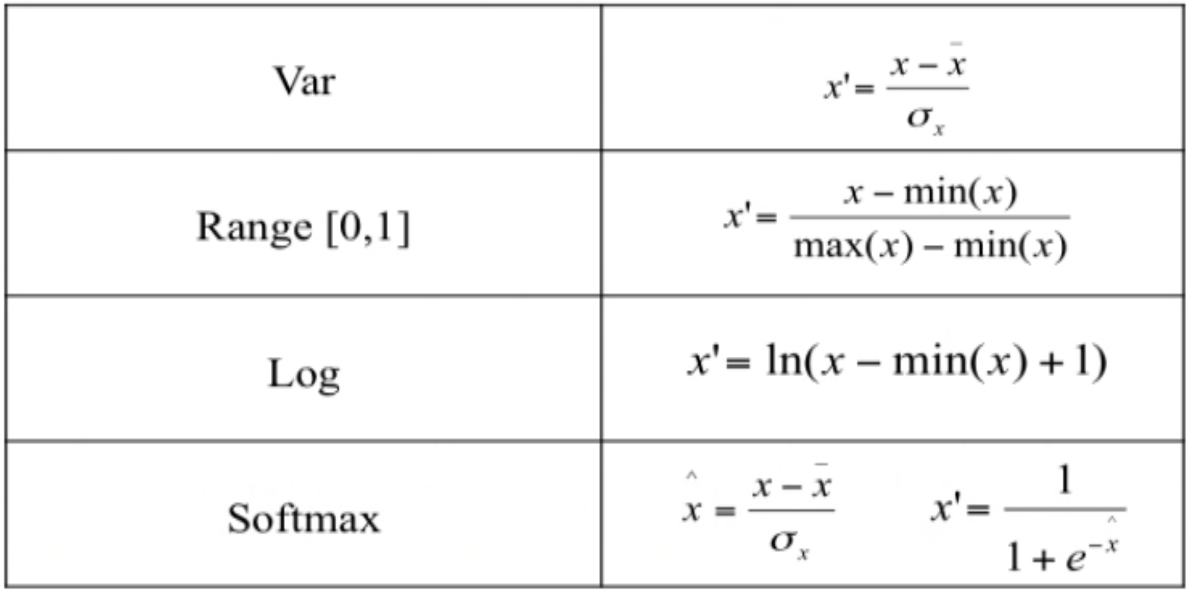
Normalization



Var	$x' = \frac{x - \bar{x}}{\sigma_x}$
Range [0,1]	$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$
Log	$x' = \ln(x - \min(x) + 1)$
Softmax	$\hat{x} = \frac{x - \bar{x}}{\sigma_x}$ $x' = \frac{1}{1 + e^{-\hat{x}}}$

VAR: the mean of each attribute of the transformed set of data points is reduced to zero by subtracting the mean of each attribute from the values of the attributes and dividing the result by the standard deviation of the attribute.

Normalization



Var	$x' = \dfrac{x - \bar{x}}{\sigma_x}$
Range [0,1]	$x' = \dfrac{x - \min(x)}{\max(x) - \min(x)}$
Log	$x' = \ln(x - \min(x) + 1)$
Softmax	$\hat{x} = \dfrac{x - \bar{x}}{\sigma_x}$ $\quad x' = \dfrac{1}{1 + e^{-\hat{x}}}$

VAR: the mean of each attribute of the transformed set of data points is reduced to zero by subtracting the mean of each attribute from the values of the attributes and dividing the result by the standard deviation of the attribute.

RANGE (Min-Max Normalization): subtracts the minimum value of an attribute from each value of the attribute and then divides the difference by the range of the attribute. It has the advantage of preserving exactly all relationship in the data, without adding any bias.

SOFTMAX: is a way of reducing the influence of extreme values or outliers in the data without removing them from the data set. It is useful when you have outlier data that you wish to include in the data set while still preserving the significance of data within a standard deviation of the mean.

Reinforcement Learning

- Supervised Learning: (input, correct output)
  Unsupervised Learning: (input)
  Reinforcement Learning: (input, *some output, grade for this output*)
- Most common applications: gaming.

In 1994, Chinook became the first program in any game to win a human World Championship. By 1996, it became clear that the program was much stronger than any human, and Chinook was retired.

Summary

- Unsupervised Learning
- Goal and Applications
- Clustering notions
- Reinforcement learning