Clustering

Léon Bottou

NEC Labs America

COS 424 – 3/4/2010

Agenda

Goals	Classification, clustering, regression, other.
Representation	Parametric vs. kernels vs. nonparametric Probabilistic vs. nonprobabilistic Linear vs. nonlinear Deep vs. shallow
Capacity Control	Explicit: architecture, feature selection Explicit: regularization, priors Implicit: approximate optimization Implicit: bayesian averaging, ensembles
Operational Considerations	Loss functions Budget constraints Online vs. offline
Computational Considerations	Exact algorithms for small datasets. Stochastic algorithms for big datasets. Parallel algorithms.

Introduction

Clustering

Assigning observations into subsets with similar characteristics.

Applications

– medecine, biology,
– market research, data mining
– image segmentation
– search results
– topics, taxonomies
– communities

Why is clustering so attractive?

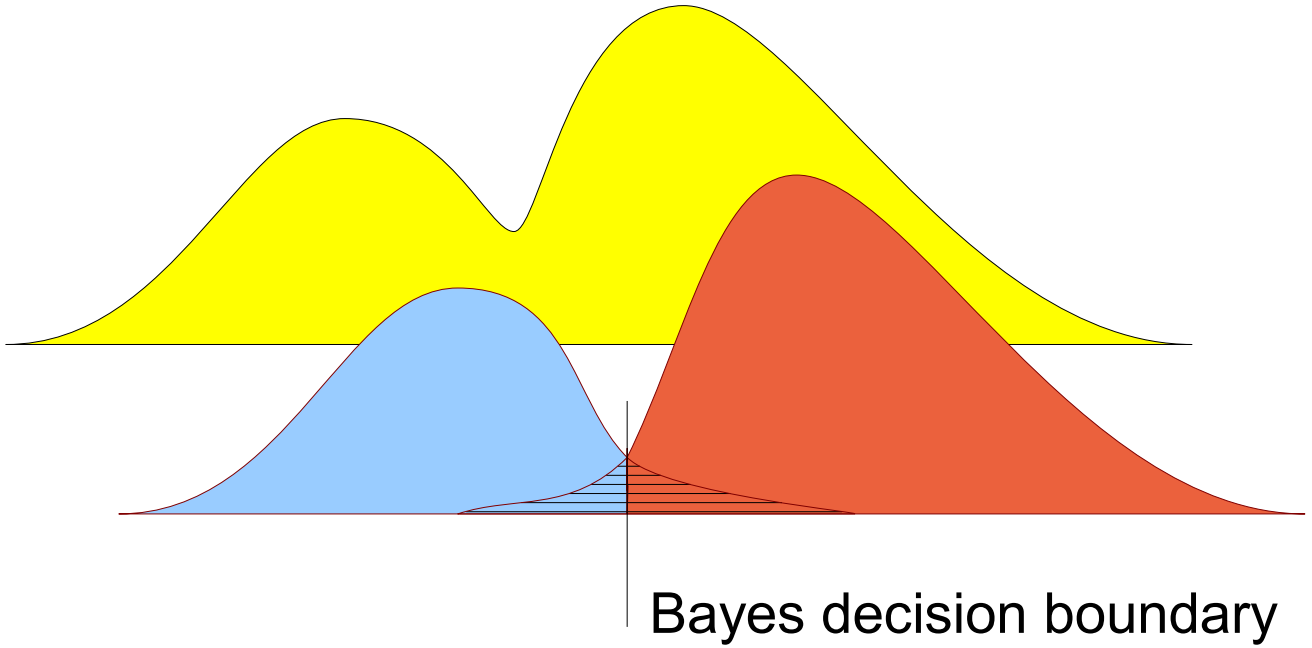
– An embodiment of Descartes' philosophy

> *"Discourse on the Method of Rightly Conducting One's Reason":*
> *"... divide each of the difficulties under examination*
> *... as might be necessary for its adequate solution."*

Summary

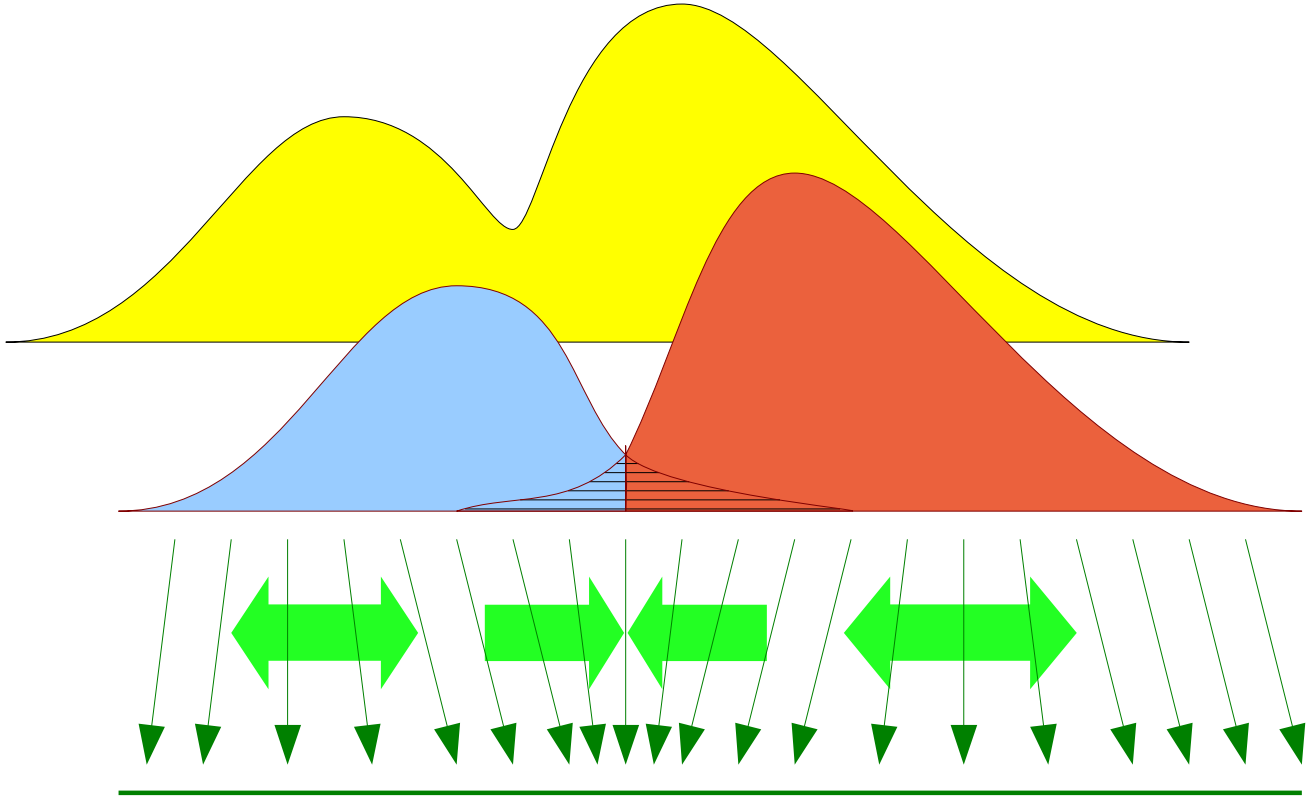
1. What is a cluster?
2. K-Means
3. Hierarchical clustering
4. Simple Gaussian mixtures

What is a cluster?



Two neatly separated classes leave a trace in $\mathbb{P}\{X\}$.

Input space transformations



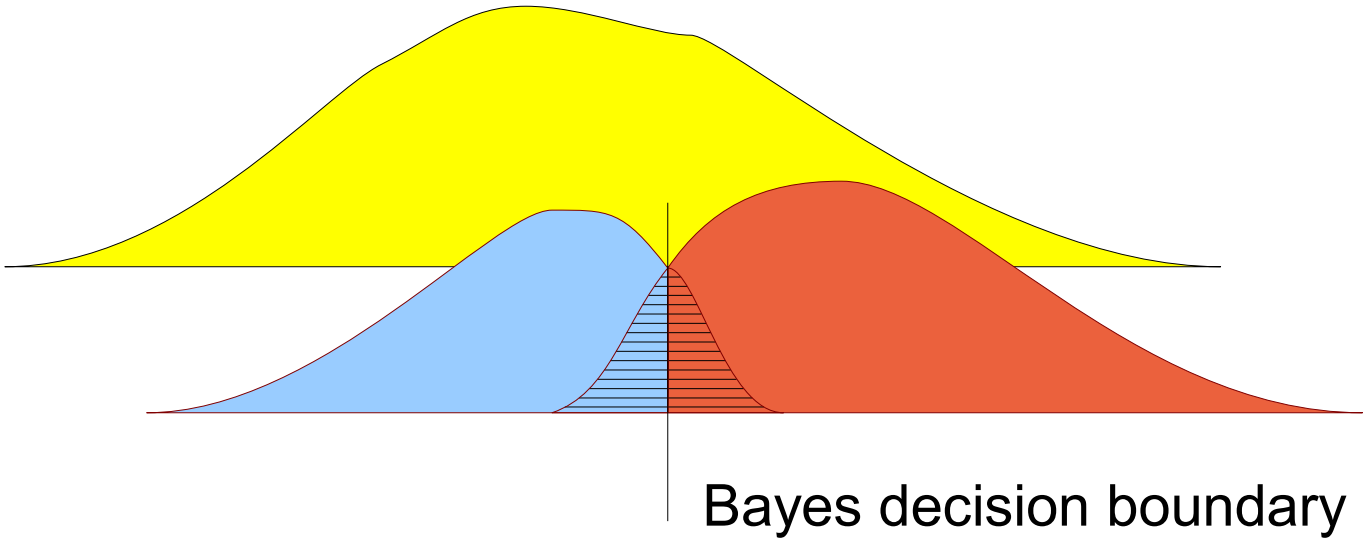
Input space is often an arbitrary decision.

For instance: camera pixels versus retina pixels.

What happens if we apply a reversible transformation to the inputs?

Input space transformations

The Bayes optimal decision boundary moves with the transformation.
The Bayes optimal error rate is unchanged.
The neatly separated clusters are gone!



Clustering depends on the arbitrary definition of the input space!

This is very different from classification, regression, etc.

K-Means

The K-Means problem

– Given observations $x_1 \dots x_n$, determine K centroids $w_1 \dots w_k$

that minimize the distortion $C(w) = \sum_{i=1}^{n} \min_k \|x_i - w_k\|^2$.

Interpretation

– Minimize the discretization error.

Properties

– Non convex objective.
– Finding the global minimum is NP-hard in general.
– Finding acceptable local minima is surprisingly easy.
– Initialization dependent.

Offline K-Means

Lloyd's algorithm

initialize centroids w_k

repeat

- assign points to classes:

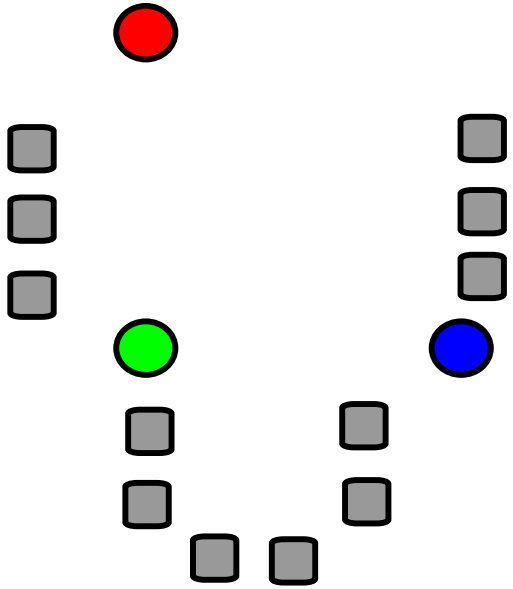
  $\forall i, \quad s_i \leftarrow \arg\min_k \|x_i - w_k\|^2. \qquad S_k \leftarrow \{i \ : \ s_i = k\}.$

- recompute centroids:

  $$\forall k, \quad w_k \leftarrow \arg\min_w \sum_{i \in S_k} \|x_i - w\|^2 = \frac{1}{\text{card}(S_k)} \sum_{i \in S_k} x_i.$$

until convergence.

Lloyd's algorithm – Illustration

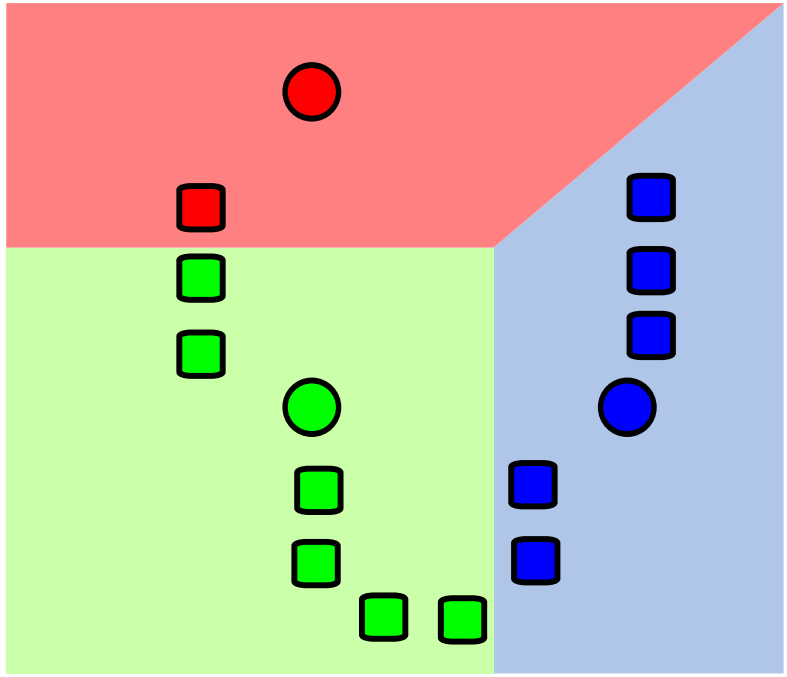


Initial state:

– Squares = data points.

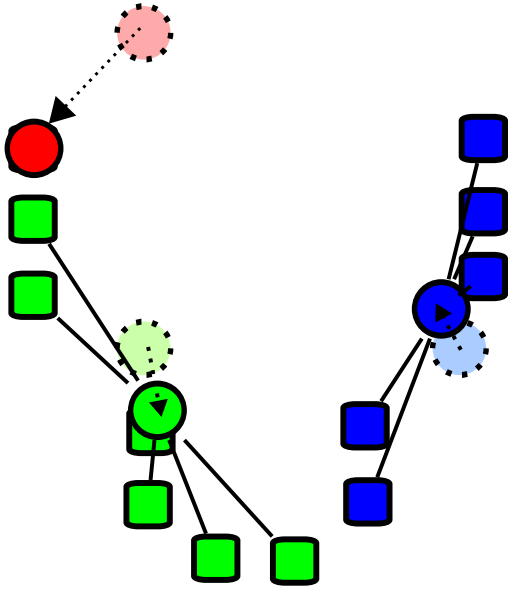
– Circles = centroids.

Lloyd's algorithm – Illustration



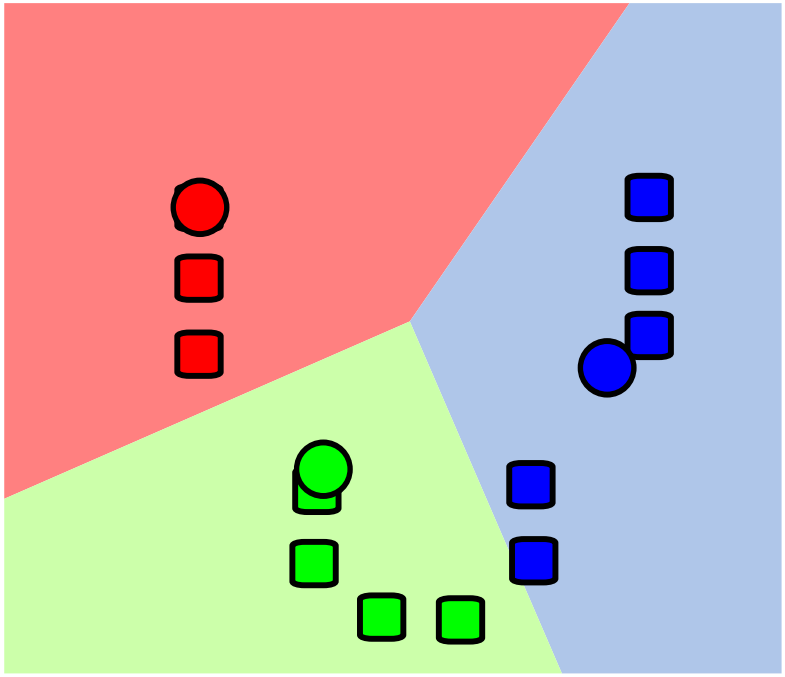
1. Assign data points to clusters.

Lloyd's algorithm – Illustration



2. Recompute centroids.

Lloyd's algorithm – Illustration



Assign data points to clusters. . .

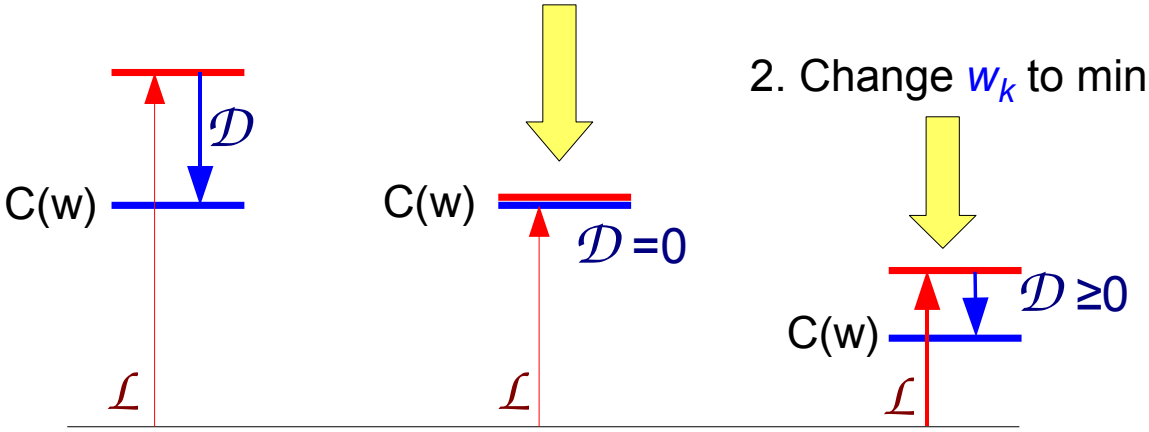
Why does Lloyd's algorithm work?

Consider an arbitrary cluster assignment s_i.

$$C(w) = \sum_{i=1}^{n} \min_k \|x_i - w_k\|^2 = \underbrace{\sum_{i=1}^{n} \|x_i - w_{s_i}\|^2}_{\mathcal{L}(s,w)} - \underbrace{\sum_{i=1}^{n} \|x_i - w_{s_i}\|^2 - \min_k \|x_i - w_k\|^2}_{\mathcal{D}(s,w) \geq 0}$$

1. Change s_i to minimize $\mathcal{D}$ leaving C(w) unchanged.

2. Change w_k to minimize $\mathcal{L}$. Meanwhile $\mathcal{D}$ can only increase.



Online K-Means

MacQueen's algorithm

initialize centroids w_k and $n_k = 0$.

repeat

- pick an observation x_t and determine cluster $s_t = \arg\min_k \|x_t - w_k\|^2$.
- update centroid s_t: $n_{s_t} \leftarrow n_{s_t} + 1$. $\quad w_{s_t} \leftarrow w_{s_t} + \frac{1}{n_{s_t}}(x_t - w_{s_t})$.

until satisfaction.

Comments

– MacQueen's algorithm finds decent clusters much faster.

– Final convergence could be slow. Do we really care?

– Just perform one or two passes over the randomly shuffled observations.

Why does MacQueen's algorithm work?

Explanation 1: Recursive averages.

– Let $u_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then $u_n = u_{n-1} + \frac{1}{n}(x_n - u_{n-1})$.

Explanation 2: Stochastic gradient.

– Apply stochastic gradient to $C(w) = \frac{1}{2n}\sum_{i=1}^{n} \min_k \|x_i - w_k\|^2$:

$$w_{s_t} \leftarrow w_{s_t} + \gamma_t (x_t - w_{s_t})$$

Explanation 3: Stochastic gradient + Newton.

– The Hessian H of $C(w)$ is diagonal and contains the fraction of observations assigned to each cluster.

$$w_{s_t} \leftarrow w_{s_t} + \frac{1}{t} H^{-1} (x_t - w_{s_t}) \;=\; w_{s_t} + \frac{1}{n_{s_t}} (x_t - w_{s_t})$$

Example: Color quantization of images

Problem

– Convert a 24 bit RGB image into a indexed image
  with a palette of K colors.

Solution

– The (r, g, b) values of the pixels are the observations x_i
– The (r, g, b) values of the K palette colors are the centroids w_k.
– Initialize the w_k with an all-purpose palette
– Alternatively, initialize the w_k with the color of random pixels.
– Perform one pass of MacQueen's algorithm
– Eliminate centroids with no observations.
– You are done.

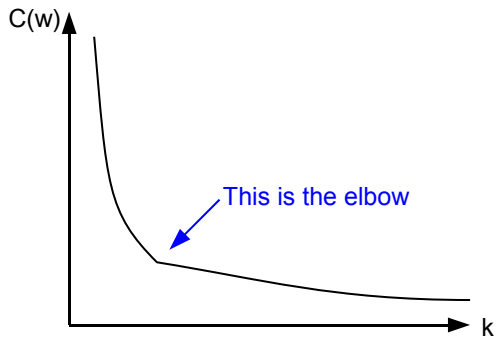
How many clusters?

Rules of thumb ?

– $K = 10$, $K = \sqrt{n}$, ...

The Elbow method ?

– Measure the distortion on a validation set.
– The distortion decreases when k increases.
– Sometimes there is no elbow, or several elbows
– Local minima mess the picture.



Rate-distortion

– Each additional cluster reduces the distortion.
– Cost of additional cluster vs. cost of distortion.
– Just another way to select K.

Conclusion

– Clustering is a very subjective matter.

Hierarchical clustering

Agglomerative clustering

– Initialization: each observation is its own cluster.

– Repeatedly merge the closest clusters

  – single linkage $D(A,B) = \min_{x \in A,\, y \in B} d(x,y)$

  – complete linkage $D(A,B) = \max_{x \in A,\, y \in B} d(x,y)$

  – distortion estimates, etc.

Divisive clustering

– Initialization: one cluster contains all observations.

– Repeatedly divide the largest cluster, e.g. 2-Means.

– Lots of variants.

K-Means plus Agglomerative Clustering

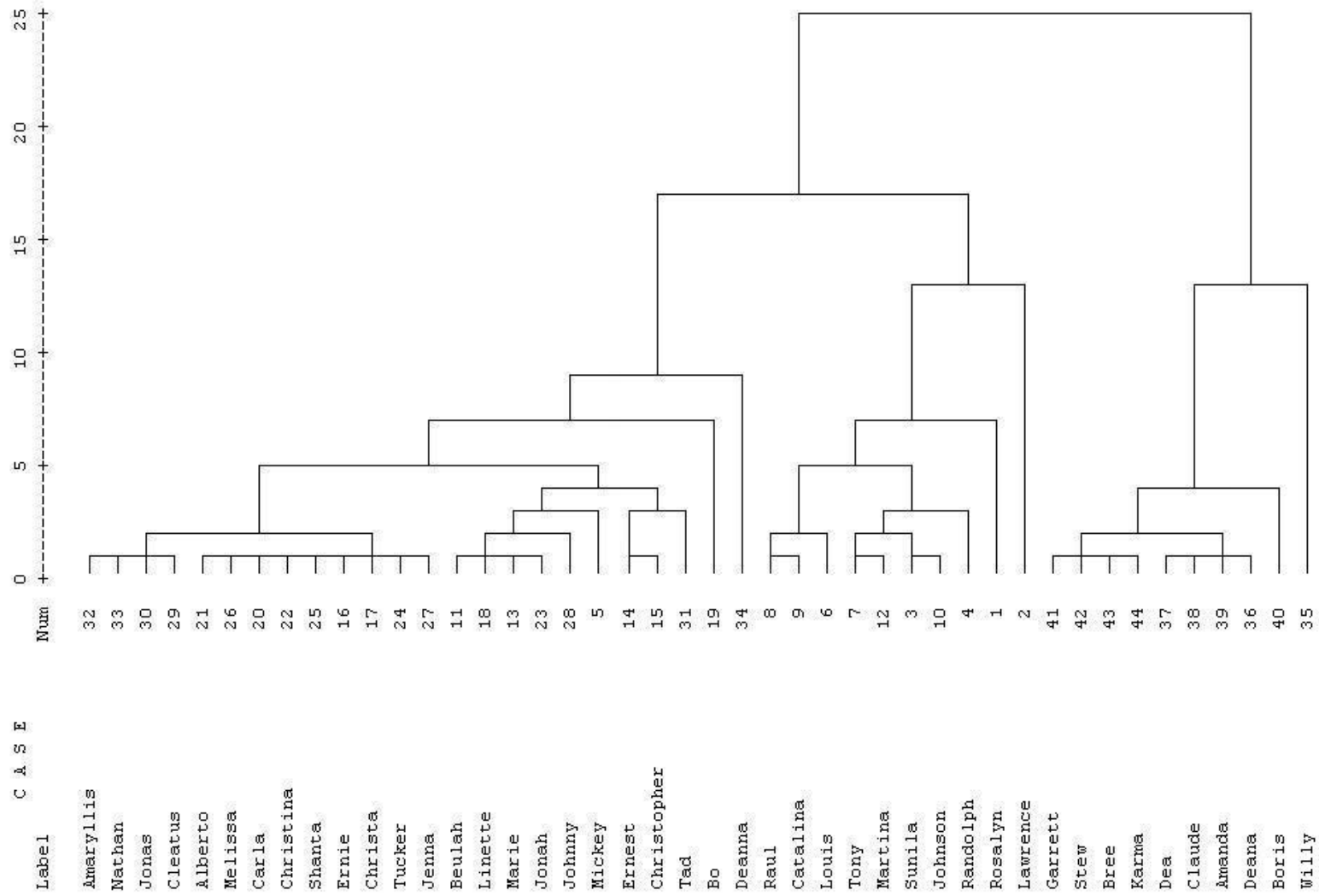
Algorithm

– Run K-Means with a large K.
– Count the number of observation for each cluster.
– Merge the closest clusters according to the following metric.

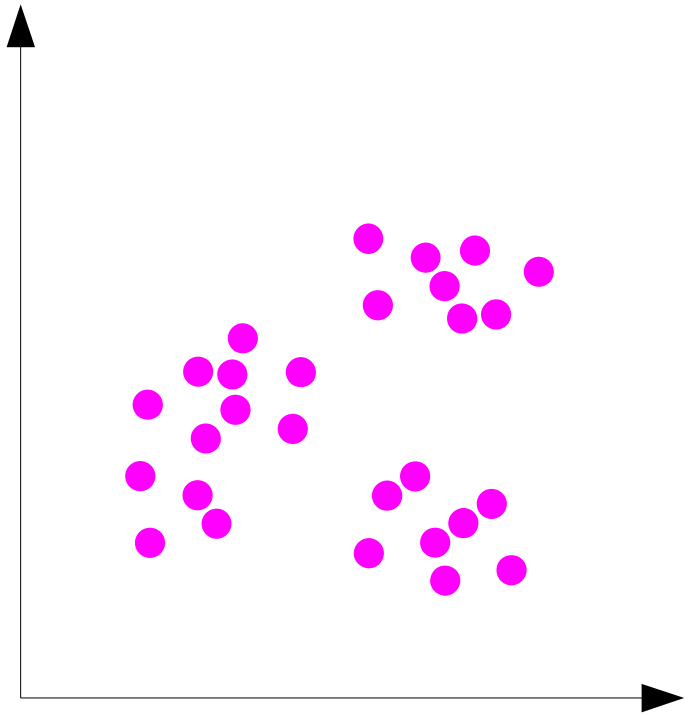
Let A be a cluster with n_A members and centroid w_A.
Let B be a cluster with n_B members and centroid w_B.
The putative center of $A \cup B$ is $w_{AB} = (n_A w_A + n_b w_B)/(n_A + n_B)$.
Quick estimate of the distortion increase:

$$\begin{aligned} d(A,B) &= \sum_{x \in A \cup B} \|x - w_{AB}\|^2 - \sum_{x \in A} \|x - w_A\|^2 - \sum_{x \in B} \|x - w_B\|^2 \\ &= n_A \|w_A - w_{AB}\|^2 + n_B \|w_B - w_{AB}\|^2 \end{aligned}$$

Dendogram


0
5
10
15
20
25
Num
CASE
Label
32 Amaryllis
33 Nathan
30 Jonas
29 Cleatus
21 Alberto
26 Melissa
20 Carla
22 Christina
25 Shanta
16 Ernie
17 Christa
24 Tucker
27 Jenna
11 Beulah
18 Linette
13 Marie
23 Jonah
28 Johnny
5 Mickey
14 Ernest
15 Christopher
31 Tad
19 Bo
34 Deanna
8 Raul
9 Catalina
6 Louis
7 Tony
12 Martina
3 Sunila
10 Johnson
4 Randolph
1 Rosalyn
2 Lawrence
41 Garrett
42 Stew
43 Bree
44 Karma
37 Dea
38 Claude
39 Amanda
36 Deana
40 Boris
35 Willy

Simple Gaussian mixture (1)



Clustering via density estimation.

– Pick a parametric model $\mathbb{P}_\theta(X)$.

– Maximize likelihood.

Pick a parametric model

– There are K components

– To generate an observation:

a.) pick a component k
with probabilities $\lambda_1 \dots \lambda_K$.

b.) generate x from component k
with probability $\mathcal{N}(\mu_i, \sigma)$.

Notes

– Same standard deviation σ (for now).

– That's why I write "Simple GMM".

Simple Gaussian mixture (2)

Parameters: $\theta = (\lambda_1, \mu_1, \ldots, \lambda_K, \mu_K)$

Model: $P_\theta(Y = y) = \lambda_y$. $P_\theta(X = x|Y = y) = \frac{1}{\sigma\,(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}\left(\frac{x-\mu_y}{\sigma}\right)^2}$.

Likelihood

$$\log L(\theta) = \sum_{i=1}^{n} \log P_\theta(X = x_i) = \sum_{i=1}^{n} \log \sum_{y=1}^{K} P_\theta(Y = y) P_\theta(X = x_i|Y = y) = \ldots$$

Maximize!

– This is non convex.

– There are $k!$ copies of each minimum (local or global).

– Conjugate gradients or Newton works.

Expectation-Maximization

Fortunately there is a simpler solution.

– We observe X

– We do not observe Y.

– Things would be simpler if we knew Y.

Decomposition

– For a given X, guess a distribution $Q(Y|X)$.

– Regardless of our guess, $\log L(\theta) = \mathcal{L}(Q,\theta) - \mathcal{M}(Q,\theta) + \mathcal{D}(Q,\theta)$

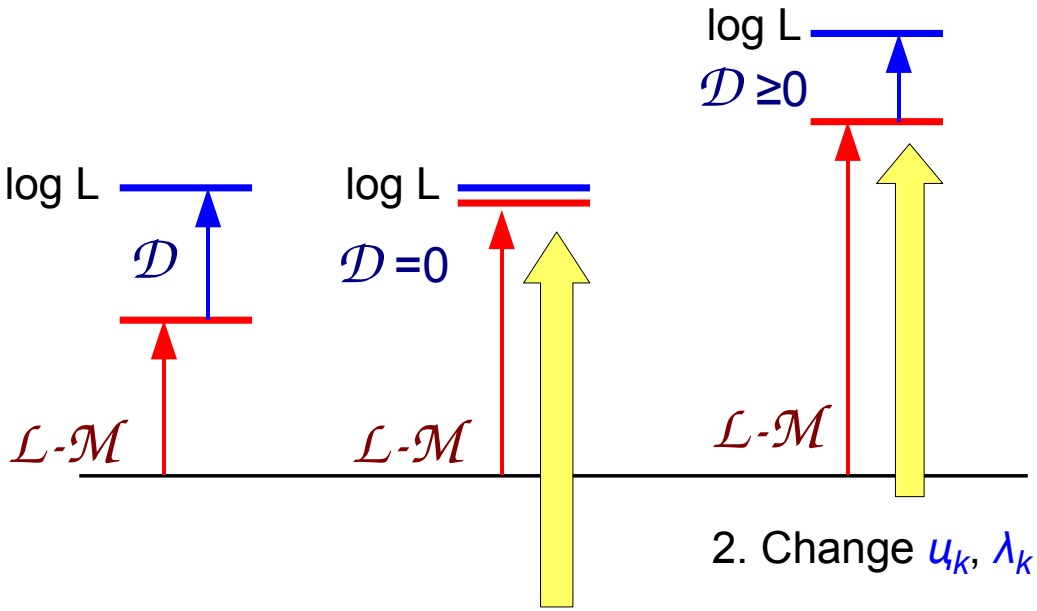
$$\mathcal{L}(Q,\theta) = \sum_{i=1}^{n}\sum_{y=1}^{K} Q(y|x_i) \log P_\theta(x_i|y) \qquad \text{Gaussian log-likelihood}$$

$$\mathcal{M}(Q,\theta) = \sum_{i=1}^{n}\sum_{y=1}^{K} Q(y|x_i) \log \frac{Q(y|x_i)}{P_\theta(y)} \qquad \text{KL divergence } D(P_Y \| Q_{Y|X})$$

$$\mathcal{D}(Q,\theta) = \sum_{i=1}^{n}\sum_{y=1}^{K} Q(y|x_i) \log \frac{Q(y|x_i)}{P_\theta(y|x_i)} \qquad \text{KL divergence } D(Q_{Y|X} \| P_{Y|X})$$

Expectation-Maximization

Remember Lloyd's algorithm for K-Means?



2. Change u_k, λ_k to maximize $\mathcal{L}$-$\mathcal{M}$. Meanwhile $\mathcal{D}$ can increase.

1. Change Q to minimize $\mathcal{D}$ leaving log L unchanged.

E-Step Soft assignments $\quad q_{ik} \leftarrow \lambda_k \, e^{-\frac{1}{2}\left(\frac{x_i - \mu_k}{\sigma}\right)^2}$

M-Step Update parameters $\quad \mu_k \leftarrow \frac{\sum_i q_{ik} \, x_i}{\sum_i q_{ik}}. \quad \lambda_k \leftarrow \frac{\sum_i q_{ik}}{\sum_{iy} q_{iy}}.$

Simple Gaussian mixture (3)

Relation with K-Means

– Like K-Means but with soft assignments.

– Limit to K-Means when $\sigma \rightarrow 0$.

In practice

– Clearly slower than K-Means.

– More robust to local minima.

– Annealing σ helps.

Subtleties

– Relation between σ and the number of clusters...

– Relation between EM and Newton.

Next Lecture

Expectation Maximization in general

– EM for general Gaussian Mixtures Models.
– EM for all kinds of mixtures.
– EM for dealing missing data.