Exercises

1. Calculate Z_{i3} given X_i = AATGCAT. What (position) *i* gives the maximum Z_i value.

	0	1	2	3
A	.3	0.2	0.1	0.85
C	.2	0.1	0.7	0.05
G	.2	0.1	0.1	0.05
T	.3	0.6	0.1	0.05

Z_{i1} = .2*.1*.05 * .2^2 * .3^2 = 3.2e-6

Z_{i2} = .3* .2 * .1 * .05 * .3^2 * .2 = 5.4e-6

Z_{i3} = .3*.3*.6 * .1 * .05 * .3^2 = 2.43e-5

Z_{i4} = .3*.3*.3*.1 * .7 * .85 * .3 = 4.8e-4

Z_{i5} = .3*.3*.3*.2*.1 * .1 * .05 = 2.7e-6

Zsum = 5.18e-4

Z_{i3} /Zsum= 0.0469

Z_{i4} is max

2. What is the probability of T in 3rd position of motif ($p_{T,3}$) in terms of Z_{ij} given X_i = AATGCAT

$p_{T,3} = (Z_{i1} + Z_{i5})/Zsum = .012$ or $(Z_{i1} + Z_{i5} + 1)/(Zsum+4)$ with pseudo

3. The EM algorithm may not find the best motif model for a collection of sequences. Give two reasons why. local maxima, initial starting conditions, more than one motif, incorrect model (zoops vs oops)

4. Why is random sampling from probabilities for the motif position in Gibb's sampling better than taking the most likely position for the motif? better avoidance of local maxima

5. Phylogenetic footprinting: a) requires orthologous genes, b) requires conserved regulatory sequences. Answer T/F for each. a) true b) true

6. Which site is expected to have the slowest substitution rate in the binding site model given in question one. position 3

Exercises

7) In EM for a motif model. The E step calculates what probability given what?

The probability of a motif at each position in each sequence, given PWM or motif model

8) In EM for a motif model. The M step calculates what probability given what?

The probability of a motif model (a,g,c,t at each position) given the Z values, or the probabilities of a motif at each position in each sequence

9) Binding site turnover: explains divergence in sequence without divergence in function [T/F], is less common as divergence time increases [T/F], assumes gene expression/regulation changes between species [T/F], assumes chance gain and loss of redundant sites [T/F]

Today’s objectives

- Introduction to machine learning
- Types of algorithms
- Biological examples of machine learning

    - Gene expression analysis

- Hierarchical and Kmeans clustering

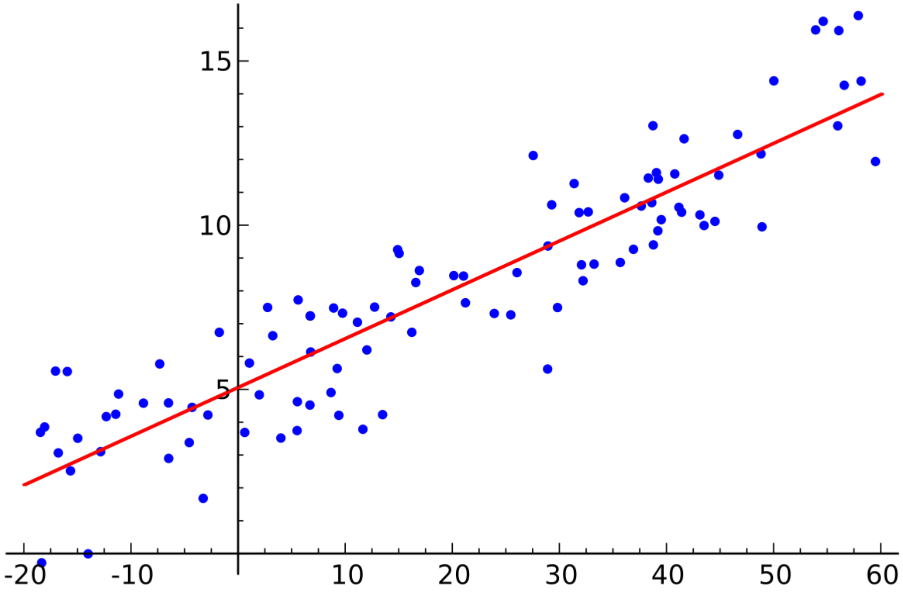
What is Machine Learning

Machine learning teaches computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed.

The construction of algorithms that can learn from and make predictions on data

Machine learning algorithms overcome strictly static program instructions by making data-driven predictions or decisions through building a model from sample inputs

Example linear regression

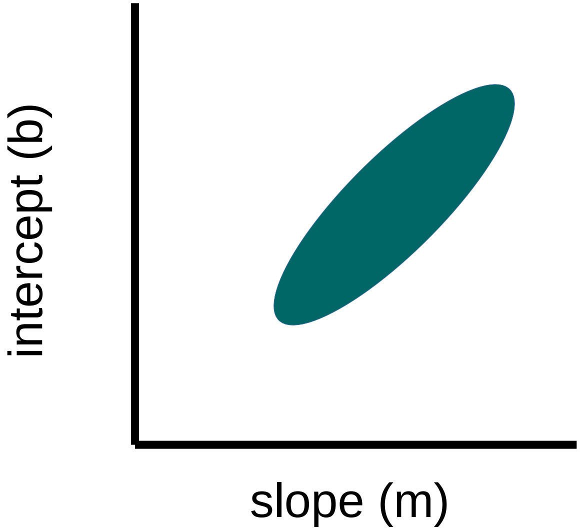


Linear model:

$$y = mx + b$$

- points are data
- line = fitted model
- m and b are 'learned' from the data
- algorithm, find m and b that:

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$



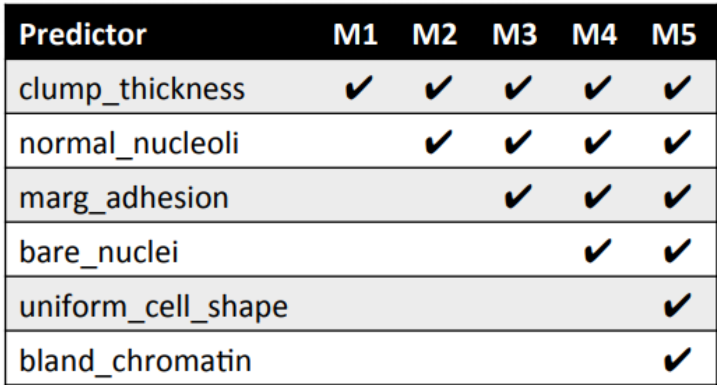
We learn them from the data
Model: given x, predict y

Predicting malignancy using logistic regression

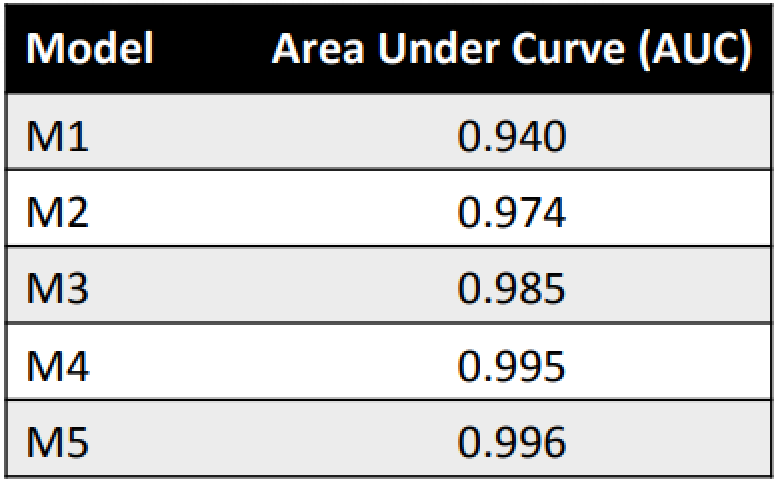
Sample	Outcome	Clump thickness	Normal nucleoli	Marginal adhesion	Bare nuclei	Uniform cell shape
1	benign	4	0	1	0	1
2	malignant	2	1	0	1	0
n	malignant	4	1	1	1	0

predicted outcome: p(outcome) ~ 0.2*Clump + 2*nulceoli – 0.1*adhesion ...

Model (variable) selection, which variables to use and coefficients

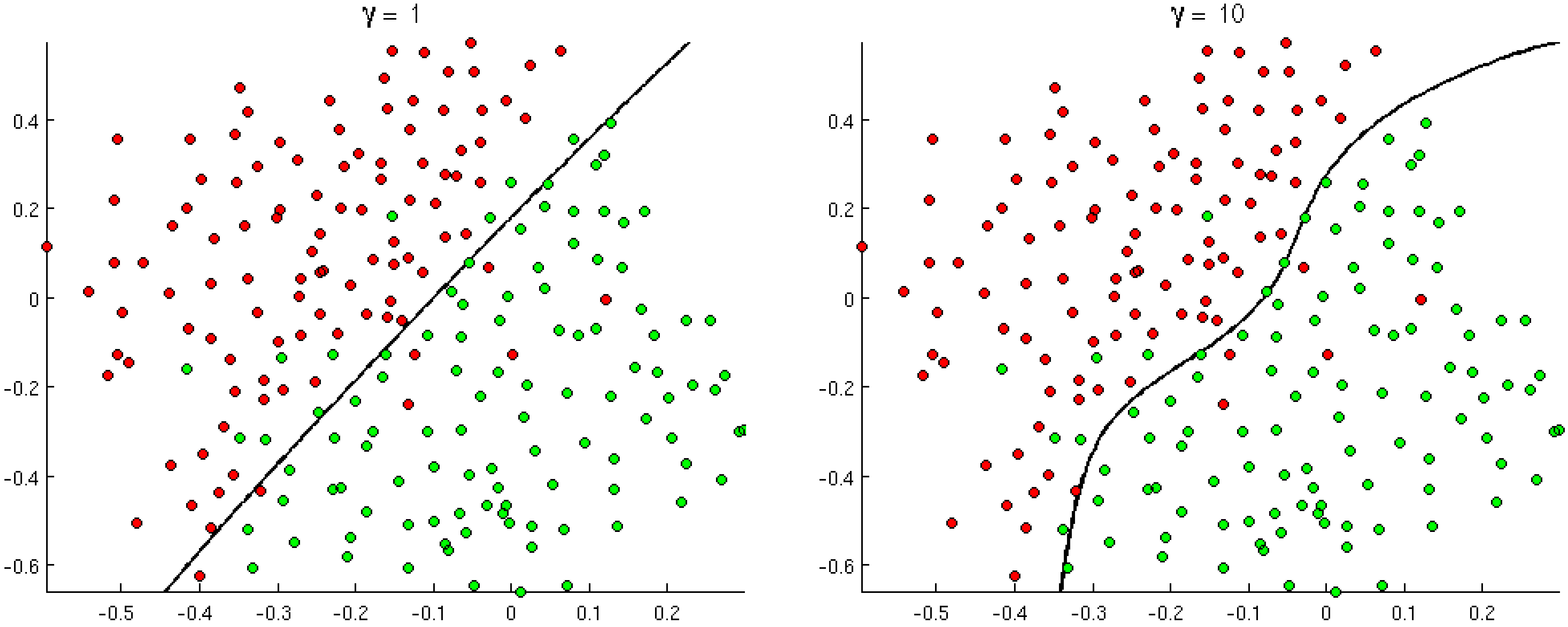


Predictor	M1	M2	M3	M4	M5
clump_thickness	✓	✓	✓	✓	✓
normal_nucleoli		✓	✓	✓	✓
marg_adhesion			✓	✓	✓
bare_nuclei				✓	✓
uniform_cell_shape					✓
bland_chromatin					✓



Model	Area Under Curve (AUC)
M1	0.940
M2	0.974
M3	0.985
M4	0.995
M5	0.996

Example: find a function to distinguish red and green data
Classification problem



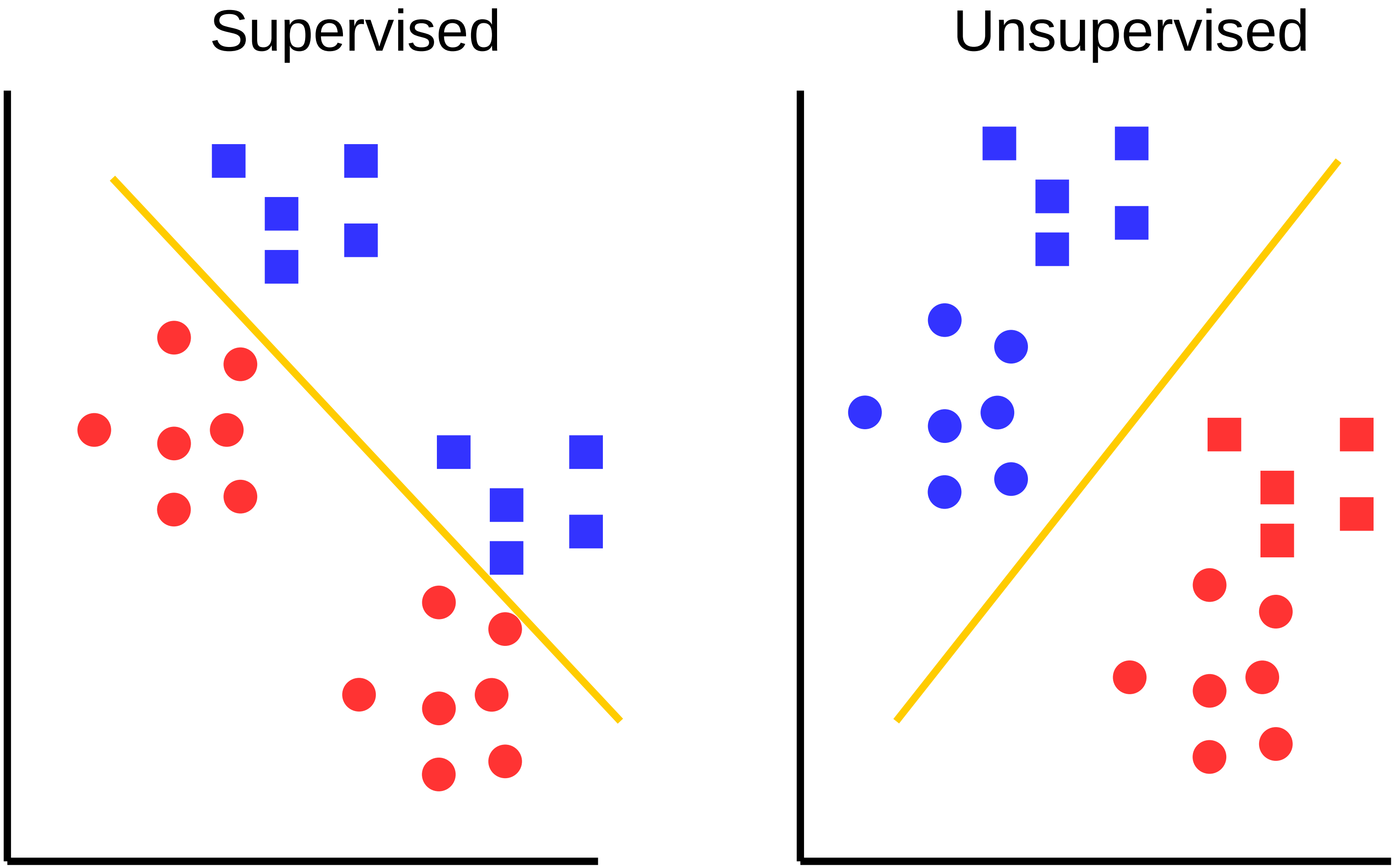
Problem: model selection, potentially many predictive variables
Solution: machine learning, algorithm builds (learns) variable function

Types of Learning (input)

Supervised learning: The computer is presented with example inputs and their desired outputs, and the goal is to learn a general rule that maps inputs to outputs.

Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data: e.g. groups) or a means towards an end (feature learning: e.g. discover low-dimensional features that capture some structure underlying the high-dimensional input data).

Supervised and unsupervised

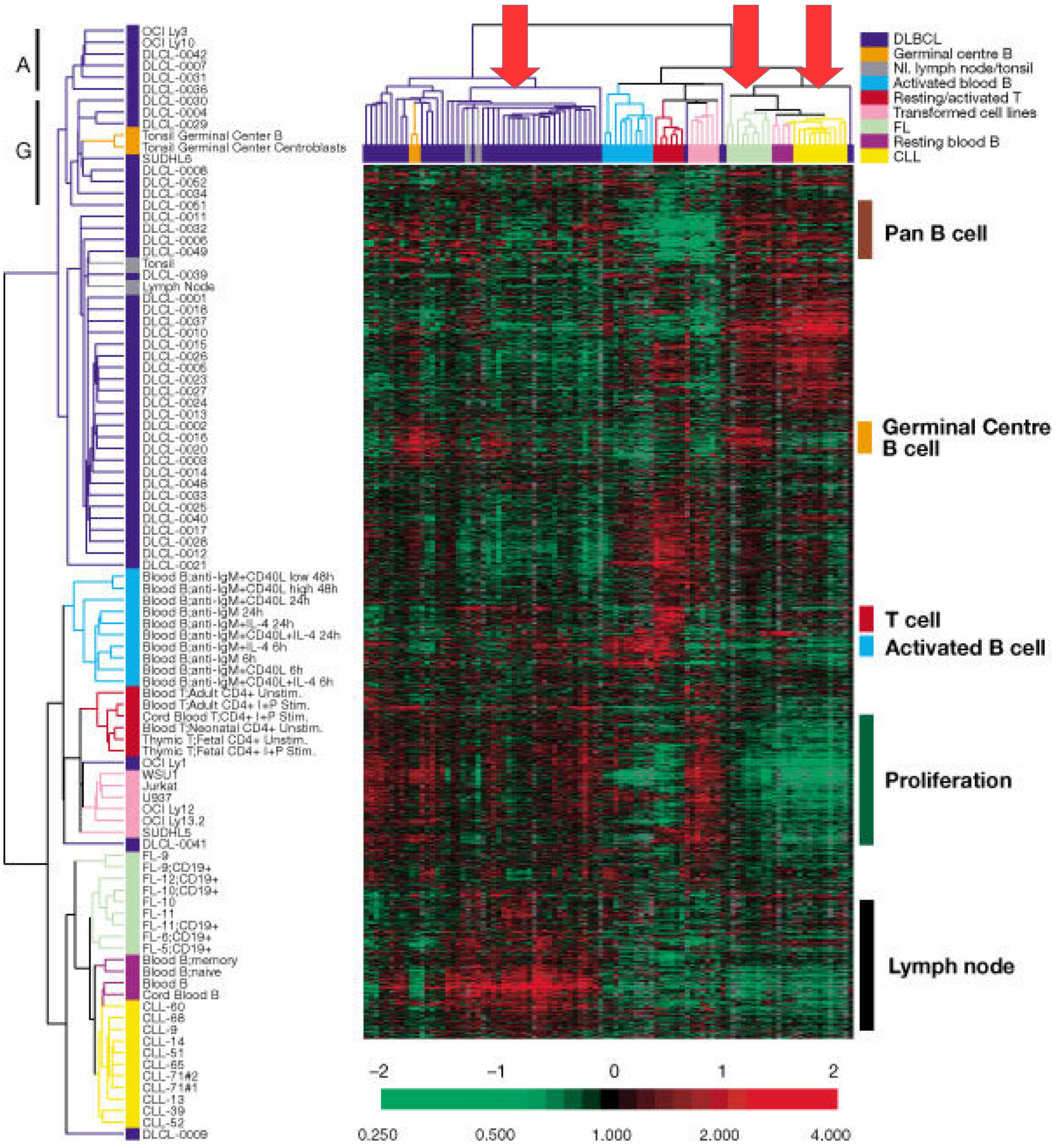


Supervised

Unsupervised

Shapes = labels
Colors = predicted groups

Big Data



- 98 normal and malignant lymocyte samples (columns)
- 3,186 genes (rows)
- heatmap of gene expression
- DLBCL, FL and CLL are malignant

Machine learning can be used to:

- visualize data and relationships (cluster)
- make (supervised model) predictions (malignant)
- identify (unsupervised model) groups and their relationships

Types of learning (output)

Categorization of machine learning tasks depend on the desired output:

Classification: the model that assigns one or more classifications to the inputs. Tumor data (input), malignant or benign (output)

Regression: the outputs are continuous rather than discrete. Smoking, age (input), life expectancy (output)

Clustering: a set of inputs is to be divided into groups/relationships, but unlike classification, the groups are not known beforehand but learned.

Machine Learning Algorithms *(sample)*

Unsupervised | **Supervised**

Continuous

Unsupervised

- Clustering & Dimensionality Reduction
  - SVD
  - PCA
  - K-means

Centroids continuous
Clusters Categorical

Supervised

- Regression
  - Linear
  - Polynomial
- Decision Trees
- Random Forests

Categorical

Unsupervised

- Association Analysis
  - Apriori
  - FP-Growth
- Hidden Markov Model

Supervised

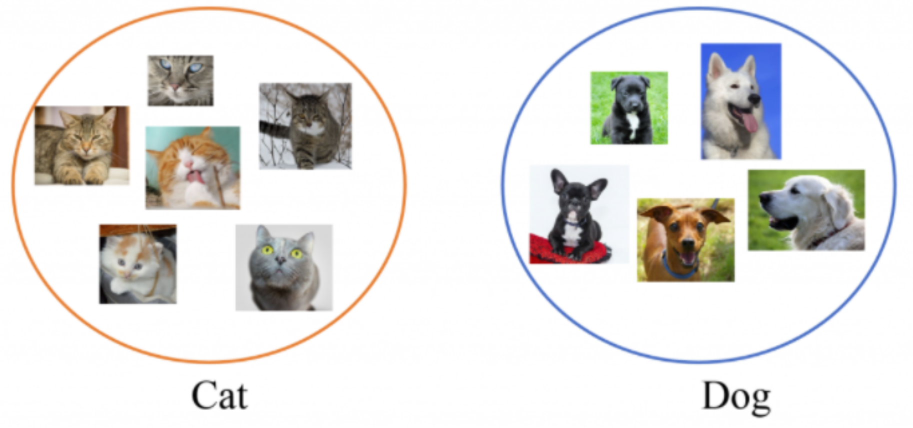
- Classification
  - KNN
  - Trees
  - Logistic Regression
  - Naive-Bayes
  - SVM

Unsupervised learning finds hidden patterns or intrinsic structures in data.

Supervised machine learning builds a model to generate predictions for the response to new data.

Overlap between supervised and unsupervised learning

Image recognition
Supervised: place a label on an image
Unsupervised: group an image with other similar images (no label)



HMM
Supervised: Trained
Unsupervised: Baum-Welch
Application of an HMM: unsupervised

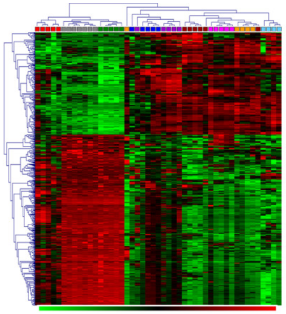
Unsupervised vs Supervised

Are the labels given/present?

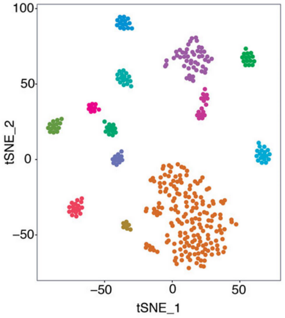
What is Clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

1) Given gene expression data, group by similarity



2) Given cell features, classify them into groups



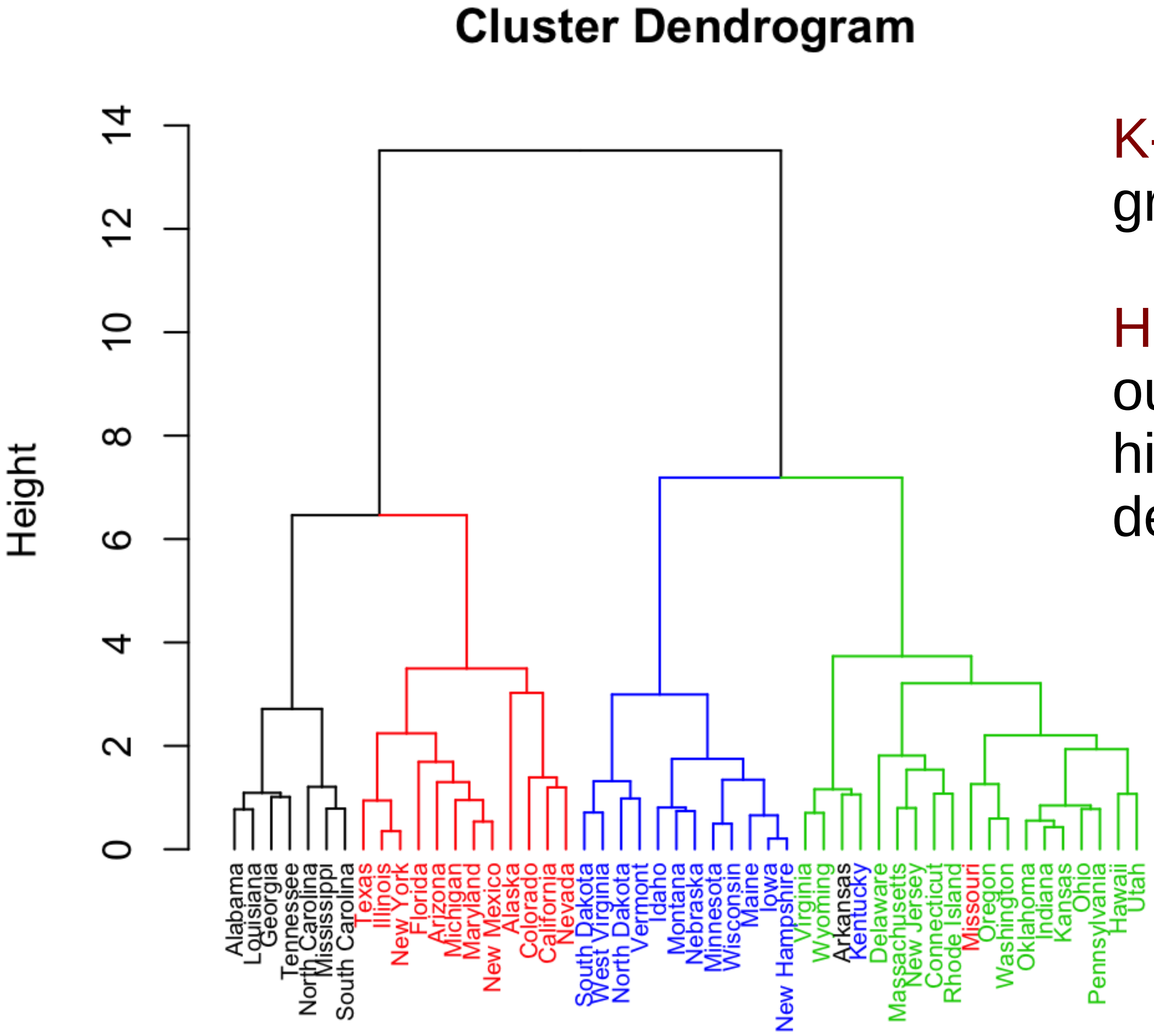
Clustering algorithms

Hierarchical clustering seeks to build a hierarchy of clusters.

- Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.
- time complexity of O(n^3) and requires O(n^2) memory (agglomerative)

Centroid-based clustering (K-means clustering), clusters are represented by a central vector. When the number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

Hierarchical vs. K-means



K-means outputs k=4 groups shown by colors

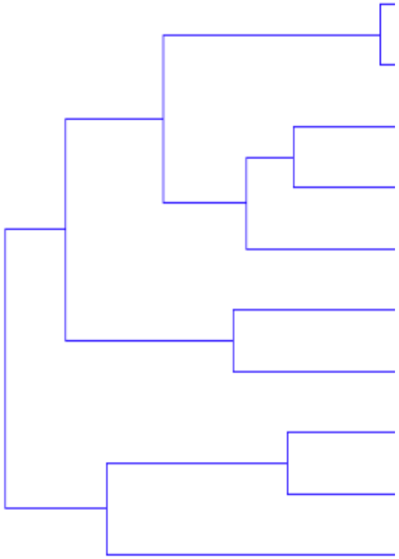
Hierarchical clustering outputs nested cluster hierarchy shown by a dendrogram

dendrogram (from Greek dendro "tree" and gramma "drawing") is a tree diagram

Clustering input

Input is a matrix with row labels, column labels, with the remainder filled in with numeric values.

Clustering can be done on Genes (below), or Samples, or both.



Gene	Sample 1	Sample 2	Sample 3	Sample 4
Gene1				
Gene2				
Gene3				
Gene4				

Normalization

Normalization is used to generate equal weighting of data during the clustering process.

Which pair of genes are most similar?

- Depends on the sample (samples 2-4 vs sample 1)
- What is the metric (sum of the sample differences)
- Should we weight Sample 1 more than Sample 2, 3, 4?

Gene	Sample 1	Sample 2	Sample 3	Sample 4
Gene1	-10	1	1	1
Gene2	-5	0	0	0
Gene3	2	0	0	0
Gene4	3	1	1	1

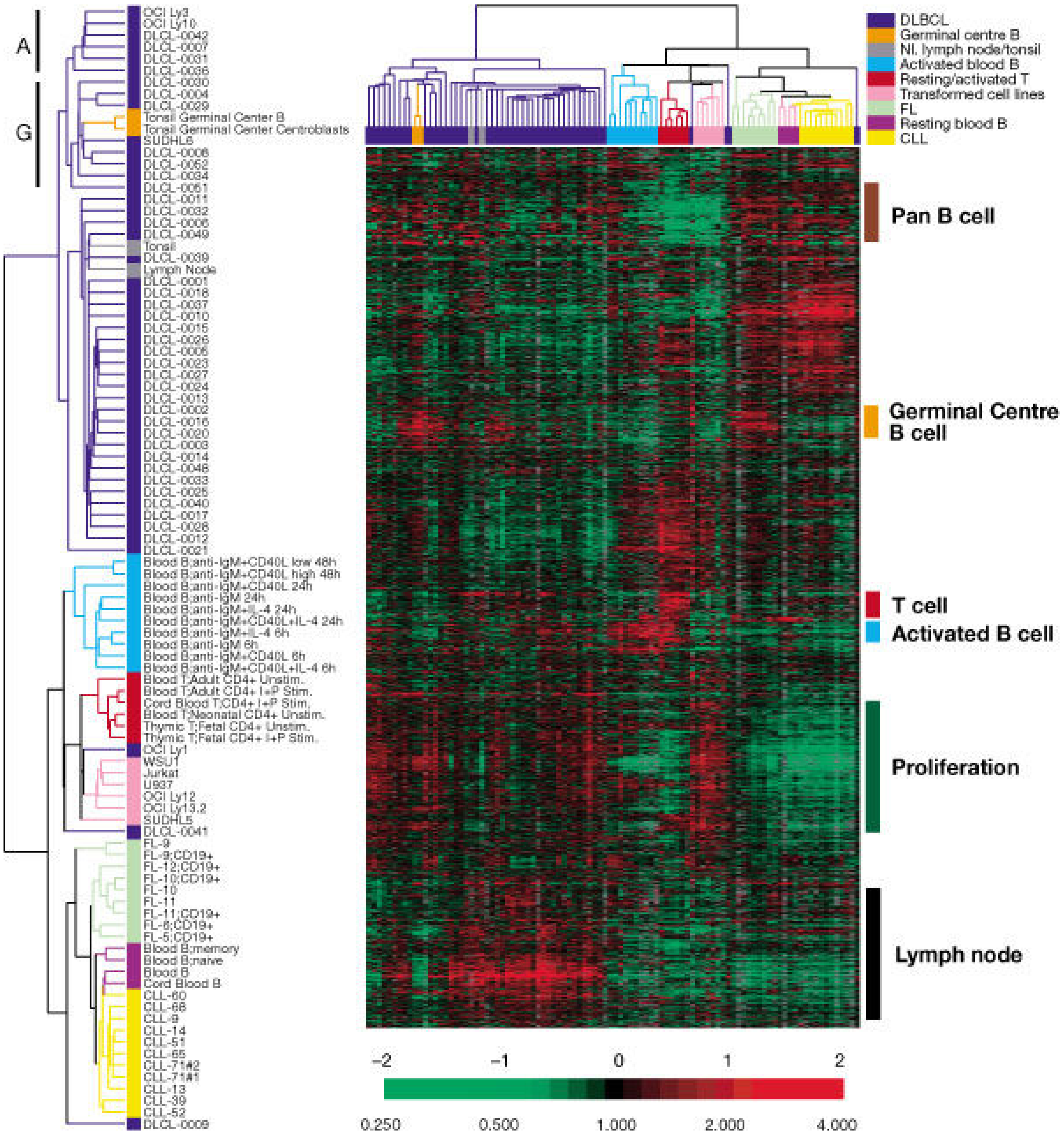
Normalization

Normalization = standard score

X' = (X-u)/σ (column normalized to cluster genes)
u = mean column
σ = standard deviation column
NB rows are often just centered (X-u) but not scaled by variance

Gene	Sample 1	Sample 2	Sample 3	Sample 4
Gene1	-1.22	0.866	0.866	0.866
Gene2	-0.407	-0.866	-0.866	-0.866
Gene3	0.733	-0.866	-0.866	-0.866
Gene4	0.896	0.866	0.866	0.866

Clustering output



- 98 normal and malignant lymocyte samples (columns)
- 3,186 genes (rows)
- DLBCL, FL and CLL are malignant

- Visualize which genes are up/down in which samples?
- Which samples are most closely related?
- Are there motifs associated with certain groups of co-regulated genes?

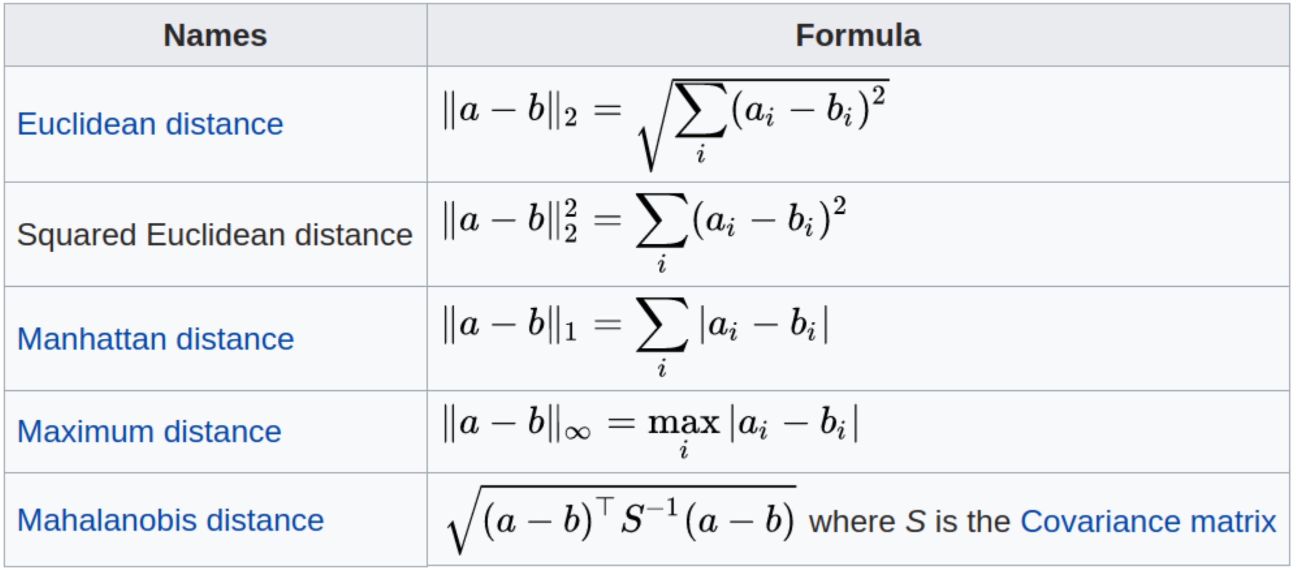
Hierarchical clustering approach

In order to decide which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of observations is required.

Hierarchical clustering uses a distance metric, a measure of distance between pairs of observations, and a linkage criterion, which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.

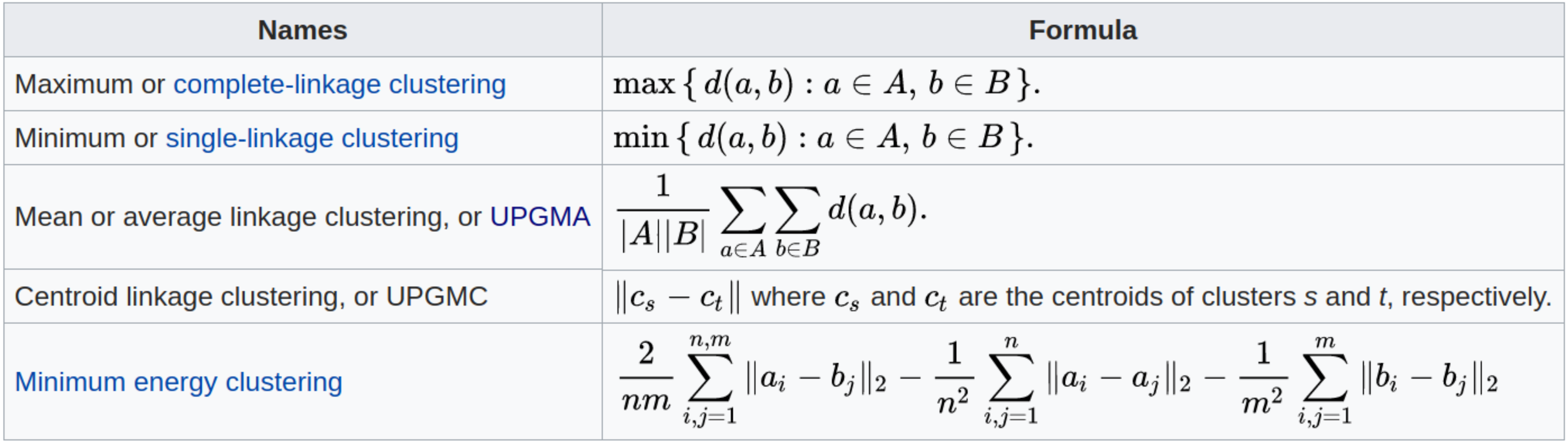
Hierarchical clustering options

Both the distance metric and a linkage criterion impact the clustering results and should be chosen based on the goals of clustering.



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

Also: correlation



Names	Formula
Maximum or complete-linkage clustering	$\max\{\, d(a,b) : a \in A,\, b \in B \,\}.$
Minimum or single-linkage clustering	$\min\{\, d(a,b) : a \in A,\, b \in B \,\}.$
Mean or average linkage clustering, or UPGMA	$\frac{1}{\|A\|\|B\|}\sum_{a\in A}\sum_{b\in B} d(a,b).$
Centroid linkage clustering, or UPGMC	$\|c_s - c_t\|$ where c_s and c_t are the centroids of clusters *s* and *t*, respectively.
Minimum energy clustering	$\frac{2}{nm}\sum_{i,j=1}^{n,m}\|a_i - b_j\|_2 - \frac{1}{n^2}\sum_{i,j=1}^{n}\|a_i - a_j\|_2 - \frac{1}{m^2}\sum_{i,j=1}^{m}\|b_i - b_j\|_2$

Algorithm

Basic algorithm (agglomerative)

- Compute the distance between each pair of input data
- Let each data be a cluster
- Repeat until only one cluster
  - Merge the two closest clusters
  - Update the distance matrix

Distance matrix calculation

Gene	Sample 1	Sample 2	Sample 3
A	1	5	3
B	3	2	5
C	5	1	4
D	4	1	4

Calculate the distance matrix for genes:

1) Normalize the data with a standard score so each sample equally weighted (specifically for clustering)

$$X_{ij} = \frac{(X_i - \mu_j)}{\sigma_j}$$

where u_j is the column mean
σ_j is the column standard deviation

2) Calculate the distance using the Euclidean norm:

$$\|x - y\|_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Normalized and distance

Normalized	Sample 1	Sample 2	Sample 3
A	-1.32	1.45	-1.22
B	-0.15	-0.13	1.22
C	1.02	-0.66	0.00
D	0.44	-0.66	0.00

	A	B	C	D
A	0.00	3.14	3.38	3.14
B	3.14	0.00	1.77	1.46
C	3.38	1.77	0.00	0.59
D	3.14	1.46	0.59	0.00

$$\|x-y\|_2=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

D(AB) = sqrt(
(-1.32 +0.15)^2 +
(1.45+0.13)^2 +
(-1.22-1.22)^2) = 3.14

Merge closest clusters and update

	A	B	C	D
A	0.00	3.14	3.38	3.14
B	3.14	0.00	1.77	1.46
C	3.38	1.77	0.00	0.59
D	3.14	1.46	0.59	0.00

	A	B	CD
A	0.00	3.14	3.26
B	3.14	0.00	1.62
CD	3.26	1.62	0.00

$d_{CD,A}$ = (3.38 + 3.14)/2 = 3.26

$d_{CD,B}$ = (1.77 + 1.46)/2 = 1.62

Linkage

$$\frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} d(a,b)$$

Where |A| is the cardinality of A (number of elements)

Joining A and B with new cluster X

$$d_{AB,X}=\frac{|A|\cdot d_{A,X}+|B|\cdot d_{B,X}}{|A|+|B|}$$

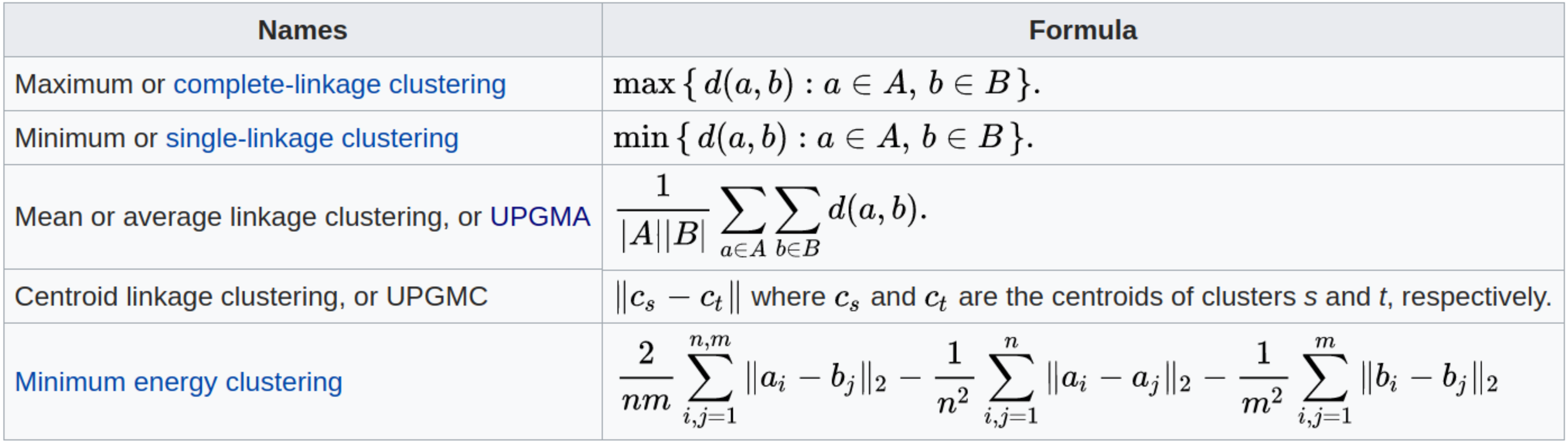
Linkage

	A	B	CD
A	0.00	3.14	3.26
B	3.14	0.00	1.62
CD	3.26	1.62	0.00

	A	BCD
A	0.00	3.22
BCD	3.22	0.00

$$d_{AB,X} = \frac{|A| \cdot d_{A,X} + |B| \cdot d_{B,X}}{|A| + |B|}$$

$d_{BCD,A}$ = (3.14*1 + 3.26*2)/3 = 3.22



Names	Formula
Maximum or complete-linkage clustering	$\max\{\,d(a,b) : a \in A,\ b \in B\,\}.$
Minimum or single-linkage clustering	$\min\{\,d(a,b) : a \in A,\ b \in B\,\}.$
Mean or average linkage clustering, or UPGMA	$\frac{1}{\|A\|\|B\|} \sum_{a \in A} \sum_{b \in B} d(a,b).$
Centroid linkage clustering, or UPGMC	$\|c_s - c_t\|$ where c_s and c_t are the centroids of clusters *s* and *t*, respectively.
Minimum energy clustering	$\frac{2}{nm} \sum_{i,j=1}^{n,m} \|a_i - b_j\|_2 - \frac{1}{n^2} \sum_{i,j=1}^{n} \|a_i - a_j\|_2 - \frac{1}{m^2} \sum_{i,j=1}^{m} \|b_i - b_j\|_2$

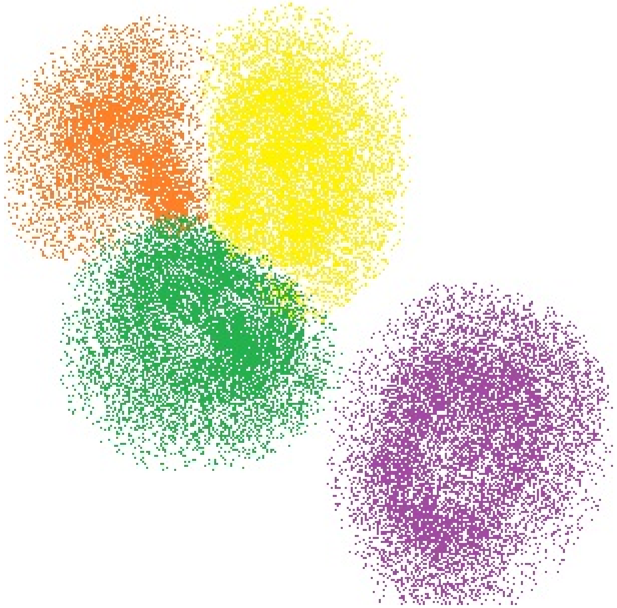
Kmeans Clustering

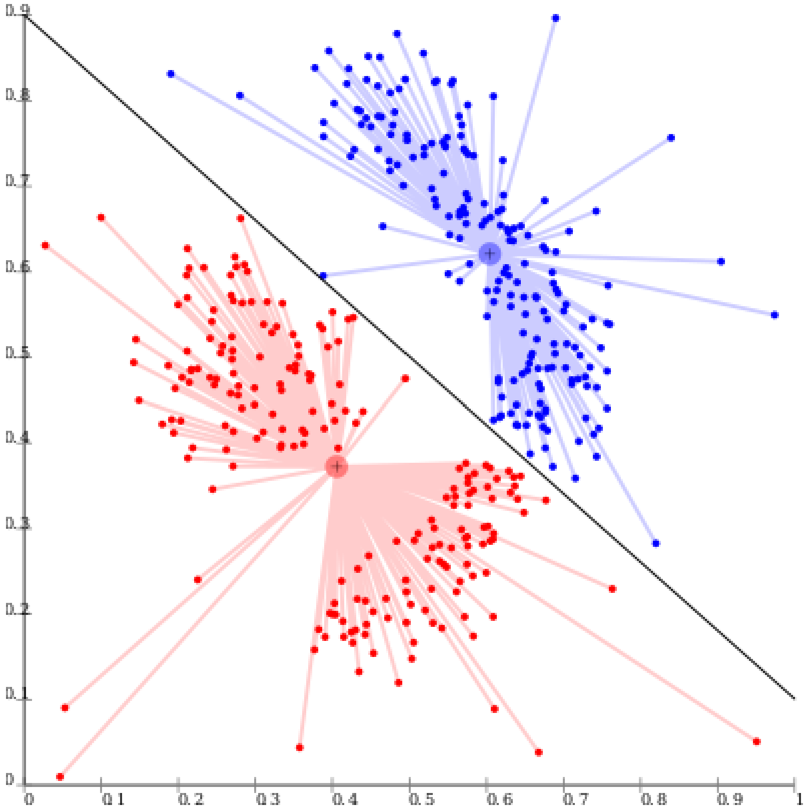
K-means clustering: partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

Given a set of observations (x_1, x_2, …, x_n), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k (≤ n)
sets S = {S_1, S_2, …, S_k} so as to minimize the within-cluster sum of squares (WCSS) (i.e. variance).

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2 = \underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

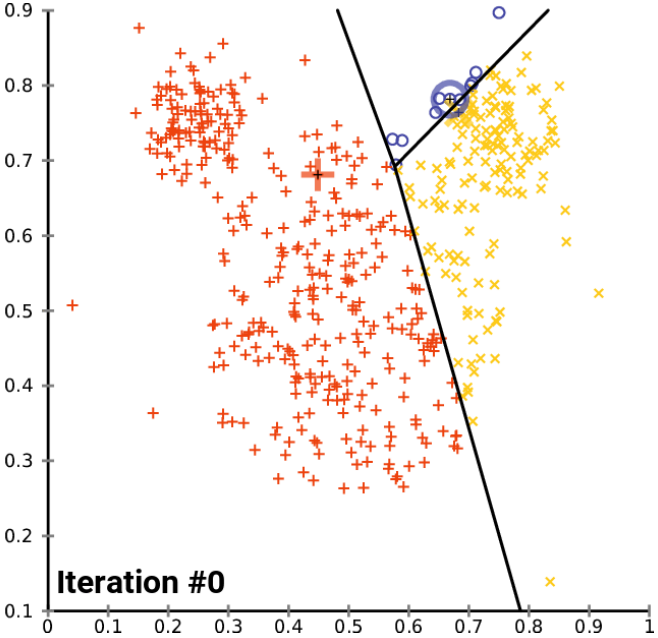
Kmeans examples





K-means algorithm

K-means clustering (K=3) iterations



The "assignment" step is also referred to as expectation step, the "update step" as maximization step, making this algorithm a variant of the generalized expectation-maximization algorithm.

Given an initial set of k means $m_1^{(1)}$, …,$m_k^{(1)}$, the algorithm proceeds by alternating between two steps:

Initialization: Random Partition randomly assigns a cluster to each observation and then proceeds to the update step

Assignment step: Assign each observation to the cluster whose mean has the least squared Euclidean distance, this is intuitively the "nearest" mean.

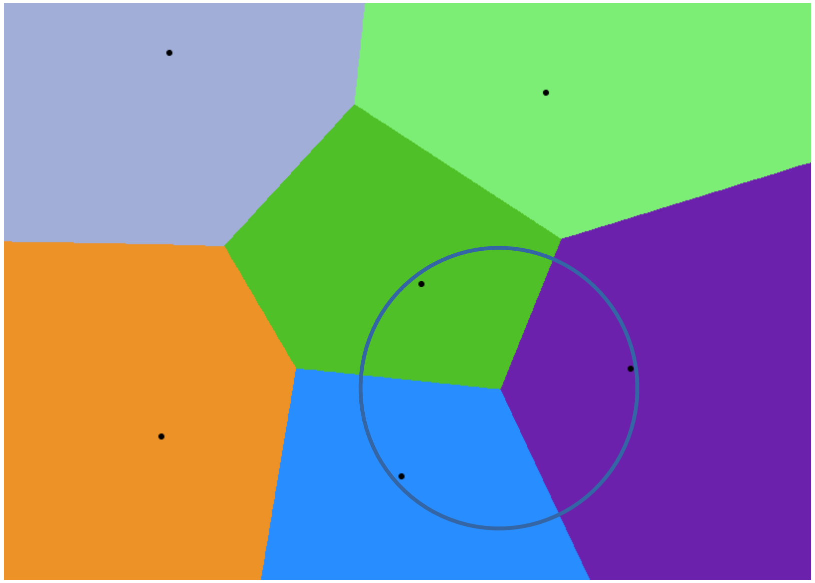
Update step: Calculate the new means to be the centroids of the observations in the new clusters. [assign or terminate]

Termination: Run until no change in assignment

By Chire - Own work, GFDL, https://commons.wikimedia.org/w/index.php?curid=59409335

K-means algorithm

Voronoi Diagram: assignment to the closest point



The "assignment" step is also referred to as expectation step, the "update step" as maximization step, making this algorithm a variant of the generalized expectation-maximization algorithm.

Given an initial set of k means $m_1^{(1)}$, …,$m_k^{(1)}$, the algorithm proceeds by alternating between two steps:

Initialization: Random Partition randomly assigns a cluster to each observation and then proceeds to the update step

Assignment step: Assign each observation to the cluster whose mean has the least squared Euclidean distance, this is intuitively the "nearest" mean.

Update step: Calculate the new means to be the centroids of the observations in the new clusters. [assign or terminate]

Termination: Run until no change in assignment

By Chire - Own work, GFDL, https://commons.wikimedia.org/w/index.php?curid=59409335

K-means algorithm

Given an initial set of k means $m_1^{(1)},\ldots,m_k^{(1)}$, the algorithm proceeds by alternating between two steps:

Assignment step: Assign each observation to the cluster whose mean has the least squared Euclidean distance, this is intuitively the "nearest" mean.

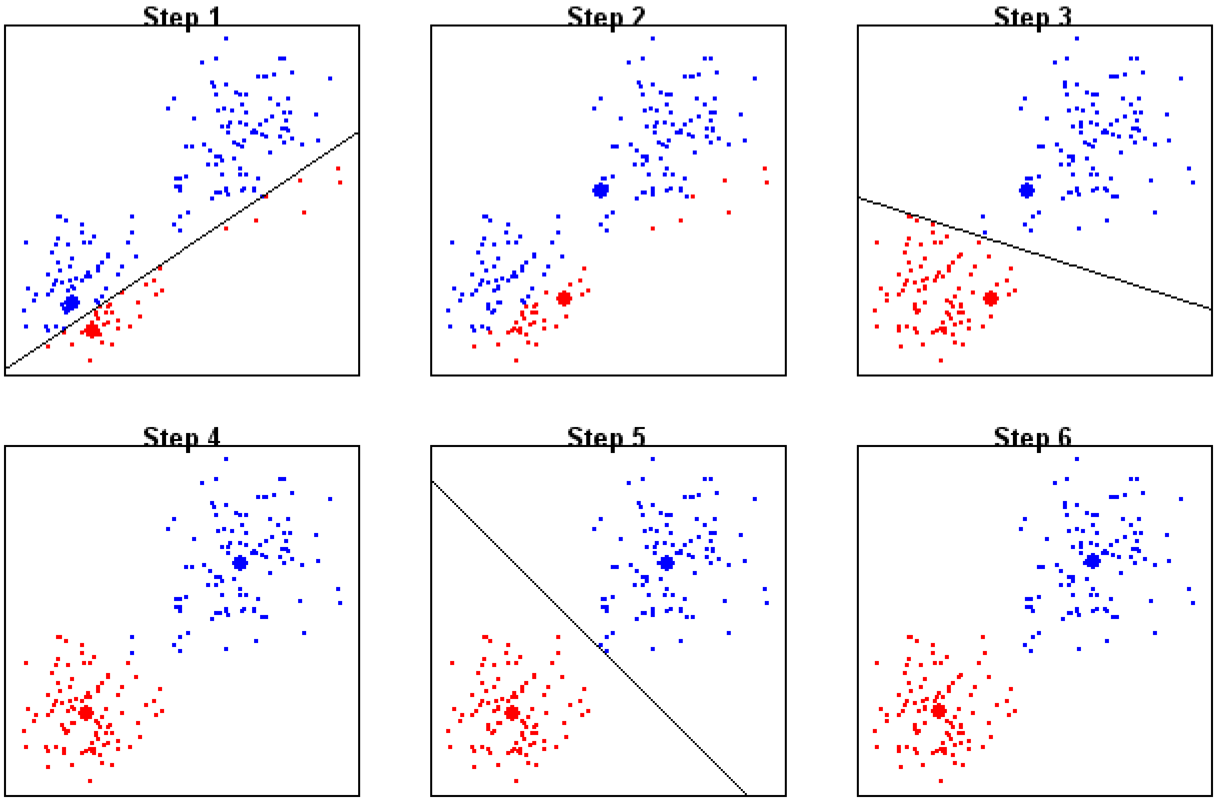
For each observation x_p, assign it to cluster S_i such that $\|x_p - m_i\|^2 \leq \|x_p - m_j\|^2$ for j = 1:k

Update step: Calculate the new means to be the centroids of the observations in the new clusters.

$$m_i = \frac{1}{|S_i|} \sum_{x_j \in S_i} x_j$$

K-Means is really just the EM (Expectation Maximization) algorithm applied to a particular naive bayes model.

Example of K-means algorithm



K-means Example

Cluster	S1	S2
1	1	5
1	3	2
1	5	1
1	4	1
2	2	5
2	5	4
2	8	2
2	5	6
Centroid 1	3.25	2.25
Centroid 2	5	4.25

Initialization (k=2):

- Randomly assign clusters, calculate centroids
- Randomly choose centroids from genes (rows), assign to cluster

Assign genes to Centroids:

- Calculate distance to each
- Pick the smaller value

Assign to clusters

Update Assignments

Cluster	S1	S2	Distance1	Distance2
1	1	5	12.625	16.5625
1	3	2	0.125	9.0625
1	5	1	4.625	10.5625
1	4	1	2.125	11.5625
1	2	5	9.125	9.5625
2	5	4	6.125	0.0625
2	8	2	22.625	14.0625
1	5	1	4.625	10.5625
Centroid 1	3.25	2.25		
Centroid 2	5	4.25		

d(Gene1,Centroid 1) = (1-3.25)^2 + (5-2.25)^2 = 12.625
d(Gene1,Centroid 2) = (1-5)^2 + (5-4.5)^2 = 16.56
Assign Gene1 to Cluster 1.

Update Centroids

Cluster	S1	S2	Distance1	Distance2
1	1	5		
1	3	2		
1	5	1		
1	4	1		
1	2	5		
2	5	4		
2	8	2		
1	5	1		
Centroid 1	3.33	2.5		
Centroid 2	6.5	3		

Centroid 1 S1 = (1+3+5+4+2+5)/6 = 3.33
Centroid 1 S2 = (5+2+1+1+5+1)/6 = 2.5

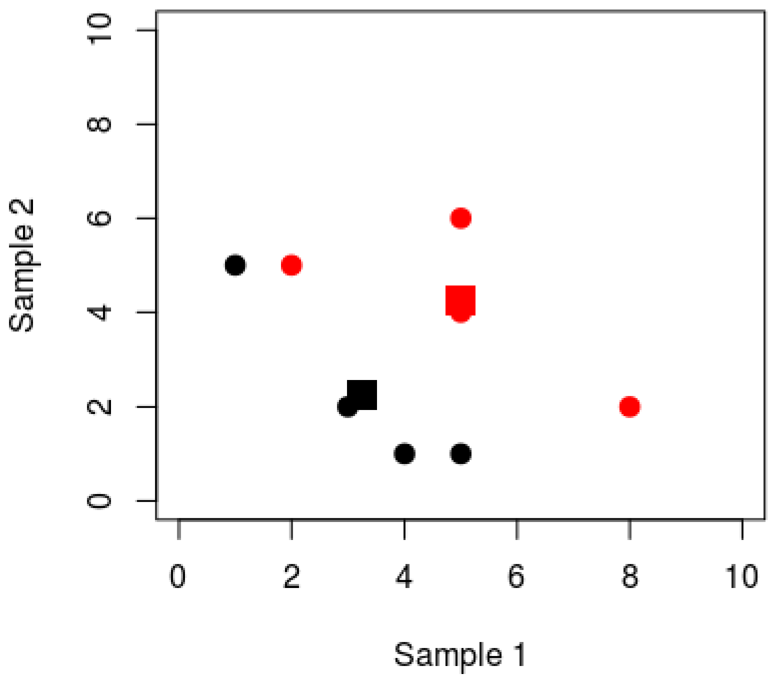
Assign to clusters

Cluster	S1	S2	Distance1	Distance2
1->1	1	5	11.69	34.25
1->1	3	2	0.36	13.25
1->1	5	1	5.03	6.25
1->1	4	1	2.69	10.25
1->1	2	5	8.03	24.25
2->2	5	4	5.03	3.25
2->2	8	2	22.03	3.25
1->1	5	1	5.03	6.25
Centroid 1	3.33	2.5		
Centroid 2	6.5	3		

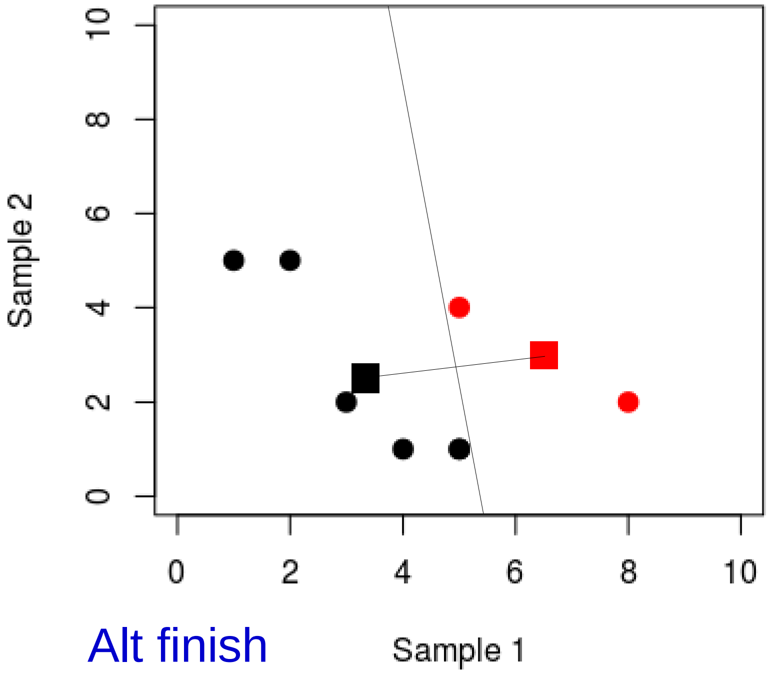
Centroid 1 S1 = (1+3+5+4+2+5)/6 = 3.33
Centroid 1 S2 = (5+2+1+1+5+1)/6 = 2.5

K-means clustering is deterministic and starting conditions can matter!

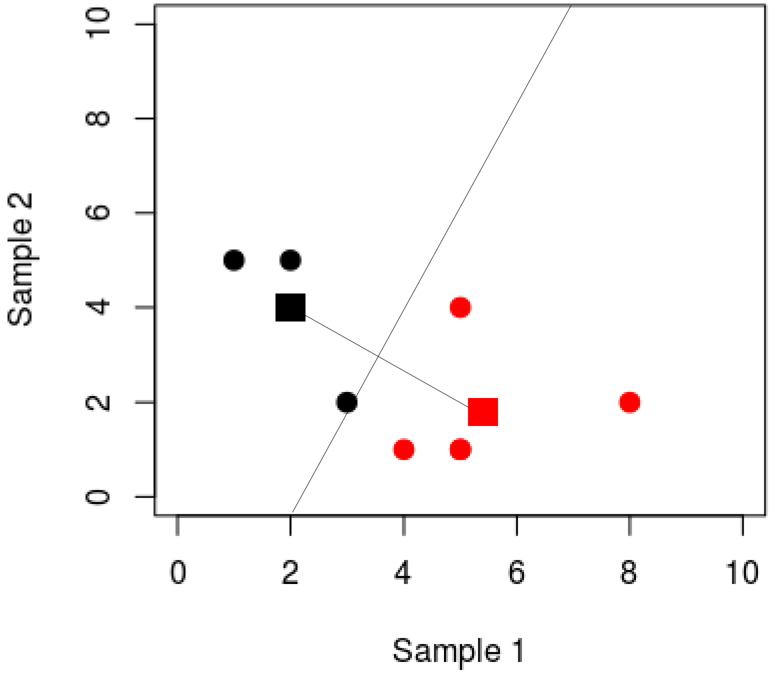
Starting conditions



Finish conditions



Alt finish



Mixture Models

A mixture model is a probabilistic model for representing the presence of subpopulations within an overall population.

Mixture models involve steps that attribute sub-population-identities (labels) to individual observations (or weights towards such sub-populations), so can be regarded as unsupervised learning

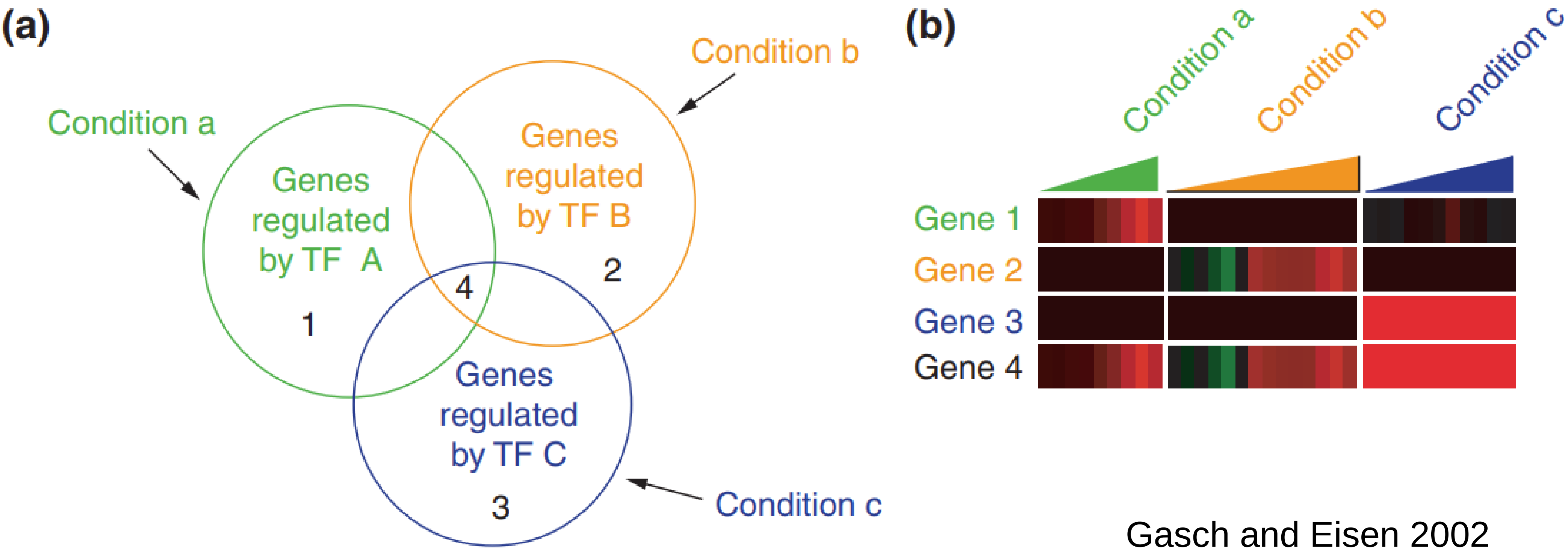
Hard labels: individuals cannot have partial membership.
Soft labels: Mixture models/fuzzy clustering

Kmeans – clusters
Motif finding – background vs motif
Population structure – a mixture of populations,
but individuals can be mixtures (admixed) of different populations
Gene expression – a mixture of transcription factor binding sites

Algorithms: EM, MCMC

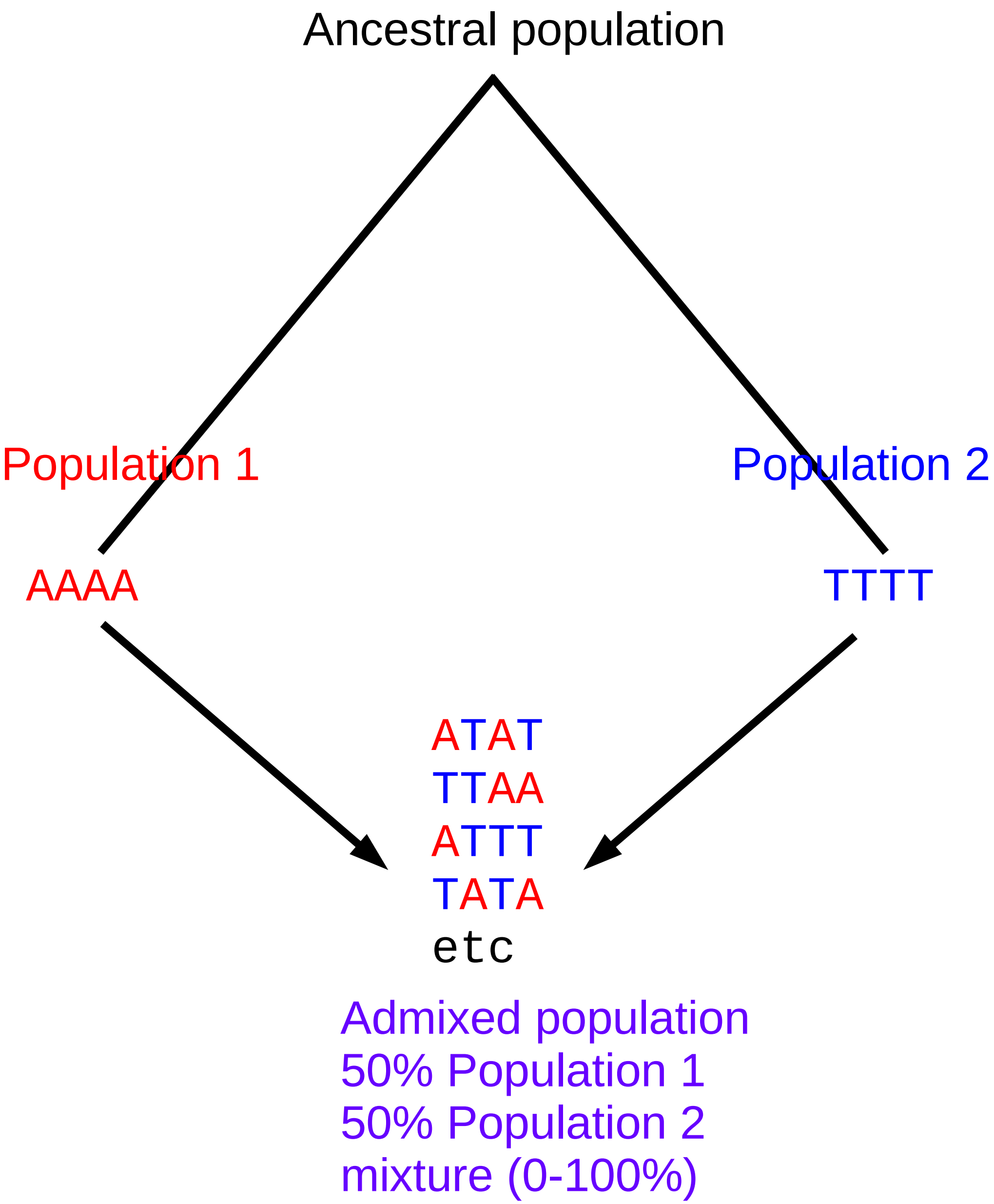
Example 1: Gene expression

Objective: cluster genes into co-regulated groups in order to find motifs
Problem: some genes share expression with multiple groups, e.g. due to multiple motifs
Solution: fuzzy or soft clustering, where genes can have membership to multiple groups



Gasch and Eisen 2002

Example 2: Admixed ancestry



Ancestral population

Population 1

Population 2

AAAA

TTTT

ATAT
TTAA
ATTT
TATA
etc

Admixed population
50% Population 1
50% Population 2
mixture (0-100%)

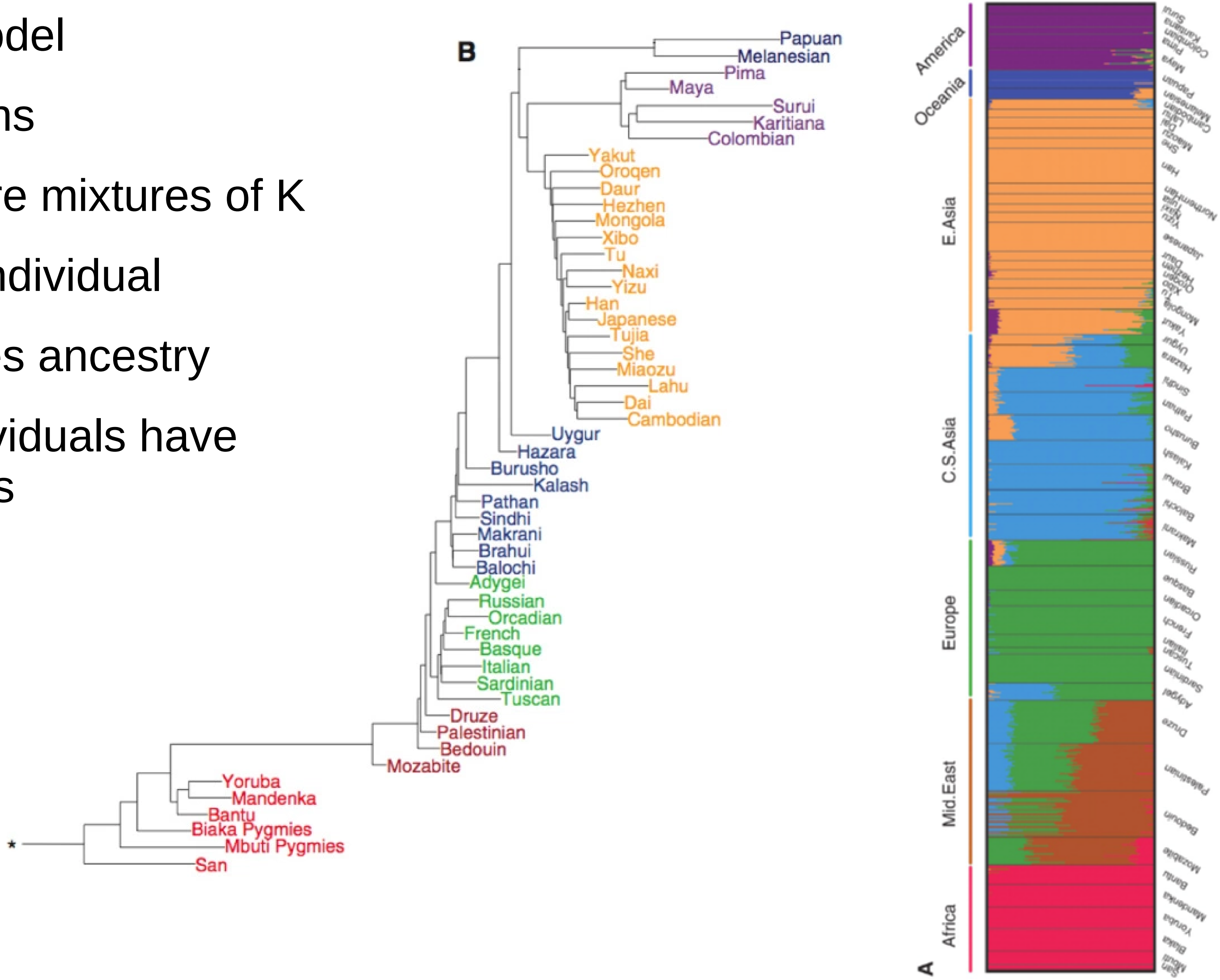
Objective: describe (model) population structure
Problem: individuals can have ancestry from multiple populations
Solution: mixture model

Parent 1   AAAA
Parent 2   TTTT

Offspring ATTT
          TATA
Mixture due to
recombination

Example 2: Admixed ancestry

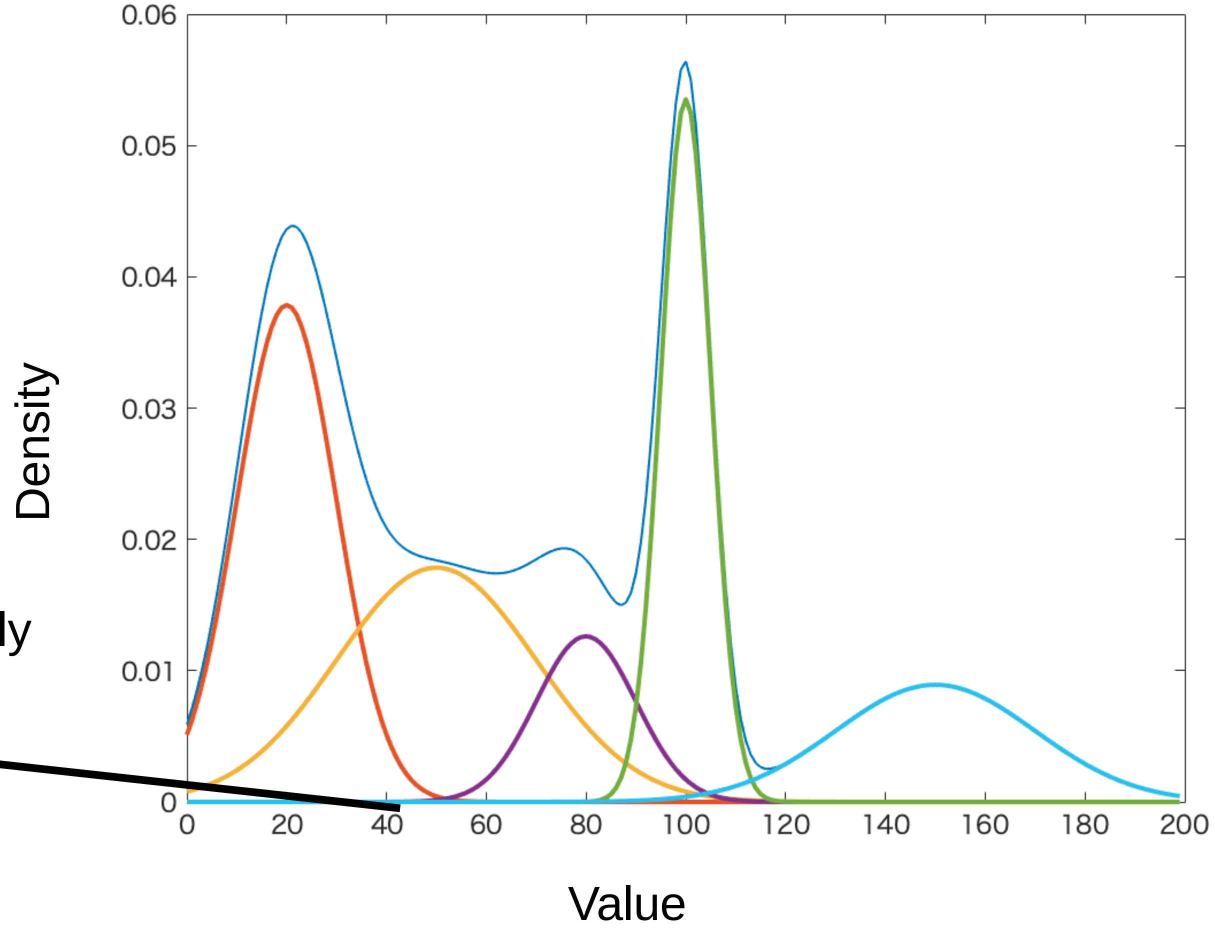
- Admixture model
  - K populations
  - individual are mixtures of K
- Each row is individual
- Color indicates ancestry
- Admixed individuals have multiple colors



Example 3: Overlapping distributions

Problem: Actual groups can be overlapping

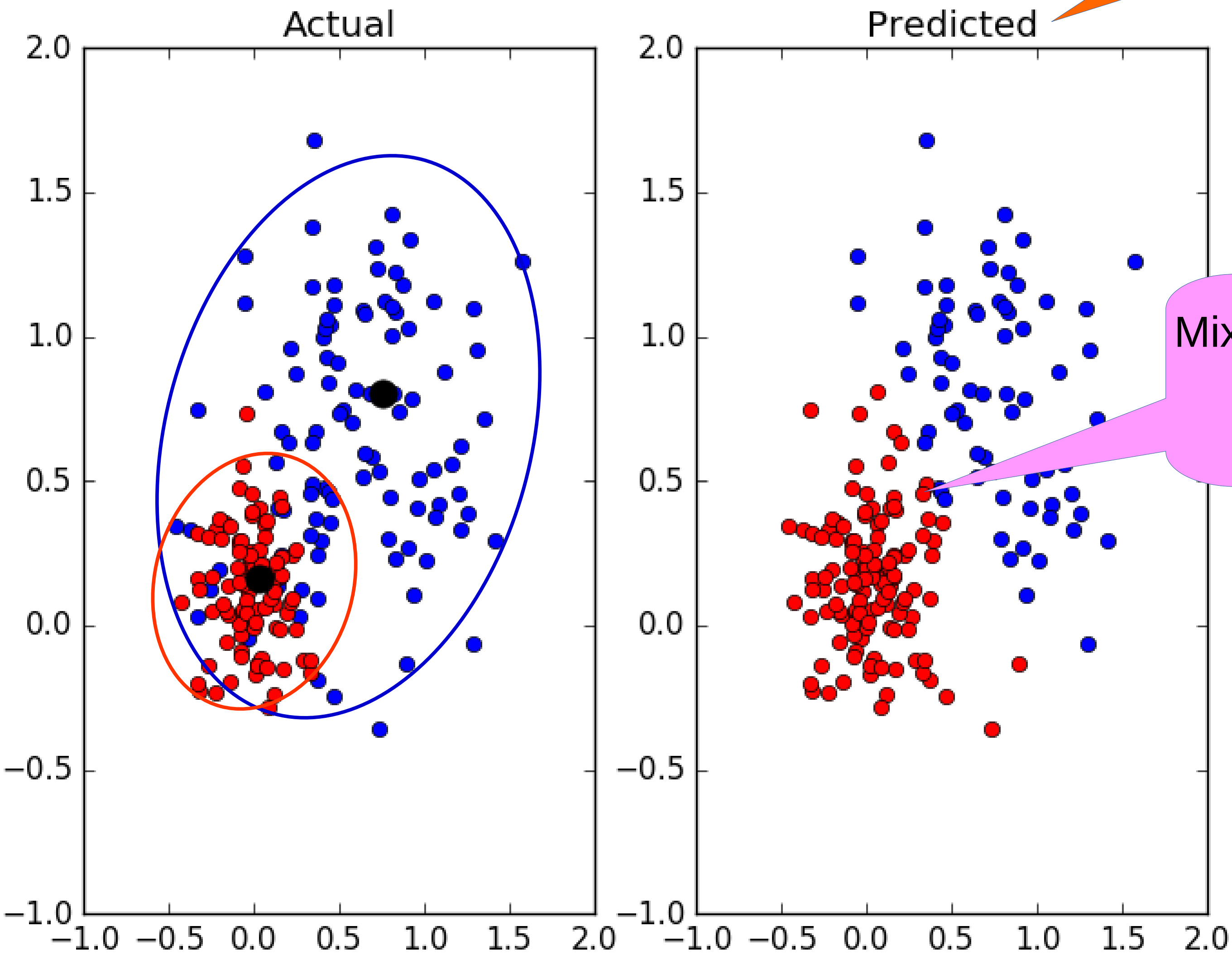
Solution: probabilistic assignment to groups (mixture model)



Value = 40

Most likely yellow, but possibly orange

Mixture models



Kmeans K=2
Hard labels

Mixture (soft labels):
50% blue
50% red

Models

Gaussian mixture model

- group mean (u), variance (width, sigma), mixing probability (pi, abundance of each group)

$$p(\mathbf{x}_n) = \sum_{k=1}^{K} p(\mathbf{x}_n|\mathbf{z})p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)$$

P(data point)

Marginalize over K groups

mixing

Gaussian
u = mean
sigma = variance

- Expectation maximization algorithm
    - Expected population assignment
    - Maximize group mean, variance

Fuzzy clustering

- centroids are *weighted* means, with weights from membership

$$\arg\min_{C} \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^{m} \left\|\mathbf{x}_i - \mathbf{c}_j\right\|^2,$$

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|\mathbf{x}_i - \mathbf{c}_j\|}{\|\mathbf{x}_i - \mathbf{c}_k\|}\right)^{\frac{2}{m-1}}}.$$

w = weights
x = data
c = centroids
m = fuzziness (>=1)
m = 1 = k-means cluster

Exercises

1) Are labels required for supervised or unsupervised machine learning algorithms?

2) Calculate the manhattan distance matrix for genes with the following observations (don't normalize):

Gene	S1	S2	S3
A	1	5	3
B	3	2	5
C	5	1	4
D	4	1	4

Exercises

3) Calculate $d_{A,BD}$ and $d_{C,BD}$ using the distance matrix and a) complete linkage as well as b) single linkage

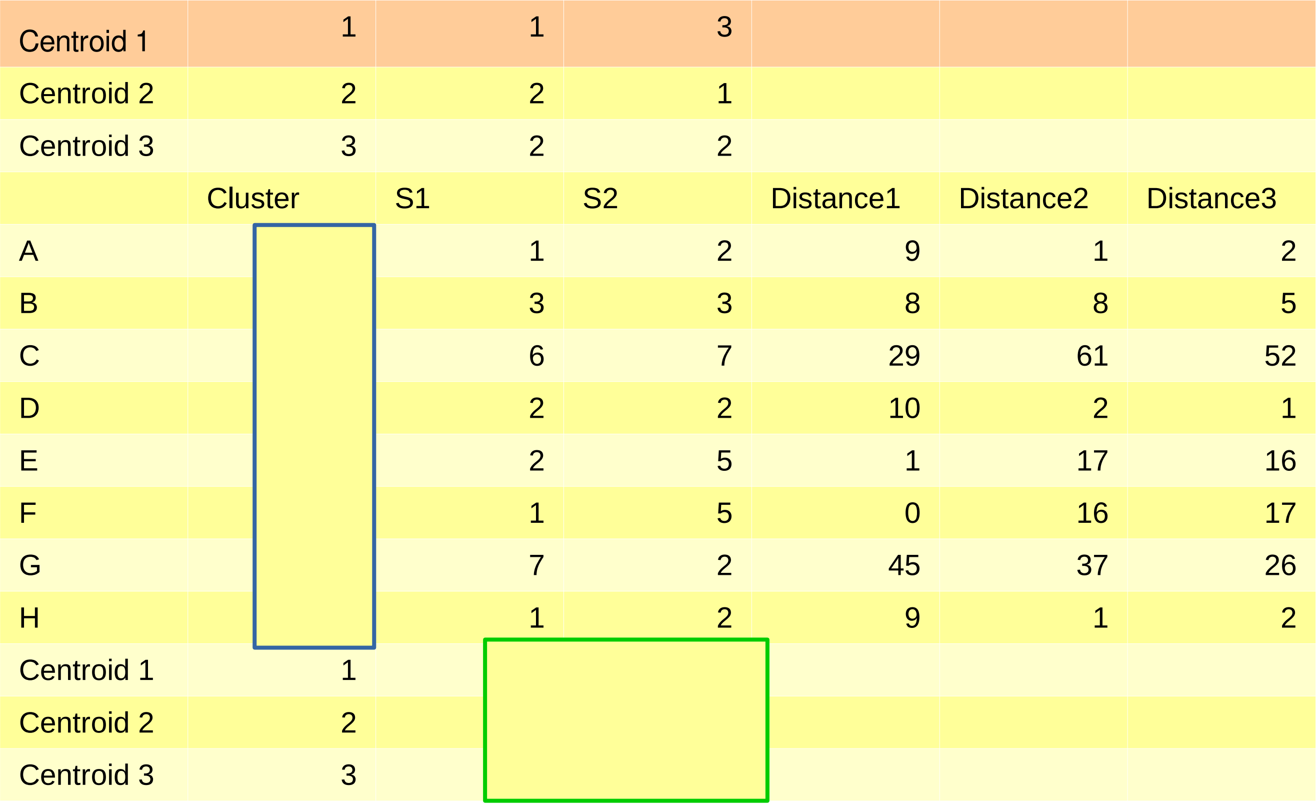
	A	B	C	D
A	0	10	19	10
B	10	0	11	4
C	19	11	0	13
D	10	4	13	0

Exercises

4) What is the hidden variable in K-means clustering?

5) Why is normalization used prior to clustering?

6) Update the cluster assignments (blue) based on the distances, and calculate new centroids (green) using this data:



Centroid 1	1	1	3			
Centroid 2	2	2	1			
Centroid 3	3	2	2			
	Cluster	S1	S2	Distance1	Distance2	Distance3
A		1	2	9	1	2
B		3	3	8	8	5
C		6	7	29	61	52
D		2	2	10	2	1
E		2	5	1	17	16
F		1	5	0	16	17
G		7	2	45	37	26
H		1	2	9	1	2
Centroid 1	1					
Centroid 2	2					
Centroid 3	3					

Exercises

7) What metric does K-means use to assign rows to clusters?

8) What cluster visualization methods would you use for the following tasks:

  a) up-regulated genes in specific samples

  b) Number of clusters as a function of distance between them

  c) Assign ancestry to an individual

9) What type of machine learning would you use to predict benign/malignant (classification, regression)?

10) What algorithm/model can accurately handle two distributions (groups/labels) that overlap?

11) What is the difference between soft and hard clustering?

12) What is the advantage of soft over hard clustering?

13) Mixture models use [hard/soft] labels such that assignments to groups is not discrete.