+

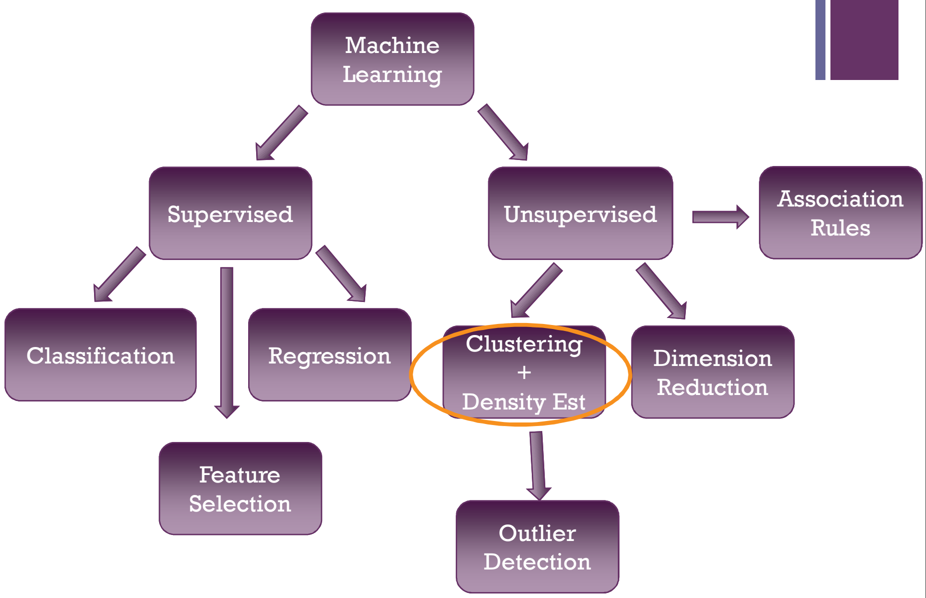
Data Mining (ECS607U)
Lecture 6 – Clustering and Density Estimation

Dr. Ioannis Patras
EECS, Queen Mary University of London

Slide thanks: Tim Hospedales

+

A Taxonomy



+ Unsupervised Learning Motivation

- Unsupervised Learning:
  - Too much data: We need to save memory/computation.
    - Reduce the data to a more manageable amount
  - Don't understand the data
    - Exploratory data analysis
    - What underlying knowledge is there?
    - Discover patterns & trends
- Dimensionality Reduction (last week)
  - Focus on dimensions (columns)
- Clustering (this week)
  - Focus on instances (rows)

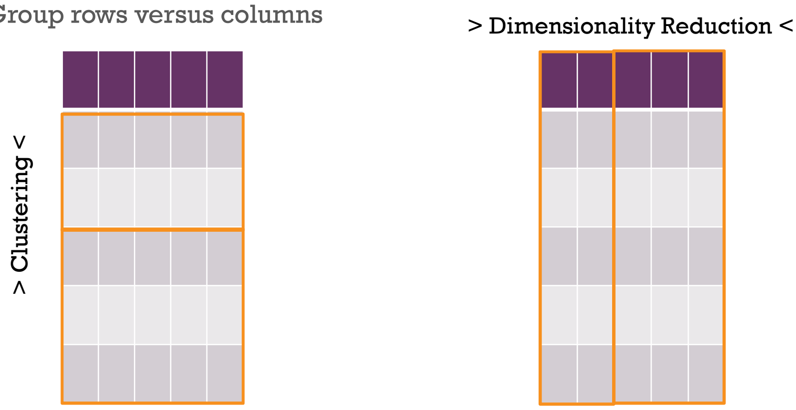
+ Overview

- Overview: What is clustering about?
- Clustering Algorithms
  - K-means
  - Hierarchical Clustering
  - (Density Estimation)
  - Gaussian Mixtures
- Further topics
  - Choosing the number of clusters
  - Advanced algorithms
- Some applications
  - ..besides marketing!

+ Clustering: Unsupervised Learning Context

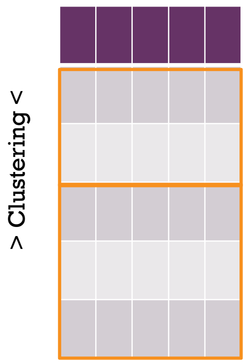
Clustering versus Dimensionality Reduction

- Both unsupervised
- Group rows versus columns

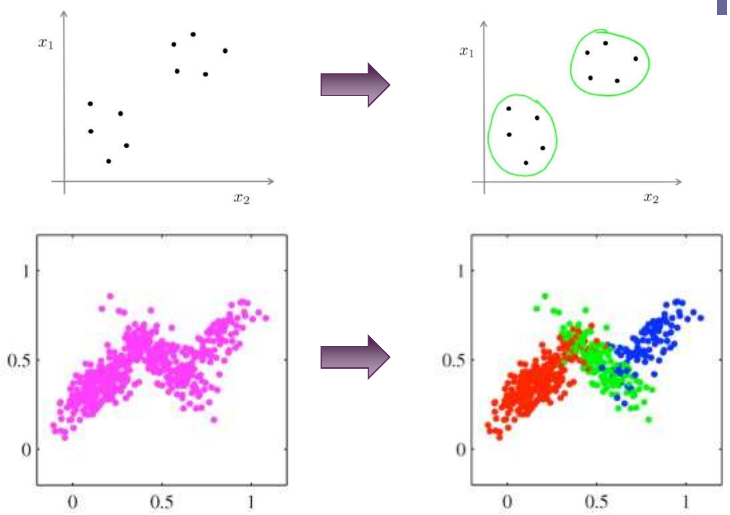


+ Clustering

- Problem definition:
  - Given a set of examples $(x_1, x_2, .., x_N)$
  - Divide them into subsets of similar examples.
- Alternative view:
  - How to do classification if you have no labeled training examples
- Conceptual & algorithmic challenges:
  - How to quantify similarity?
  - How to measure the goodness of a partition?



+ Clustering: Visual Example



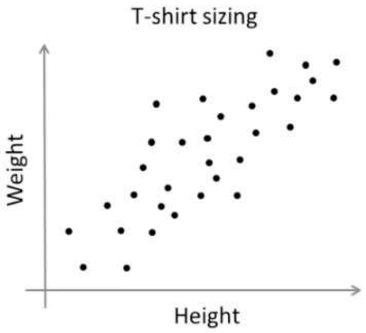
+ Clustering: Why do it?

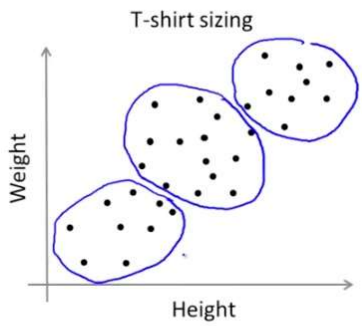
- Market segmentation
  - Teenagers, Mothers, Empty-nesters
  - Targeted products/marketing for each "cluster" of customers
- Data-center organization
- Social network analysis (find recommended friends)
- Group articles on your website or blog
- Group websites on your aggregator
- T-shirts: Given height & weight in the population:
  - How big should size of S, M, L be?
  - Should you offer S,M,L or XS, S, M, L, XL?

+ Clustering: Why do it?

- T-shirts: Given height & weight in the population:
  - How big should size of S, M, L be?
  - Should you offer S,M,L or XS, S, M, L, XL?

S, M, L

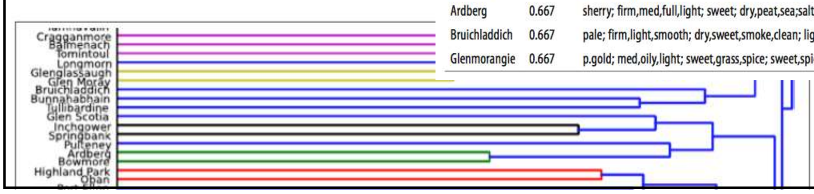




+ Clustering: Why do it?

- Stock choice in retail:
  - There are more possible products to stock than we have space for.
  - Which ones to include?
    - Try to cover the range of possible options so a consumer of every preference can find something they like.
    - How? Clustering then pick one example from each cluster.
    - Question: Have you covered the space well?

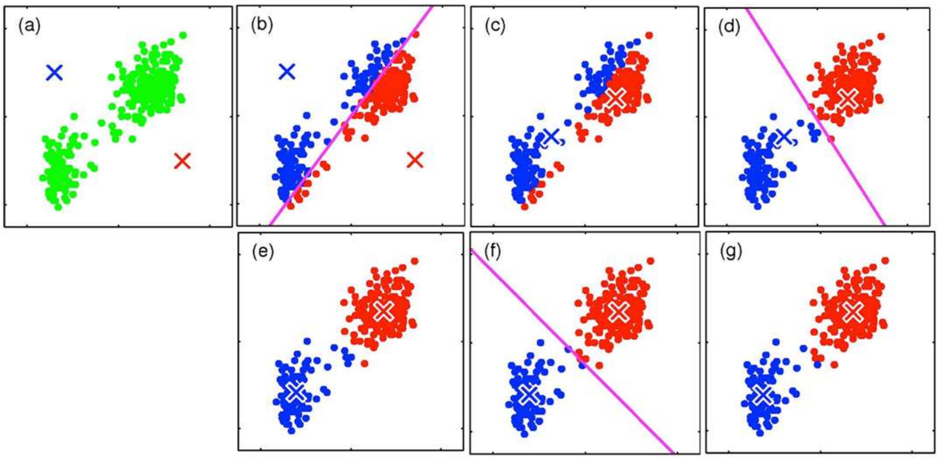
Whiskey	Distance	Descriptors
Bunnahabhain	—	gold; firm,med,light; sweet,fruit,clean; fresh,sea; full
Glenglassaugh	0.643	gold; firm,light,smooth; sweet,grass; fresh,grass
Tullibardine	0.647	gold; firm,med,smooth; sweet,fruit,full,grass,clean; sweet; big,arome,sweet
Ardberg	0.667	sherry; firm,med,full,light; sweet; dry,peat,sea;salt
Bruichladdich	0.667	pale; firm,light,smooth; dry,sweet,smoke,clean; light; full
Glenmorangie	0.667	p.gold; med,oily,light; sweet,grass,spice; sweet,spicy,grass,sea,fresh; full,long



Overview

- Overview: What is clustering about?
- Clustering Algorithms
  - K-means
  - Hierarchical Clustering
  - (Density Estimation)
  - Gaussian Mixtures
- Further topics
  - Choosing the number of clusters
  - Advanced algorithms
- Some applications
  - ..besides marketing!

K-Means Algorithm - illustration



K-Means Algorithm

- Input:
  - N point dataset, $D=\{x_1,x_2,\ldots x_n\}$,
  - Number of clusters K.
- Initialize randomly K centers $Ctr_1,..,Ctr_k$
- Repeat
  - For i=1:N
    - $Labels_i$=Cluster centroid closest to x_i
  - For k=1:K
    - Ctr_k = average of points assigned to k

Formalizing K-means

- Cost Function:
  - Find cluster centers $u_{1:k}$ and cluster assignments $c_{1:N}$ so as to minimize the sum squared distances of points from assigned clusters:

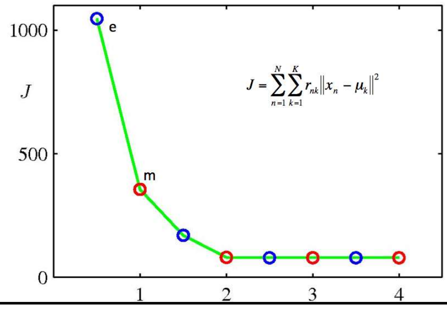
$$E_{KM}(D, c_{1:N}, \mu_{1:K}) = \sum_{i=1}^{N} \left(x_i - \mu_{c_i}\right)^2$$

- Algorithm:
  - "E-step": Find cluster closest to each point (fix u, minimize for c)
  - "M-step": Find new center of each cluster (fix c, minimize for u)
- Aside:
  - We have seen algorithms with exact & gradient solutions to problems.
  - This is our first alternating minimization solution
  - An exact solution to each part of the problem given the other: Iterate

Formalizing K-means

- Cost Function:
  - Find cluster centers $u_{1:k}$ and cluster assignments $c_{1:N}$ so as to minimize the distance of each point from it's assigned cluster:

$$E_{KM}(D, c_{1:N}, \mu_{1:K}) = \sum_{i=1}^{N} \left(x_i - \mu_{c_i}\right)^2$$



K-means Properties

- Recall the algorithm:
  - Repeat S times:
  - "E-step": Find cluster closest to each point (fix u, minimize for c)
  - "M-step": Find new center of each cluster (fix c, minimize for u)
- Computation time?
  - O(NK), Or O(NKDS) if dimension and iterations included
  - Fast relative to O(N^2), slow relative to O(N). (i.e., if large K)

+ K-means Properties

- Distance Metric
  - Typically use Euclidean
    - May or may not be appropriate depending on data.
    - May not be robust to outliers
  - What happens if you have categorical data?
    - use 1-of-N encoding

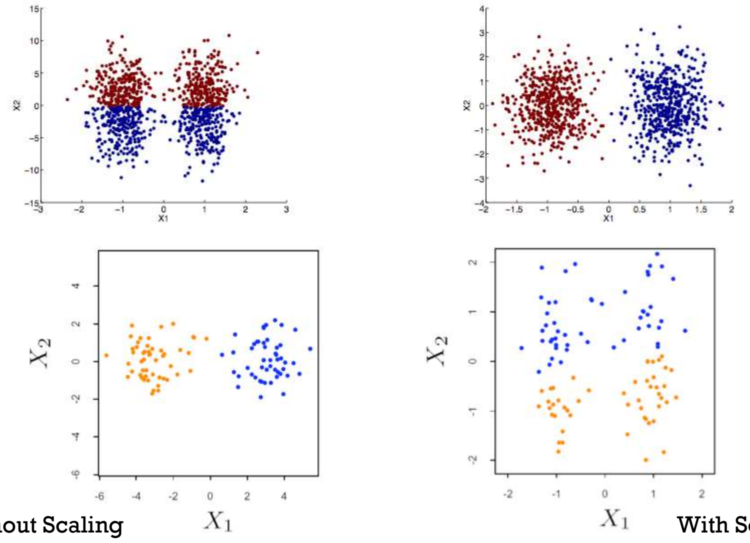
$$E_{KM}(D, c_{1:N}, \mu_{1:K}) = \sum_{i=1}^{N} \left(x_i - \mu_{c_i}\right)^2$$

- Convergence:
  - It converges to a local minima only
  - => In practice repeat with many random initializations and pick the best
  - Different distances lead to changes in both steps!!

+ K-means: When it (doesn't) work

- Sensitive to data scaling
  - Renormalize in [0,1] or by standard deviation

+ # Scaling?

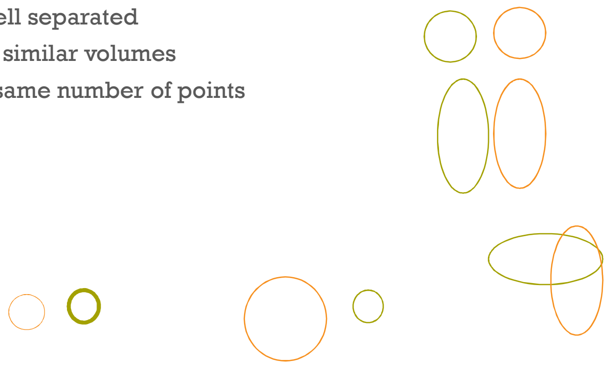


Without Scaling

With Scaling

+ # K-means: When it (doesn't) work

- Works if:
  - Clusters are spherical
  - Clusters are well separated
  - Clusters are of similar volumes
  - Clusters have same number of points

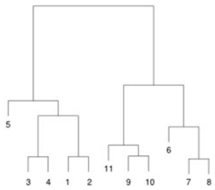


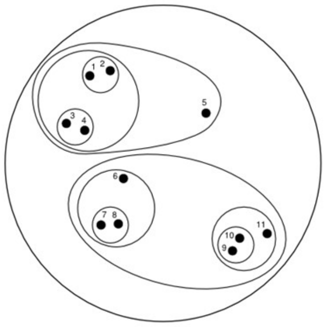
+ # Overview

- Overview: What is clustering about?
- Clustering Algorithms
  - K-means
  - Hierarchical Clustering
  - (Density Estimation)
  - Gaussian Mixtures
- Further topics
  - Choosing the number of clusters
  - Advanced algorithms
- Some applications
  - ..besides marketing!

+ # Hierarchical Clustering

- Sometimes you want a tree of similarity rather than a flat clustering
  - (And K-means clusters discovered can be sensitive to chosen K)
- Output: A dendrogram (instead of cluster centers and assignments)

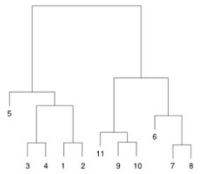


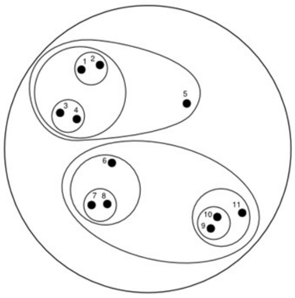


+ Hierarchical Clustering

Algorithm: Agglomerative (or Divisive)

- Start with one cluster per example
- Merge two nearest clusters
  - E.g., min, max, mean distance.
- Repeat until one cluster





+ Summary

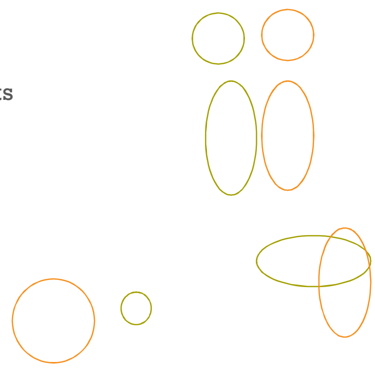
- Clustering identifies typical groups.
- Need to understand the groups.
  - What do they have in common? Two options:
  - Manually examine elements of a cluster.
  - Use a supervised classifier!
  1. Use the cluster labels as a supervision for a classifer.
  2. Run the classifier, and examine the weights on each feature. The weights will say what is unique about each cluster.

+ Overview

- Overview: What is clustering about?
- Clustering Algorithms
  - K-means
  - Hierarchical Clustering
  - (Density Estimation)
  - Gaussian Mixtures
- Further topics
  - Choosing the number of clusters
  - Advanced algorithms
- Some applications
  - ..besides marketing!

+ K-means: When it (doesn't) work

- Works if:
  - Clusters are spherical
  - Clusters are well separated
  - Clusters are of similar volumes
  - Clusters have same number of points
- Issue:
  - Hard assignments
- Motivate:
  - Mixture of Gaussians algorithm



+

Before we get to GMMs
Recap some concepts in Probability

+

Probability & Density Estimation 1

Three common probability distributions

- Binary variables: Bernoulli
  - x is 1,0 (Heads or Tails).
  - u (probability of Heads) from 0 to 1.

$$p(x;u) = u^x (1-u)^{(1-x)}$$

- Categorical variables: Multinomial
  - x is 1-of-K encoding.
  - ui (probability of outcome i) from 0 to 1. ui's sum to 1.

$$p(\mathbf{x};\mathbf{u}) = \prod u_i^{x_i}$$

- Continuous variables: Gaussian
  - x is real vector. u is a real vector. S is a matrix.

$$p(\mathbf{x};\mathbf{u},S) = \frac{1}{Z}\exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{u})^T S^{-1}(\mathbf{x}-\mathbf{u})\right)$$

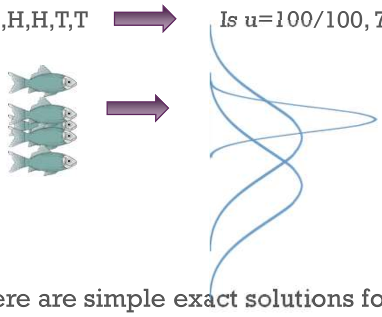
+ Probability & Density Estimation 2

- Generative Perspective:
  - Distributions tell us what data to expect according to specified parameters
  - E.g., Biased coin (u=3/4). ⇨ Expect H,H,H,T
  - E.g., Mean & var of fish length ⇨ Expect

- Density Estimation:
  - Ask what probability distribution was responsible for specified data?

+ Probability & Density Estimation 3

- Density Estimation:
  - Ask what probability distribution was responsible for specified data?
  - H,H,H,T,T ⇨ *Is* u=100/100, 75/100, 25/100, etc more likely?



- There are simple exact solutions for the best estimates of binary, categorical, and Gaussian distributions given data

Probability & Density Estimation 4

- Density Estimation:
  - Ask what probability distribution was responsible for specified data?
  - There are simple exact solutions for the best estimates of binary, categorical, and Gaussian distributions given data

$$p(x;u) = u^x (1-u)^{(1-x)} \quad \Longrightarrow \quad u = \frac{1}{N}\sum_i x_i$$

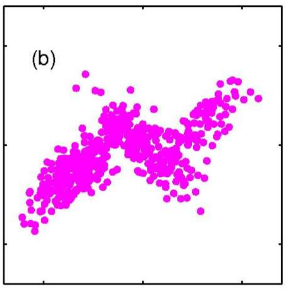
$$p(\mathbf{x};\mathbf{u}) = \prod u_k^{x_k} \quad \Longrightarrow \quad u_k = \frac{\sum_i x_{ik}}{\sum_i \sum_k x_{ik}} = \frac{N_k}{N}$$

$$p(\mathbf{x};\mathbf{u},S) = \frac{1}{Z}\exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{u})^T S^{-1}(\mathbf{x}-\mathbf{u})\right) \quad \Longrightarrow \quad \mathbf{u} = \frac{1}{N}\sum_i \mathbf{x}_i \quad S = \frac{1}{N}\sum_i (\mathbf{x}-\mathbf{u})(\mathbf{x}-\mathbf{u})^T$$

Back to GMMs

+ K-means: When it (doesn't) work

- Works if:
  - Clusters are spherical
  - Clusters are well separated
  - Clusters are of similar volumes
  - Clusters have same number of points
- Issue:
  - Hard assignments
- K-means won't work for data like this:

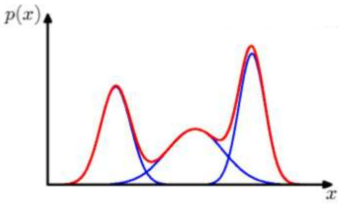


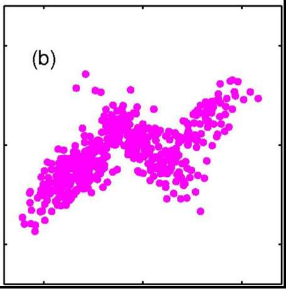
- Motivate:
  - Mixture of Gaussians algorithm

+ Gaussian Mixture Models

- Tough data
  - K-means has problems
  - Single Gaussian doesn't fit it well
- GMM solution:
  - Explain as: Weighted sum of K Gaussian densities

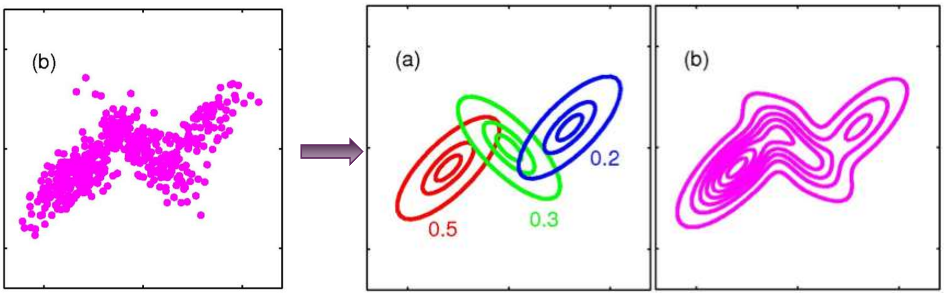
$$p(\mathbf{x}) = \sum_k \pi_k N(\mathbf{x}; \mathbf{u}_k, S_k)$$





Gaussian Mixture Models

- GMM clustering:
  - Explain as: Weighted sum of K Gaussian densities $p(\mathbf{x}) = \sum_k \pi_k N(\mathbf{x}; \mathbf{u}_k, S_k)$
- Example problem and solution



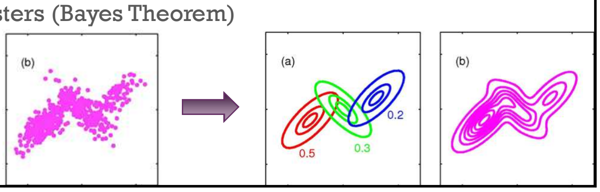
- … But what algorithm can obtain these solutions?

Gaussian Mixture Models: Solution

- Optimization Criteria: Maximum Likelihood

$$L(D; \boldsymbol{\pi}, \mathbf{u}, S) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k N(\mathbf{x}_i; \mathbf{u}_k, S_k)$$

- Solution?
  - If we knew which points belong to which clusters, we know how to fit Gaussians (Density Estimation: Gaussian)
  - If we knew which points belong to which cluster, we know the relative size of each (Density Estimation: Multinomial)
  - If we knew the Gaussians, we could find out which points belonged to which clusters (Bayes Theorem)

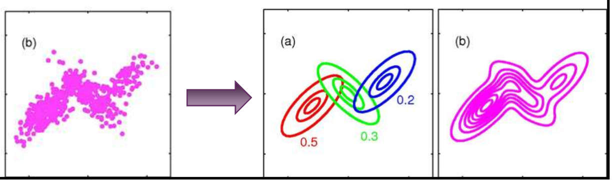


Gaussian Mixture Models: Solution

- Optimization Criteria: Maximum Likelihood

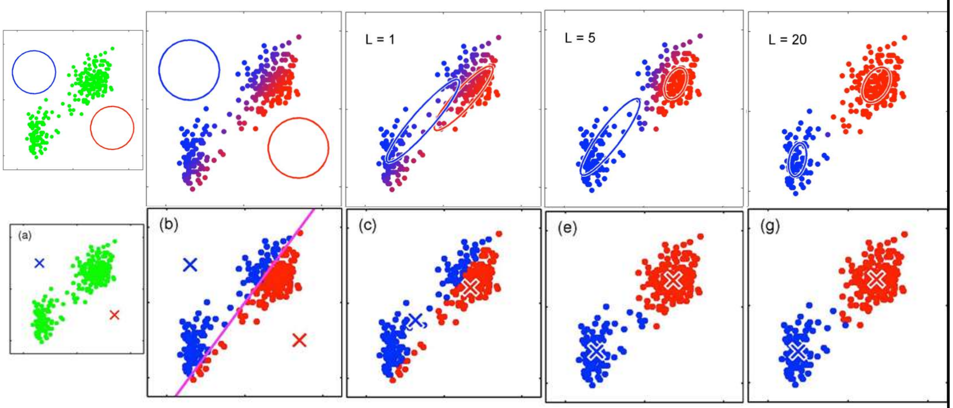
$$L(D;\boldsymbol{\pi},\mathbf{u},S)=\prod_{i=1}^{N}\sum_{k=1}^{K}\pi_k N(\mathbf{x}_i;\mathbf{u}_k,S_k)$$

- EM Algorithm Solution
  - E: Given Gaussians, infer how likely each point to each cluster.
  - M: Given soft assignments, update Gaussians and cluster prior.



GMM: EM Solution Example

What contrasts do you see?



GMM: Limitations

- Converges to local optima only
  - Can also do multiple restarts and pick the best
- Picking K is still an issue
- Cost $O(NKD^2)$
  - Data requirements >> $O(D)$ due to covariance matrix
    - (Estimating the shape information of each)

GMM versus K-Means

K-means	GMM

Algorithm:
- For i=1:N
  - c_i=Index of the cluster centroid closest to x_i
- For k=1:K
  - u_k = average of points assigned to k
	Algorithm:
	- For i = 1 : N
	  - Probability $p(k\|x_i)$ of belonging to each cluster k
	- For k=1:K
	  - Fit the Gaussian $N(u_k,S_k)$ given probabilities $p(k\|x_i)$
	  - Fit the cluster prior $p(k)$ given probabilities $p(k\|x)$.

GMM versus K-Means

K-means	GMM
Assumptions	Assumptions
■ All clusters same size	■ Cluster k of size π_k
■ All clusters spherical	■ Clusters of shape S
■ Clusters are sharply peaked	■ Clusters have spread S
■ Hard assign points to clusters	■ Soft assign points to clusters
■ Optimize:	■ Optimize:
■ Sum Squared Dist of points to clusters	■ Log-likelihood of data

$$E_{KM}(D, c_{1:N}, \mu_{1:K}) = \sum_{i=1}^{N} \left(x_i - \mu_{c_i}\right)^2$$

$$L_{GMM}(D; \boldsymbol{\pi}, \mathbf{u}, S) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k N(\mathbf{x}_i; \mathbf{u}_k, S_k)$$

GMM versus K-Means

K-means	GMM
■ Pick K is non-trivial	■ Pick K is non-trivial
■ Only local optimization	■ Only local optimization
■ Cost:	■ Cost:
■ CPU: O(NKD)	■ CPU: O(NKD^2):
	■ Data: >> O(D)

$$E_{KM}(D, c_{1:N}, \mu_{1:K}) = \sum_{i=1}^{N} \left(x_i - \mu_{c_i}\right)^2$$

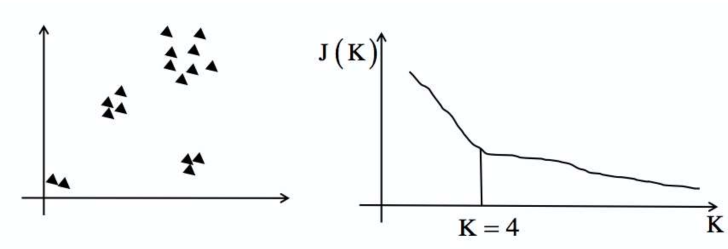
$$L_{GMM}(D; \boldsymbol{\pi}, \mathbf{u}, S) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k N(\mathbf{x}_i; \mathbf{u}_k, S_k)$$

Overview

- Overview: What is clustering about?
- Clustering Algorithms
  - K-means
  - Hierarchical Clustering
  - (Density Estimation)
  - Gaussian Mixtures
- Further topics
  - Choosing the number of clusters
  - Advanced algorithms
- Some applications
  - ..besides marketing!

Choosing Number of Clusters
For K-means and GMM

- 1. Elbow Method
  - Plot $E_{KM/GMM}$ as a function of k, and choose the elbow point.



- 2. Present results to end users see what they prefer
  - Broader or more specialized groups

+
Choosing Number of Clusters
For K-means and GMM

- 3. Cross-validation
  - For K = 1…Large
    - Learn GMM/KM clusters on a train set.
    - Evaluate Quality(K) = quality on validation set.
  - Pick K with the highest validation set quality.

+
Choosing Number of Clusters
For K-means and GMM

- 4. BIC/AIC Criterion
  - (ML people: An approximation to the integration required in the Bayesian model selection)
  - Adds a penalty to the cost that penalises more complex models.
    - P: Number of parameters in model. N: Number of data points.
  - Evaluate modified cost E^{K}_{BIC} for many values of K.
  - Pick the K with best cost

$$E^{K}{}_{BIC} = E^{K} - \frac{p}{2}\log N$$

$$E_{KM}(D, c_{1:N}, \mu_{1:K}) = \sum_{i=1}^{N}\left(x_i - \mu_{c_i}\right)^2 \qquad E_{GMM}(D; \boldsymbol{\pi}, \mathbf{u}, S) = -\log\prod_{i=1}^{N}\sum_{k=1}^{K}\pi_k N(\mathbf{x}_i; \mathbf{u}_k, S_k)$$

+
Overview

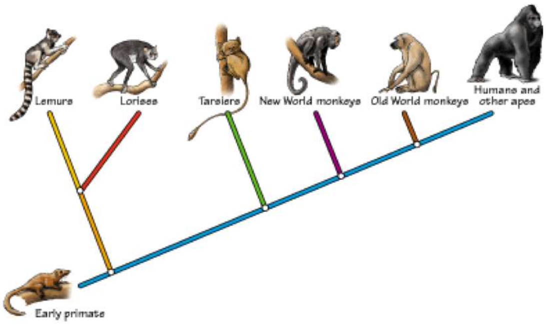
- Overview: What is clustering about?
- Clustering Algorithms
  - K-means
  - Hierarchical Clustering
  - (Density Estimation)
  - Gaussian Mixtures
- Further topics
  - Choosing the number of clusters
  - Advanced algorithms
- Some applications
  - ..besides marketing!

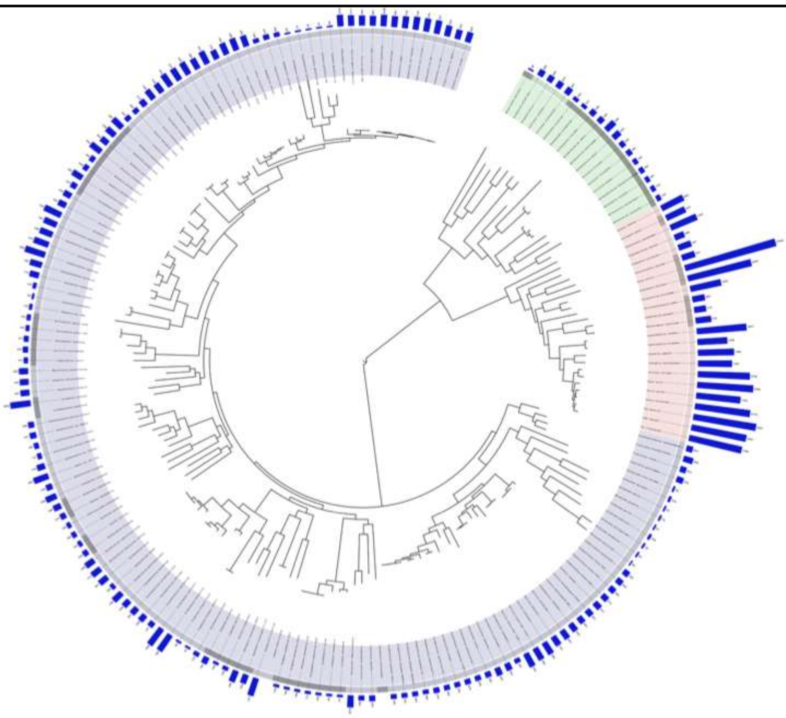
+
Applications: Bioinformatics

- Clustering of Gene Activity from Microarrays
  - Input: Gene activations
  - Output: Gene clusters
  - Discover which genes activate at the same time (appeared in the same cluster)
  - => Help discover relation between functions of dissimilar genes

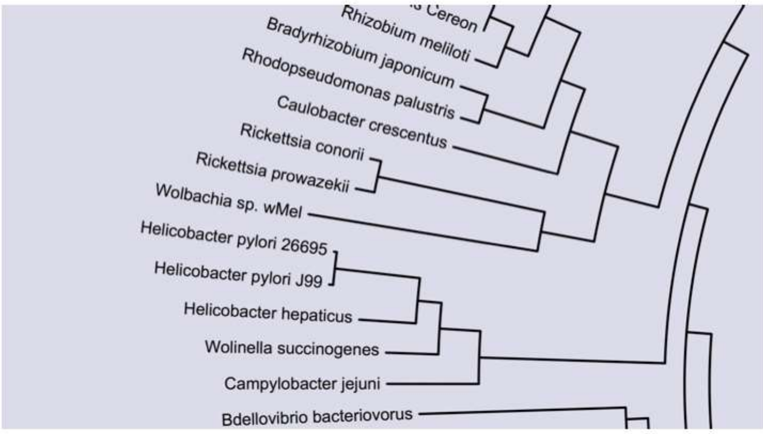
Applications: Bioinformatics for Evolutionary Biology

- Hierarchical Clustering of Genes:
  - Input: Genes of different species
  - Output: Dendrogram relating the genes
  - => Reveals evolutionary history





+ Evolutionary Biology



+ Applications: Finance

- Clustering companies
  - Input: Corporate data (e.g., financials)
  - Output: Clusters of similar companies
- Application:
  - Develop hedging/investment strategies
  - Predict if one stock price will rise or fall

Applications: News Summarisation

- Clustering News articles:
  - Input: One news article per row (E.g., as bag of words)
  - Output: Clusters of similar news articles.
- Application:
  - Get an overview of today's news by one article in each cluster.
    - No redundancy in stories

Application: Video Summarisation EECS Research ☺

- Context:
  - Very many surveillance cameras recording long periods of video.
  - Exhaustively watching all is too time-consuming.
  - Want to get an overview of what happened during the hour/day/week
- Clustering
  - Input: Video frames/clips
  - Output: A category of every frame/clip

+ Application: Video Summarisation EECS Research ☺

- Video Summarisation
  - A few clips quickly summarise the typical events
  - Everything else during the day is "more of the same"

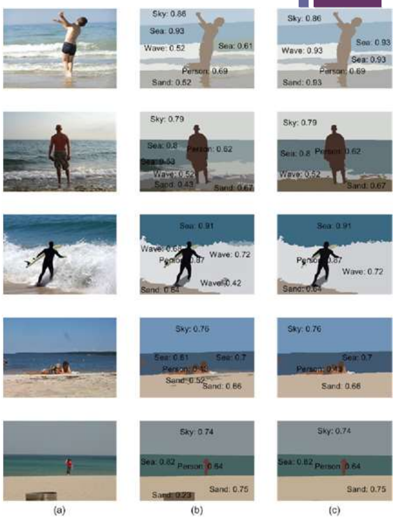
+ Applications: Recommendations (Content-based)

- Input:
  - Descriptions for each product
- Output:
  - Clusters of similar products
- Application:
  - Use discovered product similarity to choose a similar product to recommend

+ Applications: Image Processing

Photoshop filters…

- Input: Image Pixels
- Output: Segmentation (Grouping of all N pixels into K groups)



+ Applications: Politics

Clustering voting records

- Input: Votes of each MP
- Output: Clusters of voting patterns
  - Who tend to vote together / vote against each other.

You should know

- The idea and motivations for clustering
- Be able to sketch algorithms for:
  - K-means, hierarchical clustering, GMM
- Limitations of each algorithm
- Assess which of these algorithm would be suitable for a given problem