Machine Learning

What is “Machine Learning”?

- Machine learning is a branch of artificial intelligence concerned with the design and development of algorithms that allow computers to evolve behaviours based on empirical data

- Things ‘learn’ when they change their behaviour in a way that makes them perform better

- *“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”*

  Tom M. Mitchell

Why do machines need to learn?

- Some tasks cannot be defined except by example; that is, we might be able to specify input/output pairs but not a concise relationship between inputs and desired outputs.
- The amount of knowledge available about certain tasks might be too large for explicit encoding by humans. Machines that learn this knowledge might be able to capture more of it than humans could
- Computers can adjust their internal structure to produce correct outputs for a large number of sample inputs and thus suitably constrain their input/output function to approximate the relationship

Why do machines need to learn?

- It is possible that hidden among large amounts of data are important relationships and correlations. Machine learning methods can often be used to extract these relationships
- Environments change over time. Machines that can adapt to a changing environment would reduce the need for constant redesign.
- New knowledge about tasks is constantly being discovered by humans. Continuing redesign of AI systems to conform to new knowledge is impractical, but machine learning methods might be able to track much of it.

Machine Learning Applications

- *Speech recognition.* Currently available commercial systems for speech recognition all use machine learning in one fashion or another to train the system to recognize speech

- *Computer vision.* Many current vision systems, from face recognition systems, to systems that automatically classify microscope images of cells

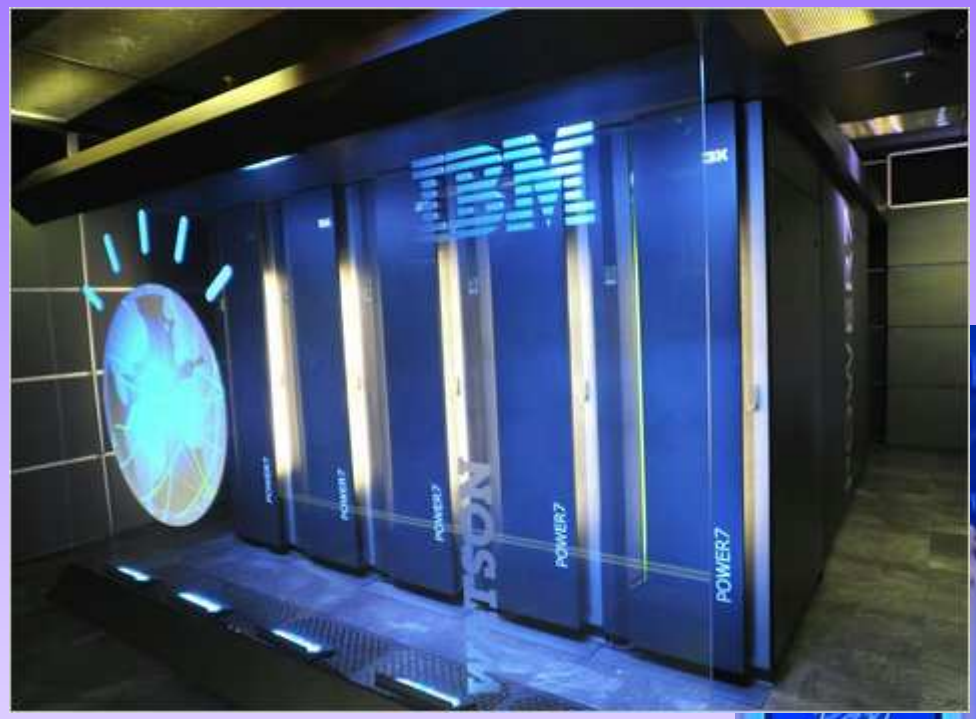
- *Bio-surveillance.* A variety of government efforts to detect and track disease outbreaks now use machine learning

- *Robot control.* Machine learning methods have been successfully used in a number of robot systems, including teaching a robot to fly and navigate a helicopter

Some other examples…

Some other examples…

Some other examples…




THINK
DENKE
$24,000
Who is Stoker?
$1,000
$77,147
Who is Bram Stoker?
$ 17,973
$21,600

US Census Data

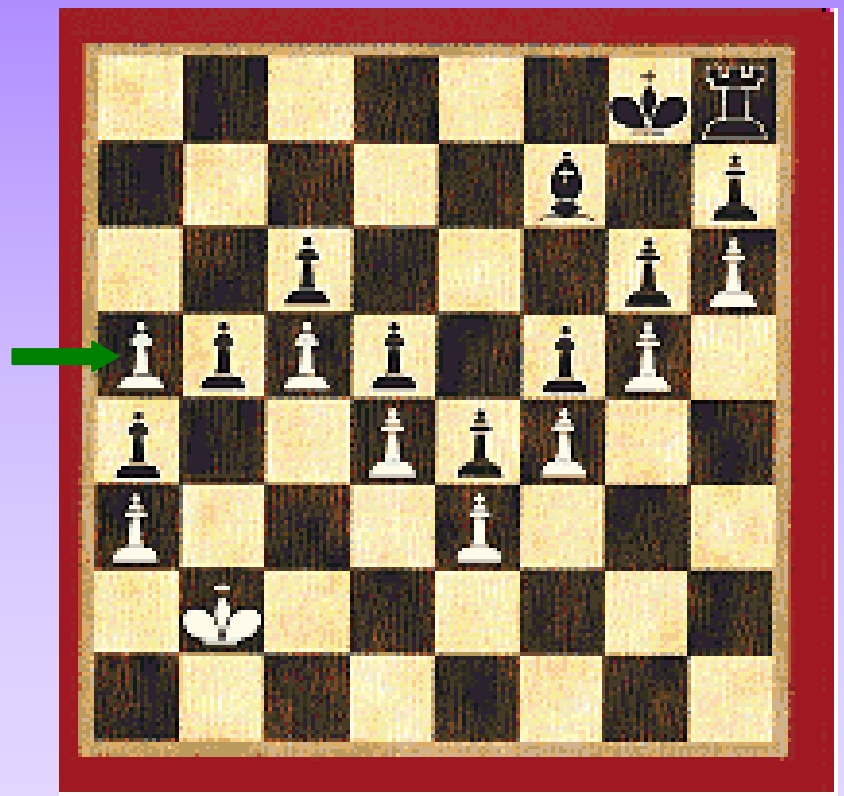
- Attempted to find unusual patterns among census data
- Majority of patterns discovered were obvious or facile
- ‘Interestingness’

The Case of the Poisoned Rook

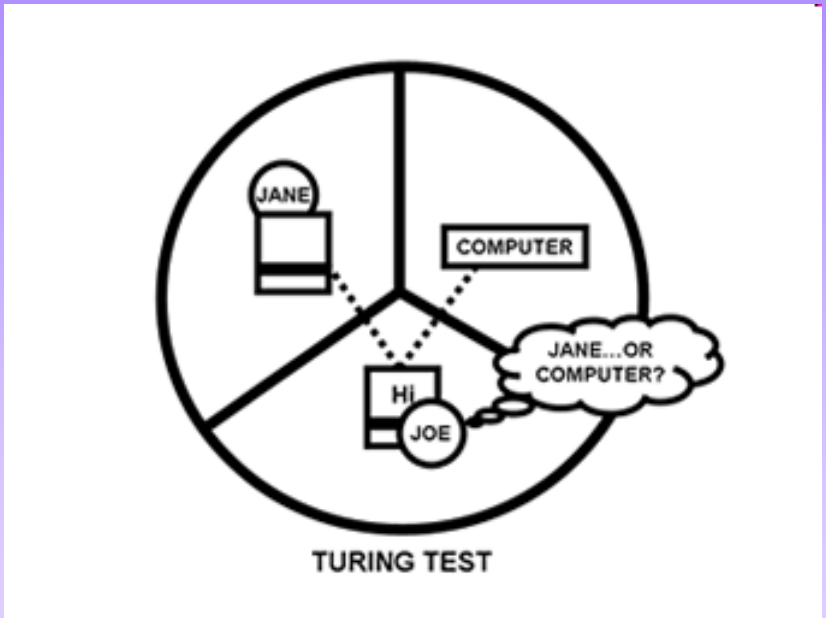
- A challenge for the Deep Thought Computer
- What is the best move for white to make?

The Case of the Poisoned Rook

- Deep Thought takes the rook and loses the game
- Needs to look 50 moves ahead to understand the strategy
- Common sense or consciousness?

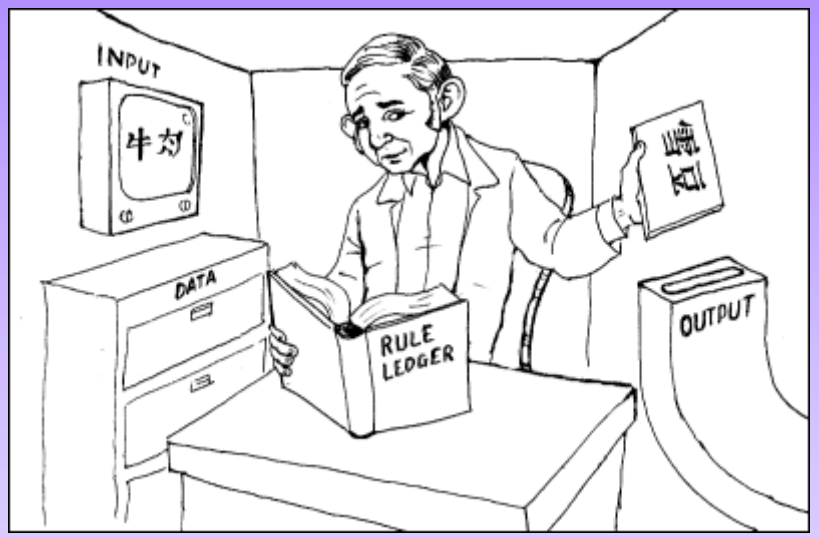


Artificial Intelligence



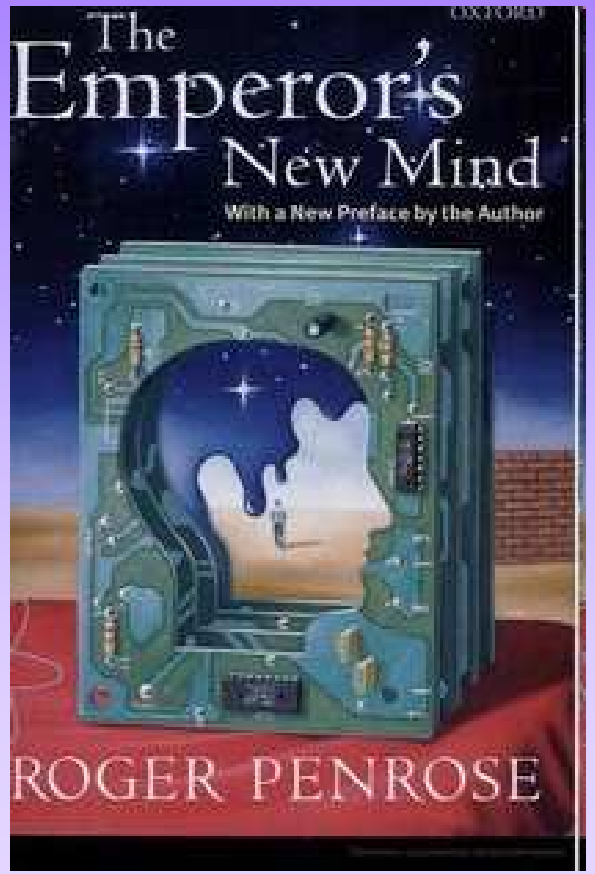
- Turing Test
- Can you tell a human from a computer?
- The ‘gold standard’ of Artificial Intelligence

Chinese Room



- Thought experiment by John Searle
- Data is processed by not understood
- Outside the room, the Turing test is satisfied

Roger Penrose



- **Human thought processes are non-computational**
- **Faster algorithms will not be able to replicate the human brain**
- **Quantum computing?**

Machine Learning

- Generally defined as ‘supervised’ or ‘unsupervised’

- Supervised algorithms
  - Associated with classification
  - Based upon a teaching signal

- Unsupervised algorithms
  - Associated with clustering
  - No teaching signal

Data Mining of Microarrays

- **Classification**
  - Normal vs Cancer expression profiles
  - Relating genotype to phenotype
  - Usually supervised

- **Clustering**
  - Co-expression of genes
  - Understanding biological pathways
  - Usually unsupervised

Clustering

- Assumed that the clusters resemble some form of natural group
- Can be a measure of similarity or dissimilarity within a dataset
- Some clusters may be overlapping or hierarchical

Vector Geometry

- The essence of cluster analysis is the ability to calculate a distance matrix
- Distance matrix exists between two entities in multi-dimensional space
- An equation is needed to measure distance

Distance calculation

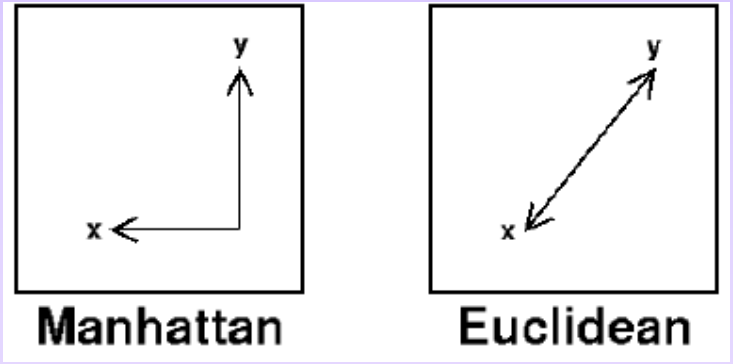
Assuming vectors are (x_1, x_2, x_3…) and (y_1, y_2, y_3…)

Manhattan distance (L_1 metric)

$$Manh(x,y) = \sum_{i=1}^{d} |x_i - y_i|$$

Euclidean distance (L_2 metric)

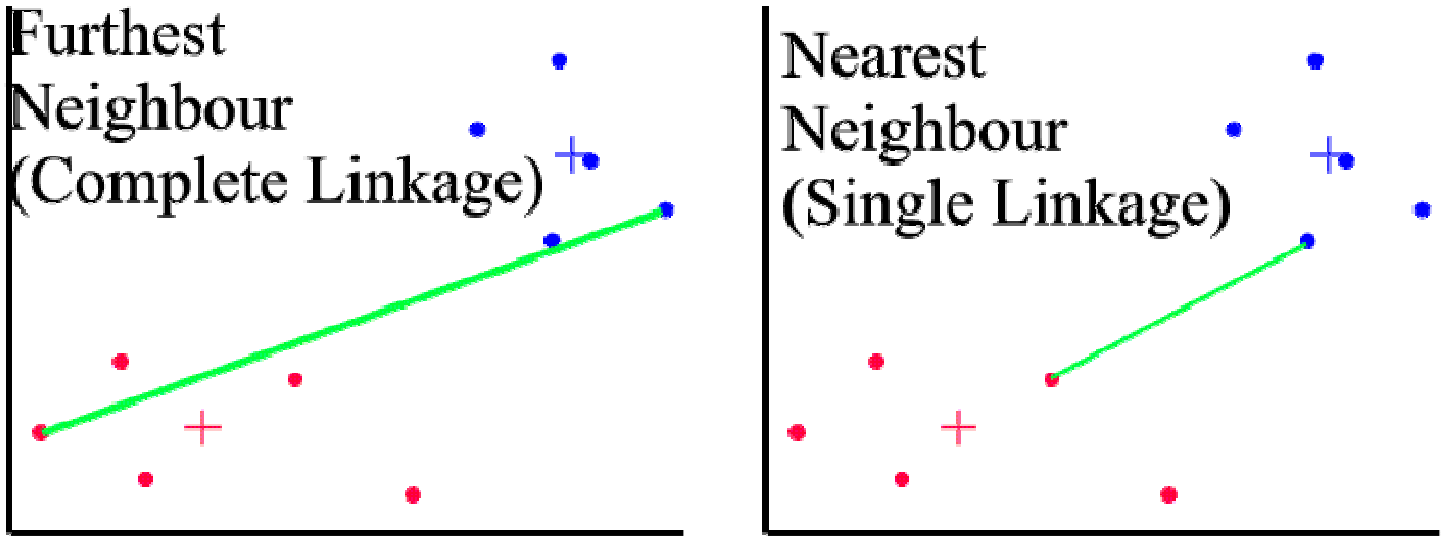
$$Eucl(x,y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$



Selecting Distance

- Numerous other metrics can be used to determine distance
- Power terms can accentuate larger distances in comparison to smaller distances
- Euclidean distance is reckoned to be a good compromise

Measuring distance between clusters



Clustering Mechanism

- Agglomerative – Starts with each instance as a single cluster
- Divisive – One single cluster that splits into many

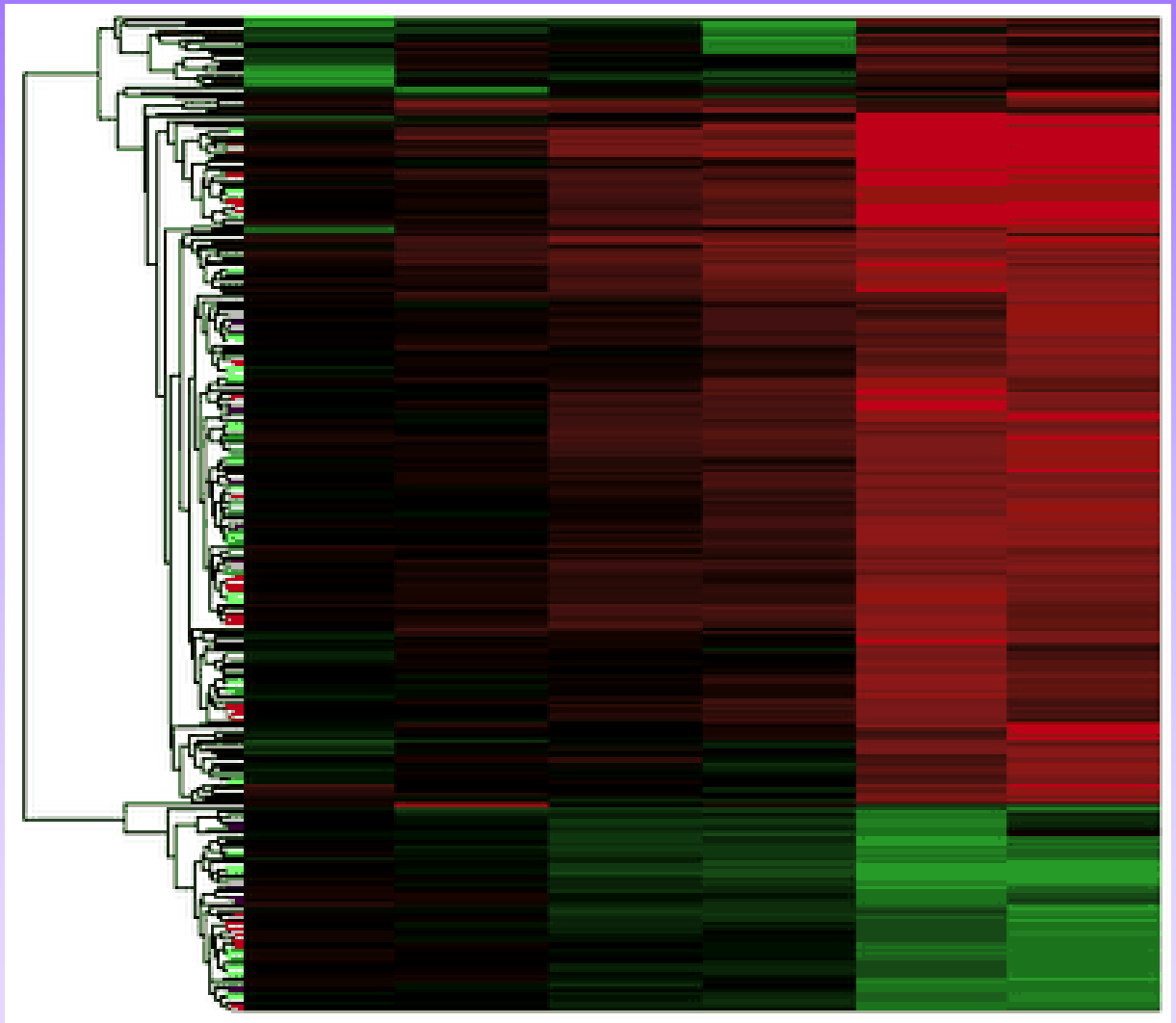
Hierarchical Clustering

- Splits from root depending on how cluster is related

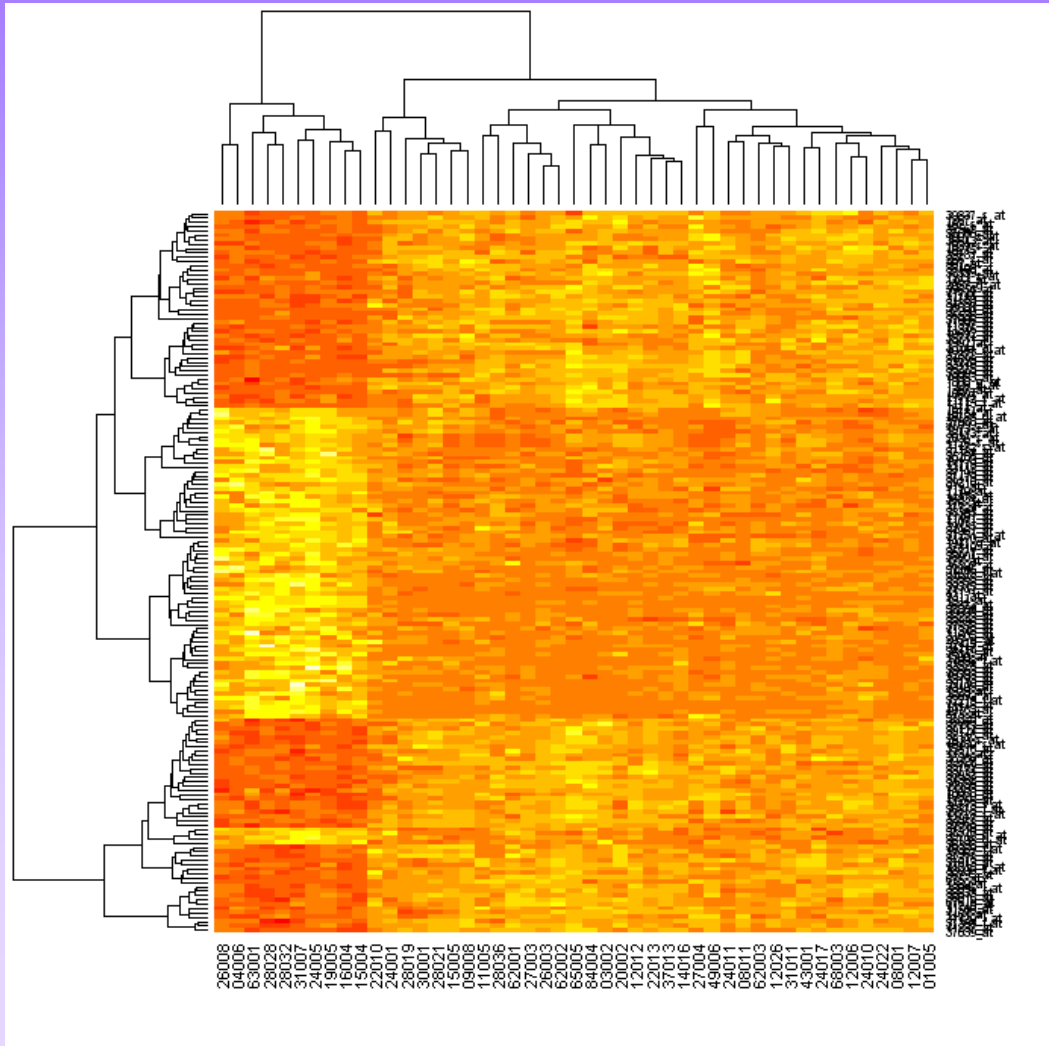
- Dendogram can be drawn so length of horizontal branches denote a measure of dissimilarity between classes

Gene clustering

- Dendogram shows a relationship between clusters



Heat Map

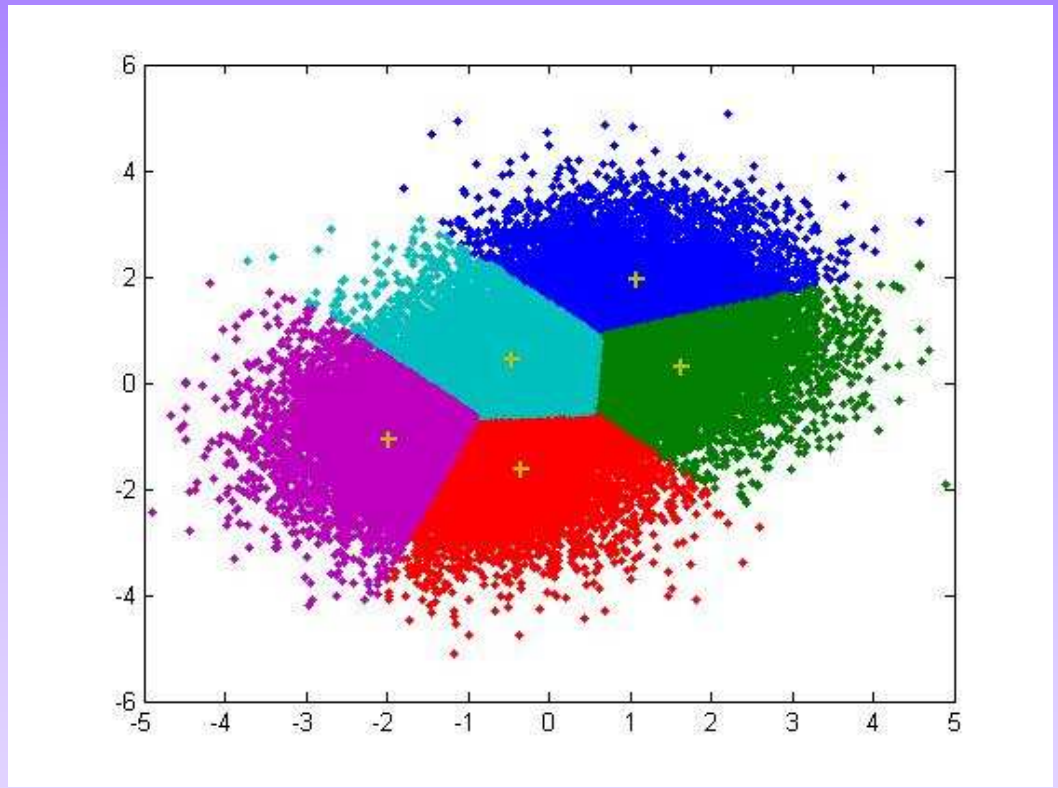


- Two way hierarchical clustering of both samples and genes
- Clear visualisation of expression values between many samples over a number of genes

K means clustering

- Partitioning the data into *k* number of clusters (*k* specified by user)
- k points are initially chosen at random
- Following clustering, new centroids are calculated and process begins again

K means clustering



- Iterative process
- Moves towards stable conformation
- Minimises total squared distance to cluster centres

Limitations of Clustering

- Does not make use of phenotypic information and therefore can not answer research questions related to a phenotype
- Cluster analysis is not based on any model and has no statistical or probabilistic basis
- No accepted estimation procedures which can be used to determine the number of clusters or prove that there is actually more than one cluster

Classification

- Each sample belongs to a predefined class
- Goal: discover a relationship which allows us to predict the class of an example, given its predictor attributes
- The relationship is discovered by using a **training set** which is then used to predict the class of examples in the **test set**

Data partitioning for classification

Training set
(known-class examples)

	…		class
			yes
			no
			no
			yes
			yes
			no

Test set
(unknown-class examples)

	…		class
			?
			?
			?
			?
			?
			?

Classification involves induction of a classification model from the training set and its application to new data in the test set (unseen during training)

Training and Test Data

- A common approach is to split data into 2/3 training and 1/3 test
- You may be unlucky with the distribution
- You may have very few samples
- Data Stratification

Cross-validation

- Decide on a fixed number of folds (partitions) of the data
- If k=5, the data is split into five approximately equally sized, stratified parts
- Each part in turn is used as the test set, the remaining four fifths are used for training
- Error estimates are averaged to give an overall error estimate over the five partitions

“Leave one out” Cross Validation

- Training the data by removing one sample at a time (k fold cross fold validation where k is equal to number of samples)
- Deterministic method (no random sampling)
- Makes fullest possible use of the data
- Data is inherently non-stratified

Bootstrapping

- Sampling with replacement to create a training set equal to the size of the original set
- Training set: draw N samples randomly from the original data set some samples may be left out, some may appear multiple times
- Test Set: Samples that are not drawn

Bootstrapping

- Bootstrapping is sometimes called 0.632 bootstrap

$$\left(1-\frac{1}{n}\right)^{n} \approx e^{-1} = 0.368$$

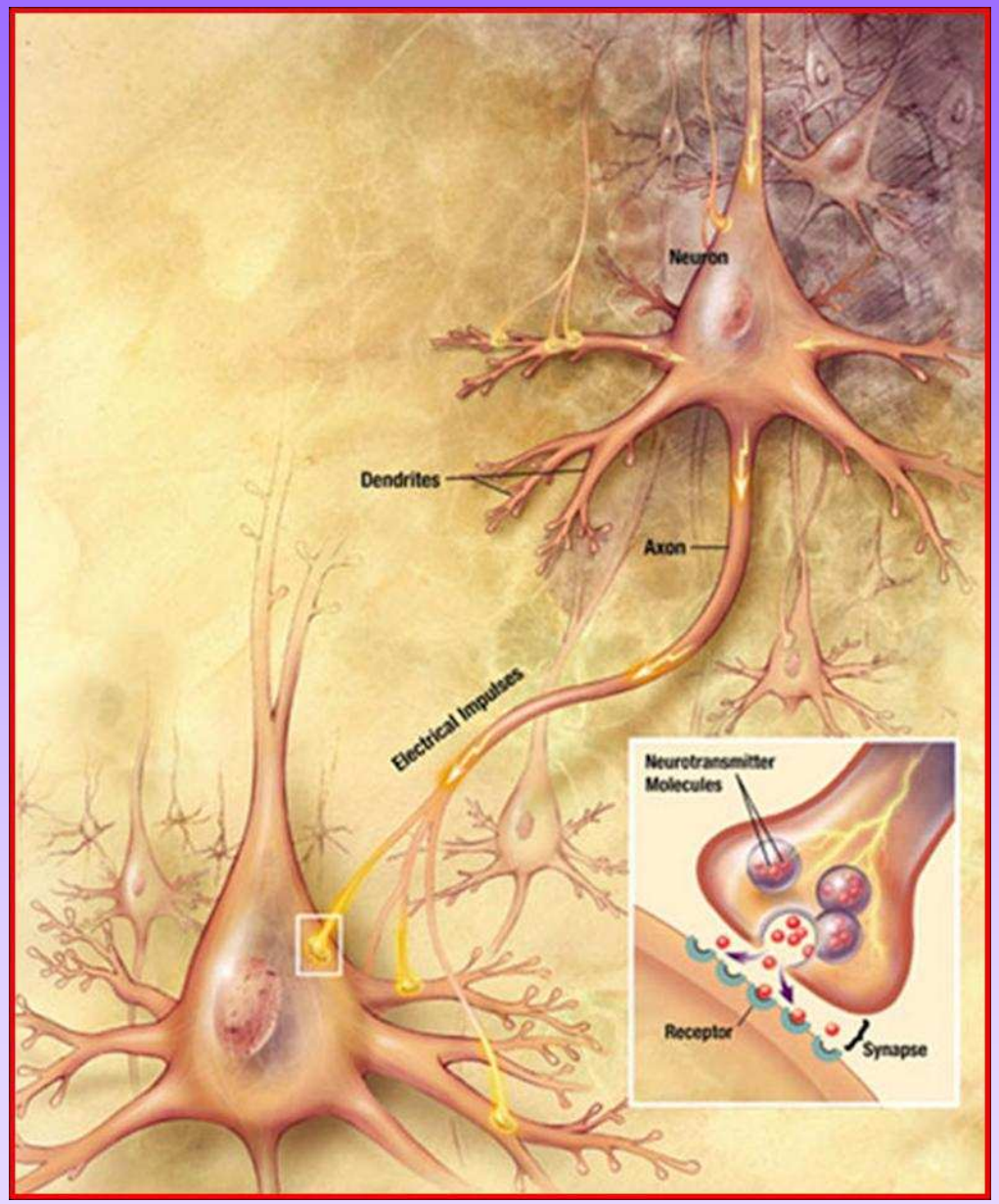
- The probability of not being picked by sampling is 36.8%

- Higher than average error rate if only 63.2% of the set is used for training

Monte Carlo Cross Validation

- Take several random drawings of training and test sets and evaluate the mean of generalization error
- Can use any number of random drawings
- Generalization error estimate is the mean of the errors computed by each random draw
- Because of the random drawings, it is not certain that each data item is used as often for training as testing

Biological Neural Networks



Dendrites
(Input layer)

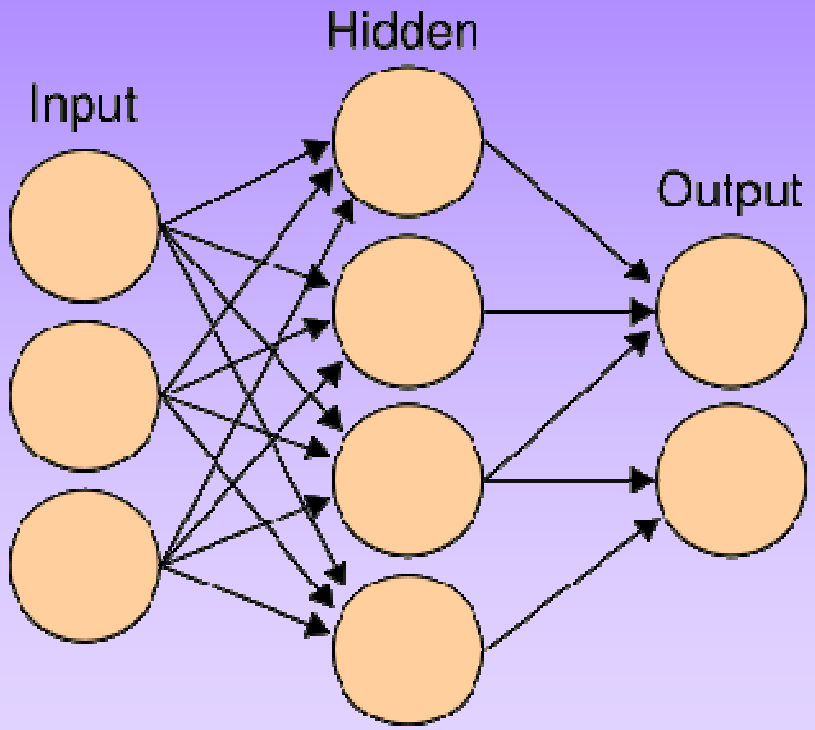
↓

Connected via synapses
to other neurons
(Connecting layer)

↓

Single axon
(Output layer)

Artificial Neural Networks

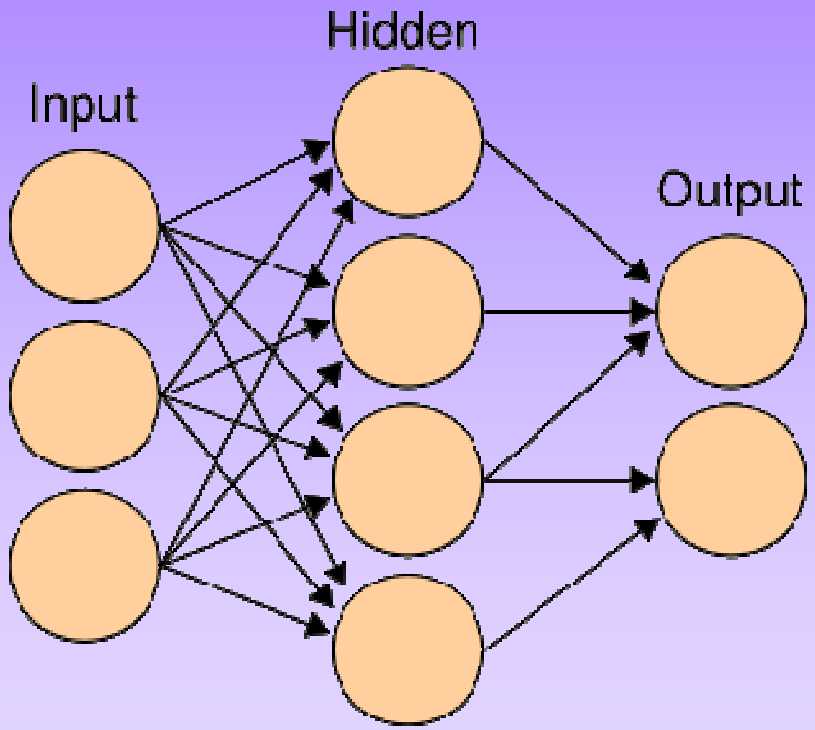


- Based loosely on biological network
- Series of nodes that process information using a connectionist approach
- Output node – Function of all input nodes

Artificial Neural Networks

- Many inputs determining one output
- Weights adjusted in hidden layer to gain desired output
- Classification operation is maximised through training and tested for accuracy

Artificial Neural Networks

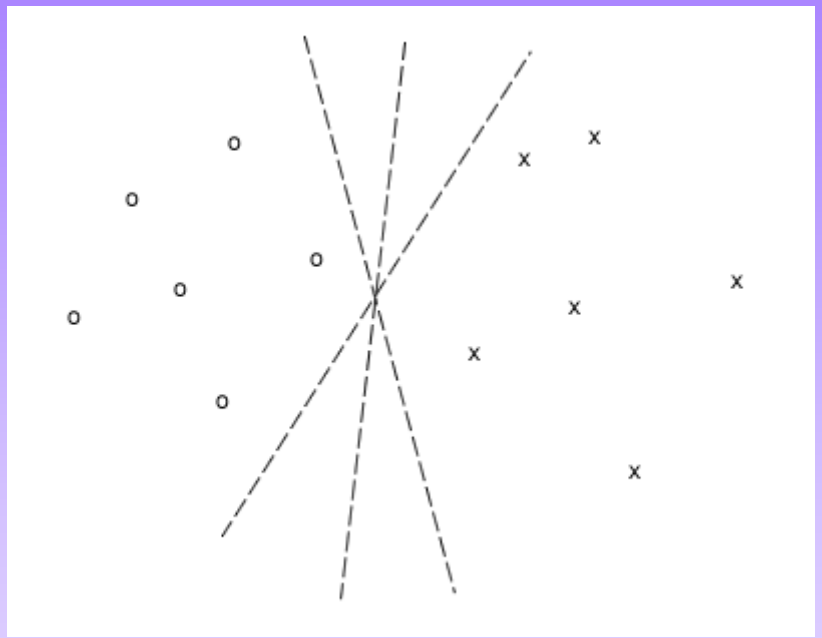


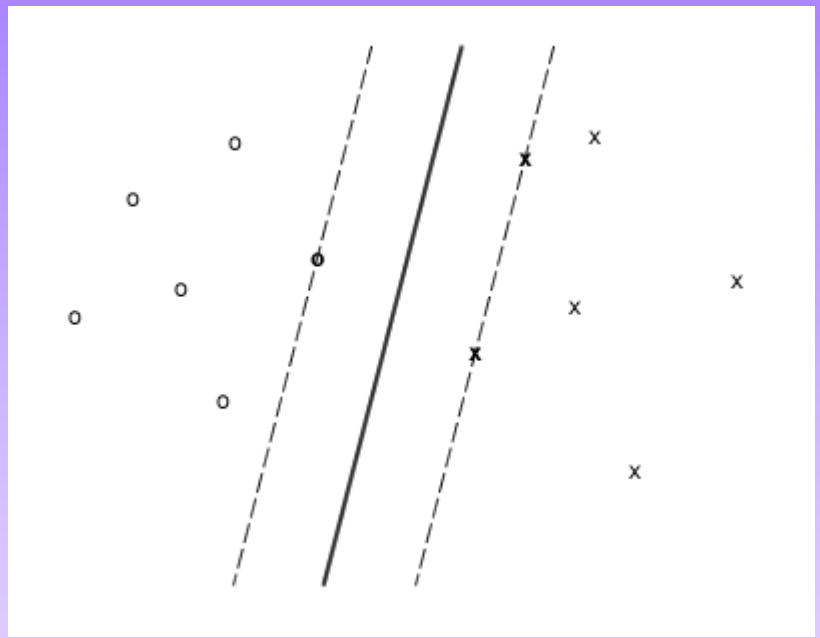
- Pick random weight values
- Run input variable and observe output
- Algorithm gets more complex with multiple hidden layers

Support Vector Machines

- Support Vector Machines are two state classifier - ‘Yes’ and ‘No’
- Classifier finds maximal separation between the two states
- Uses linear model to separate non-linear class boundaries

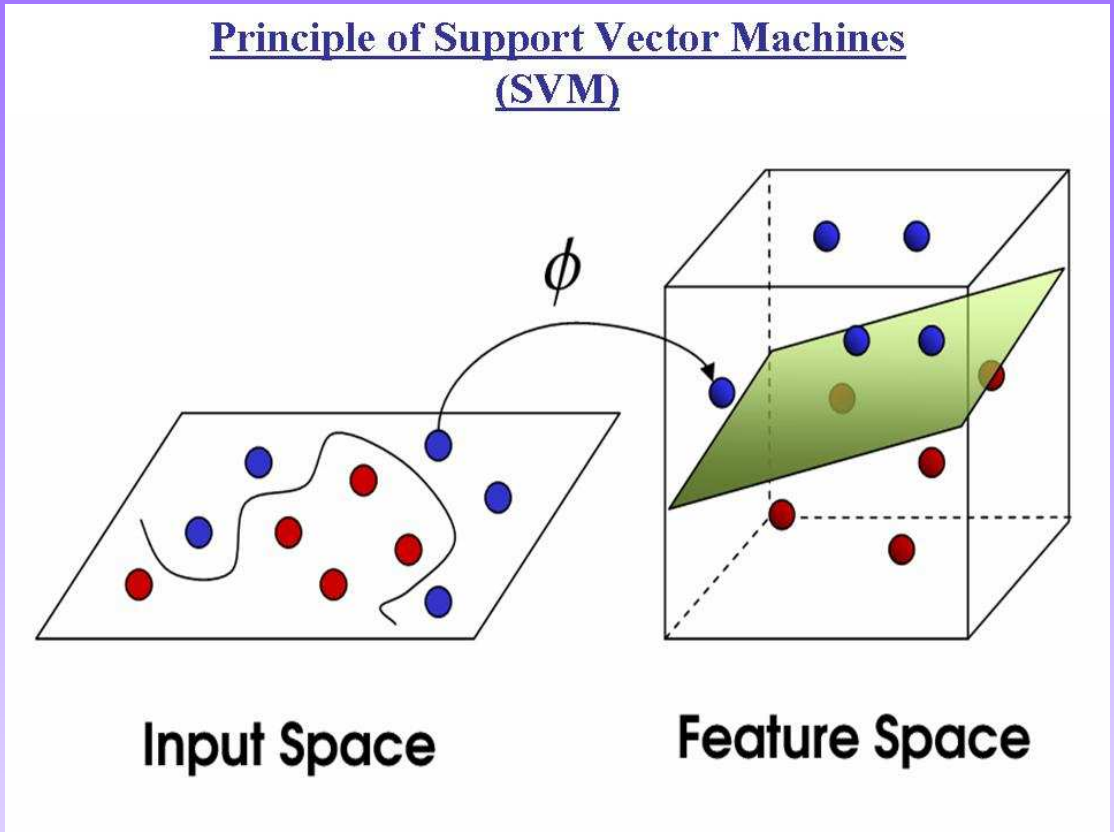
Linear Separation





- The margins of the hyperplane touch a limited number of special points - ‘ Support Vectors’.

Support Vector Machines



Tranforms non-linear data into abstract ‘feature space’

Support Vector Machines

- An SVM will construct a separating hyperplane in that space, one which maximizes the margin between the two data sets.
- To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are "pushed up against" the two data sets.
- Good separation is achieved by the hyperplane that has the largest distance to the neighbouring datapoints of both classes,
- In general, the larger the margin, the lower the generalization error of the classifier

Hyperparameter Tuning

- The performance of an SVM can improve dramatically through tuning
- Tuning of is usually done by minimizing an estimate of generalization error
- This quantifies the amount of violation of the hyperplane allowed by the learning sample
- Increasing the penalty punishes violations more severely, forcing the hyperplane to separate the learning sample (and thus probably producing overfitting).

Limits of Classification

- Requires predefined classes
- “Black box” approach
- Limited comprehensibility

Microarray Data

- For most dataset, if n is large, the training and testing procedure can be done with a simple splitting.
- For microarray data, the sample sizes n is usually very small
- Serious problems when estimating prediction accuracy and when constructing a prediction rule based on the available data

CMA package

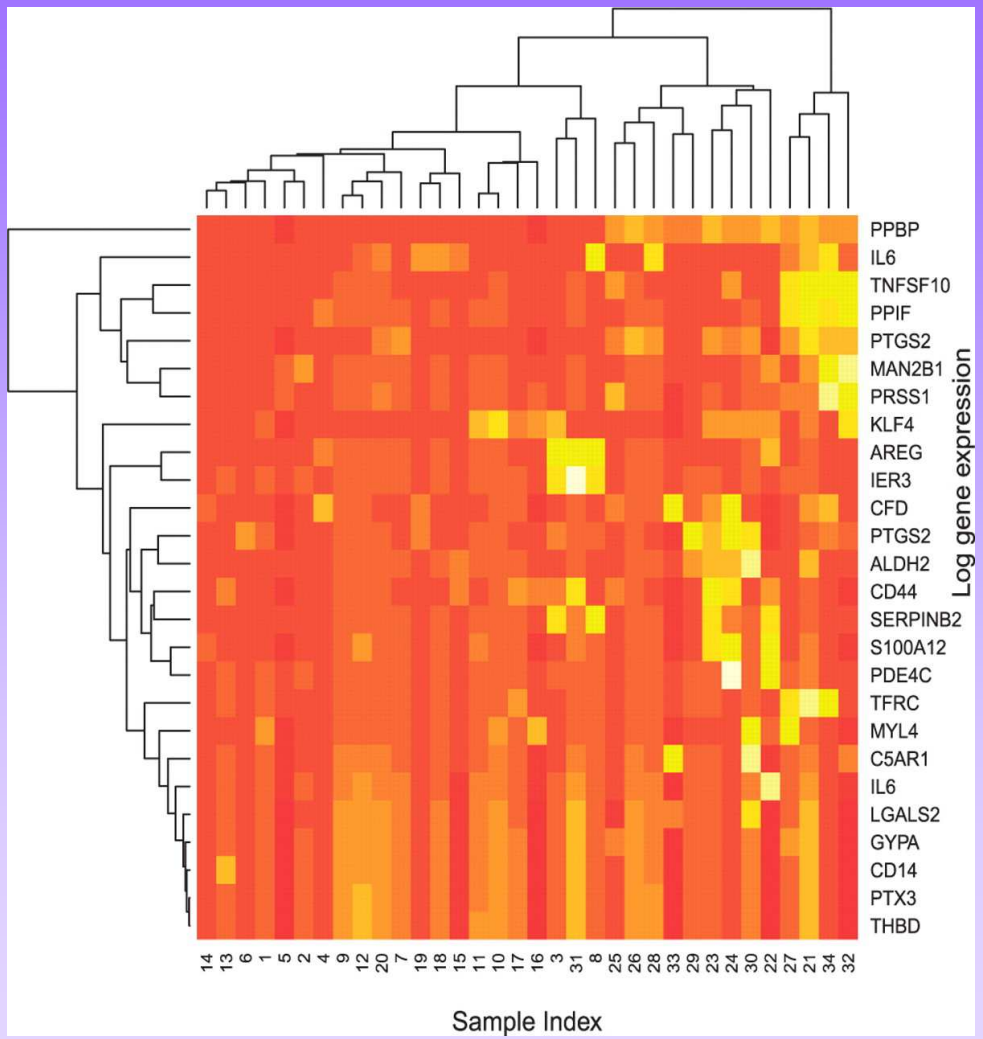
- CMA applies machine learning techniques to microarray data
- The program splits the dataset into parts L (learning/training sample) and T (test sample)
- Model selection done only with L and evaluating the resulting decision function f (·) only on T

Golub Data Set (Golub et al. 1999)

- The sample consists of 38 observations in total, from which 27 belong to class 0 (acute lymphoblastic leukemia) and 11 to class 1 (acute myeloid leukemia).

- The aim of the analysis is to identify genes whose expression discriminate acute lymphoblastic leukemia (ALL) patients form acute myeloid leukemia (AML) patients.

Heatmap of Golub dataset



Dendrogram of the clustering on the observations (on top)

Dendrogram of the clustering on the selected genes (on left-hand side)

Stingo F C , Vannucci M Bioinformatics 2011;27:495-501

Techniques

- Loading the Data
- Generate Learning Sets
- Apply Hyperparameter tuning
- Run the classifier