Text Clustering, K-Means, Gaussian Mixture Models, Expectation-Maximization, Hierarchical Clustering

Sameer Maskey

Week 3, Sept 19, 2012

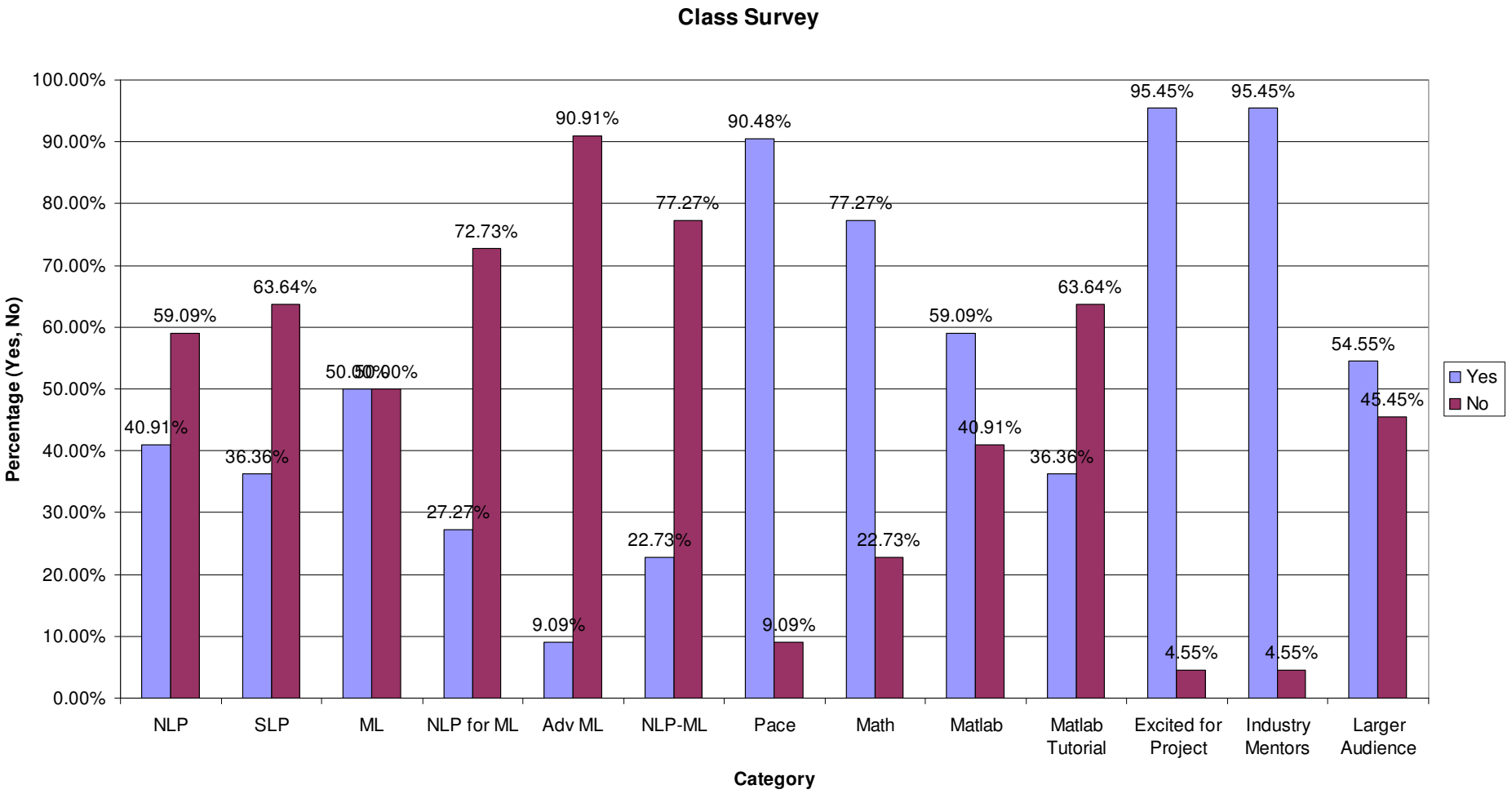
Topics for Today

- Text Clustering
- Gaussian Mixture Models
- K-Means
- Expectation Maximization
- Hierarchical Clustering

Announcement

- Proposal Due tonight (11:59pm) – not graded
- **Feedback by Friday**
- Final Proposal due (11:59pm) next Wednesday
  - 5% of the project grade
  - Email me the proposal with the title
    - "Project Proposal : Statistical NLP for the Web"
- Homework 1 is out
  - Due October 4th (11:59pm) Thursday
- Please use courseworks

Course Initial Survey



Perceptron Algorithm

We are given (x_i, y_i)
Initialize w
Do until converged
    if error$(\mathrm{y}_i, sign(w.x_i)) == TRUE$
        $w \leftarrow w + y_i x_i$
    end if
End do

If predicted class is wrong, subtract or add that point to weight vector

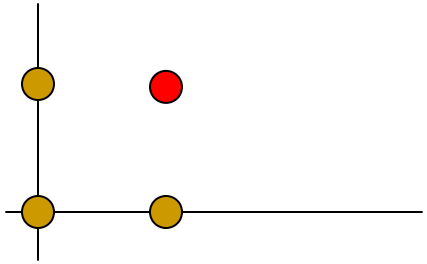
Perceptron (cont.)

Y is prediction based on weights and it's either 0 or 1 in this case

$$y_j(t) = f[w(t).x_j]$$

$$w_i(t+1) = w_i(t) + \alpha(d_j - y_j(t))x_{i,j}$$

Error is either 1, 0 or -1



Input				Initial weights			Output					Error	Correction	Final weights		
Sensor values			Desired output				Per sensor			Sum	Network					
x_0	x_1	x_2	z	w_0	w_1	w_2	c_0	c_1	c_2	s	n	e	d	w_0	w_1	w_2
							$x_0 * w_0$	$x_1 * w_1$	$x_2 * w_2$	$c_0 + c_1 + c_2$	if $s > t$ then 1, else 0	$z - n$	$r * e$	$\Delta(x_0 * d)$	$\Delta(x_1 * d)$	$\Delta(x_2 * d)$
1	0	0	1	0.4	0	0.1	0.4	0	0	0.4	0	1	+0.1	0.5	0	0.1
1	0	1	1	0.5	0	0.1	0.5	0	0.1	0.6	1	0	0	0.5	0	0.1
1	1	0	1	0.5	0	0.1	0.5	0	0	0.5	0	1	+0.1	0.6	0.1	0.1
1	1	1	0	0.6	0.1	0.1	0.6	0.1	0.1	0.8	1	-1	-0.1	0.5	0	0

Example from Wikipedia

Naïve Bayes Classifier for Text

$$P(Y = y_k | X_1, X_2, ..., X_N) = \frac{P(Y=y_k) P(X_1, X_2, .., X_N | Y=y_k)}{\sum_j P(Y=y_j) P(X_1, X_2, .., X_N | Y=y_j)}$$

$$= \frac{P(Y=y_k) \Pi_i P(X_i | Y=y_k)}{\sum_j P(Y=y_j) \Pi_i P(X_i | Y=y_j)}$$

$$Y \leftarrow argmax_{y_k} P(Y = y_k) \Pi_i P(X_i | Y = y_k)$$

Naïve Bayes Classifier for Text

- Given the training data what are the parameters to be estimated?

$$P(Y) \qquad P(X|Y_1) \qquad P(X|Y_2)$$

Diabetes : 0.8
Hepatitis : 0.2

the: 0.001
diabetic : 0.02
blood : 0.0015
sugar : 0.02
weight : 0.018
…

the: 0.001
diabetic : 0.0001
water : 0.0118
fever : 0.01
weight : 0.008
…

$$Y \leftarrow argmax_{y_k} P(Y = y_k)\Pi_i P(X_i|Y = y_k)$$

Data without Labels

Data with corresponding **Human Scores**

- "…is writing a paper"
- "… has flu ☹"
- "… is happy, yankees won!"

0.5
0.1
0.87

Regression

Data with corresponding **Human Class Labels**

- "…is writing a paper"
- "… has flu ☹"
- "… is happy, yankees won!"

SAD
SAD
HAPPY

Perceptron

Naïve Bayes

Fisher's Linear Discriminant

Data with NO corresponding Labels

- "…is writing a paper"
- "… has flu ☹"
- "… is happy, yankees won!"

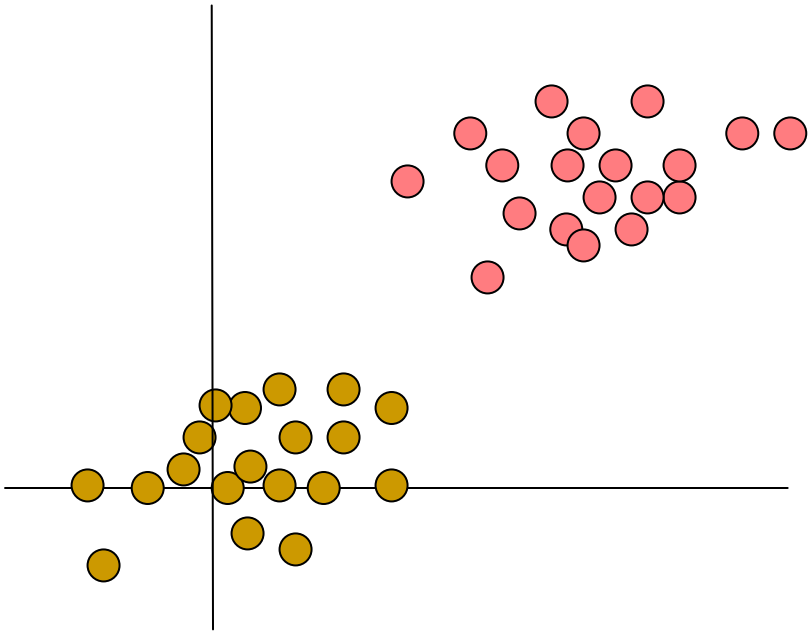
-
-
-

?

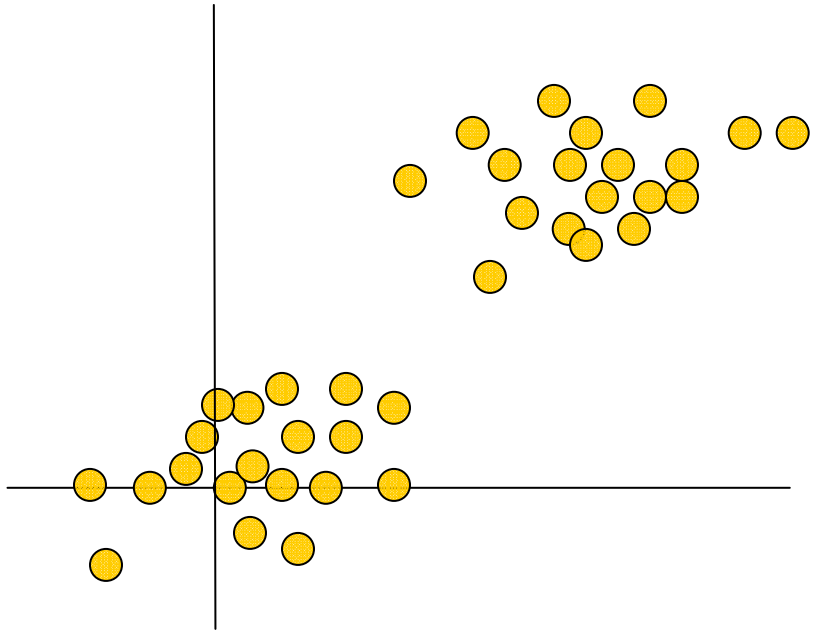
Document Clustering

- Previously we classified Documents into Two Classes
  - Diabetes (Class1) and Hepatitis (Class2)
- We had human labeled data
  - **Supervised learning**
- What if we do not have manually tagged documents
  - Can we still classify documents?
    - Document clustering
    - **Unsupervised Learning**

Classification vs. Clustering

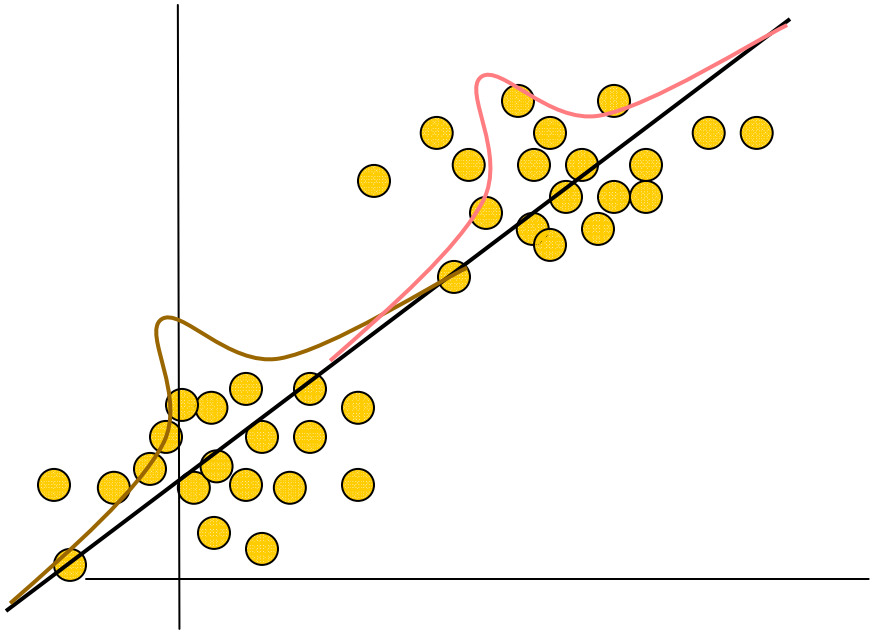


Supervised Training
of Classification Algorithm



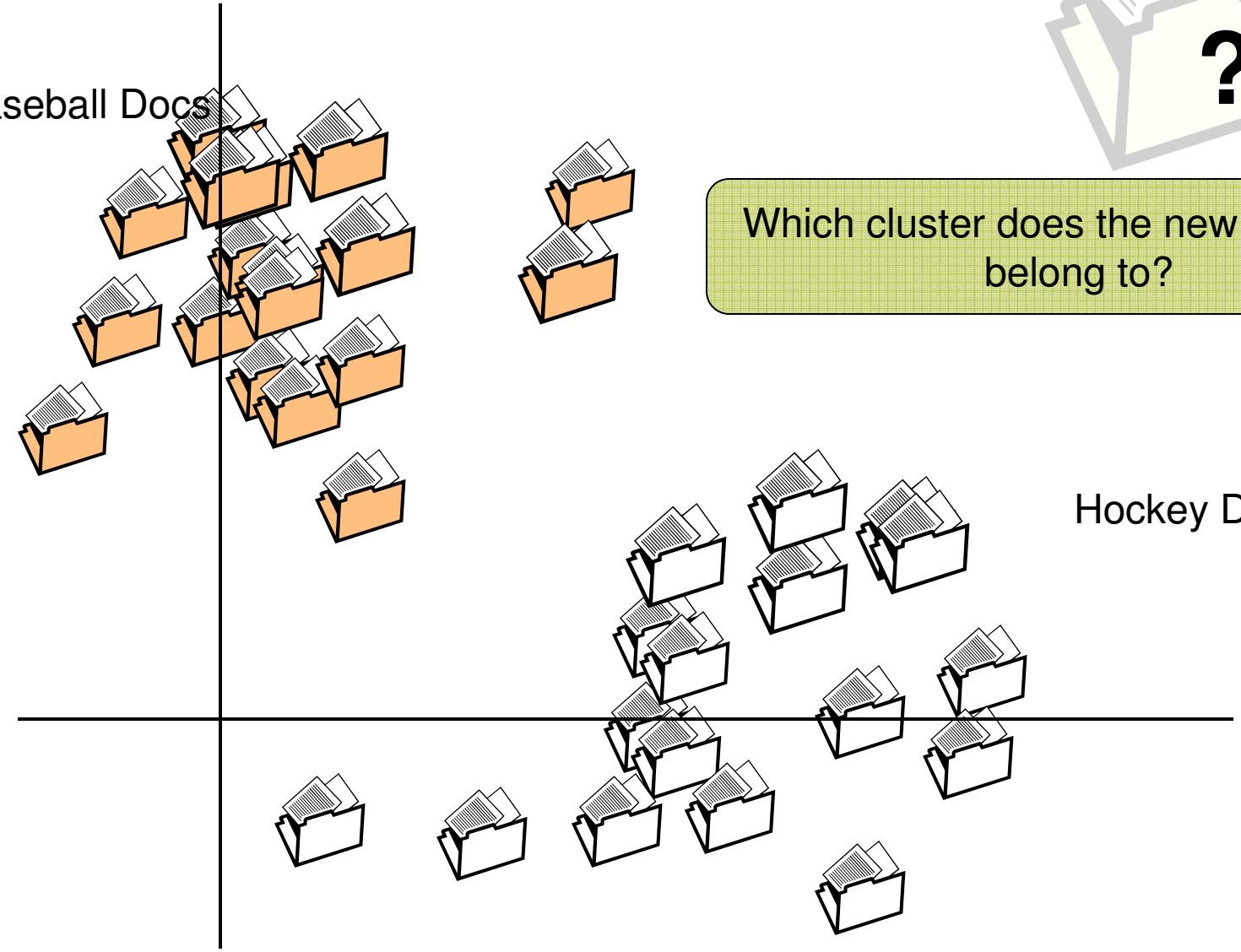
Unsupervised Training
of Clustering Algorithm

Clusters for Classification



Automatically Found Clusters
can be used for Classification

Document Clustering

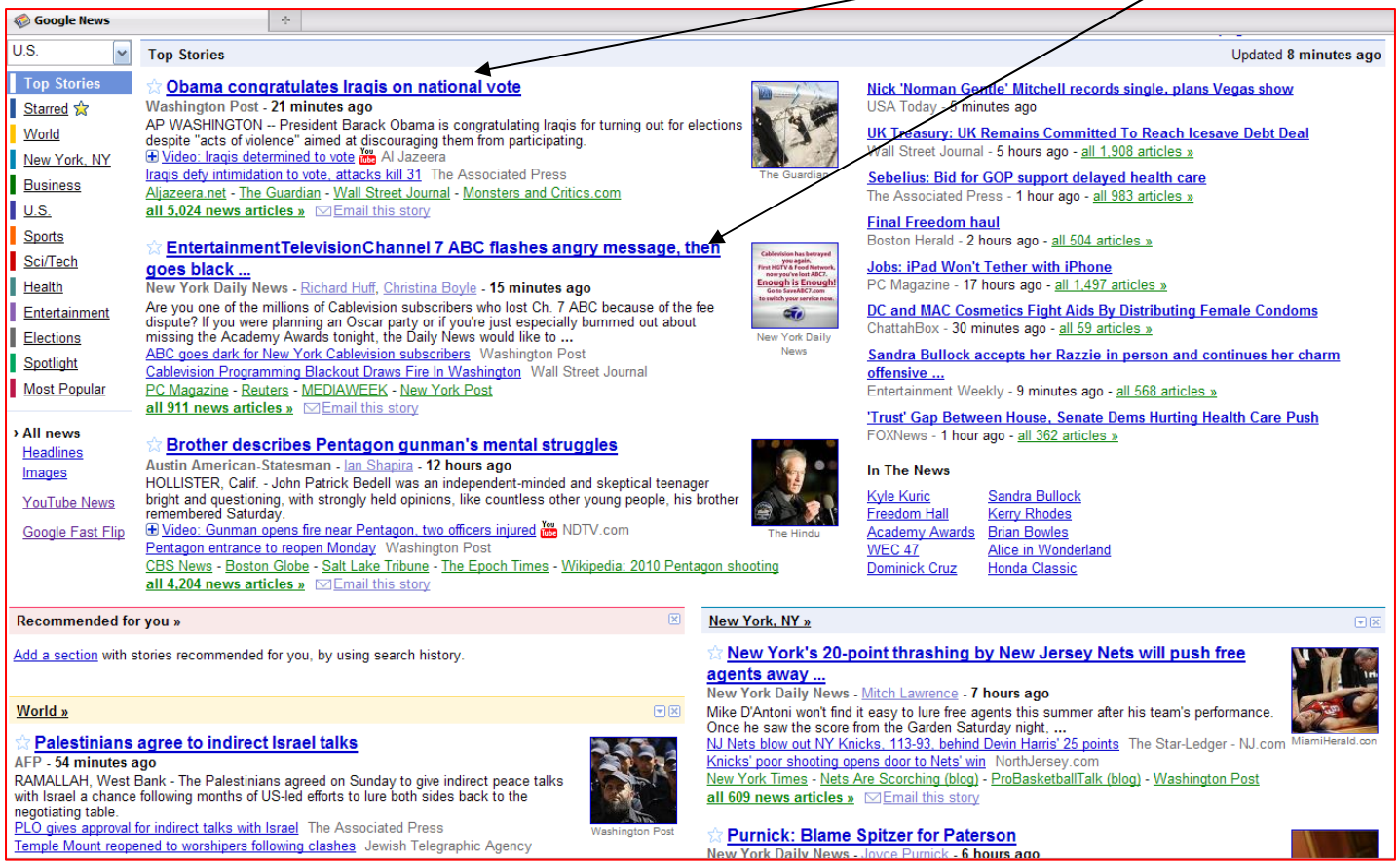


Document Clustering

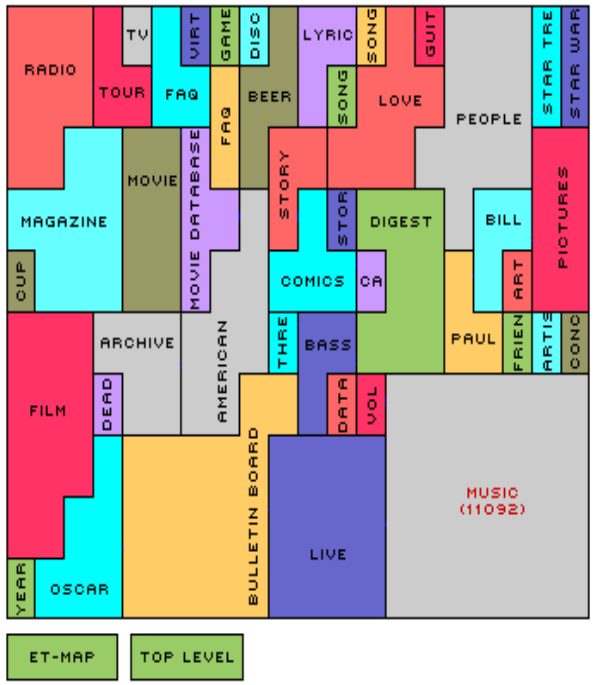
- Cluster the documents in ‘N’ clusters/categories
- For classification we were able to estimate parameters using labeled data
  - Perceptrons – find the parameters that decide the separating hyperplane
  - Naïve Bayes – count the number of times word occurs in the given class and normalize
- Not evident on how to find separating hyperplane when no labeled data available
- Not evident how many classes we have for data when we do not have labels

Document Clustering Application

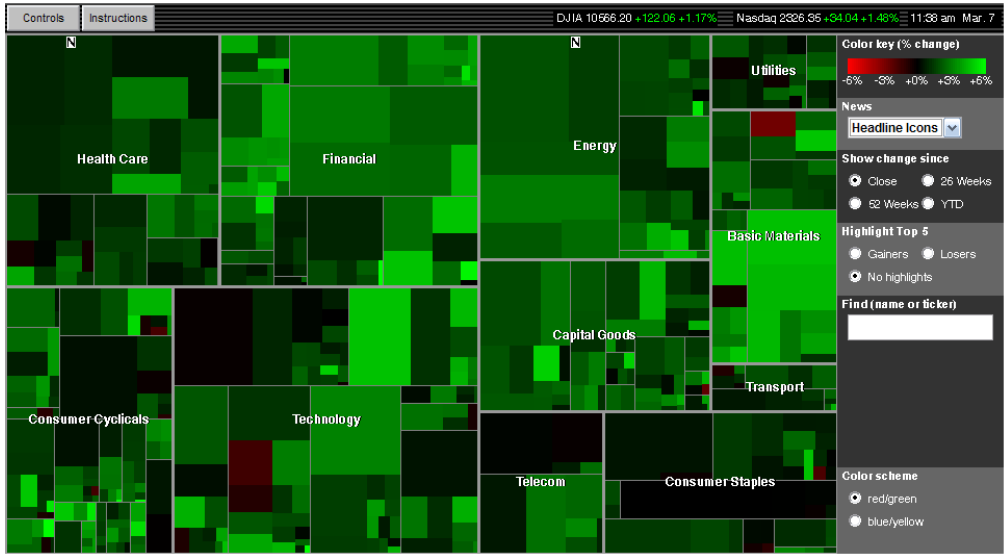
- Even though we do not know human labels automatically induced clusters could be useful



Document Clustering Application



A Map of Yahoo!, Mappa.Mundi Magazine, February 2000.



Map of the Market with Headlines Smartmoney [2]

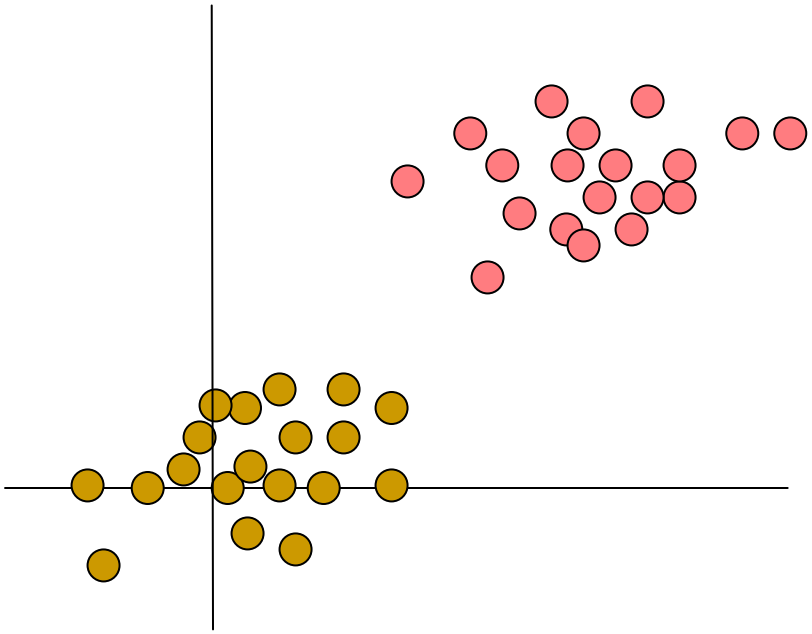
How to Cluster Documents with No Labeled Data?

- Treat cluster IDs or class labels as hidden variables
- Maximize the likelihood of the unlabeled data
- Cannot simply count for MLE as we do not know which point belongs to which class
  - User Iterative Algorithm such as K-Means, EM

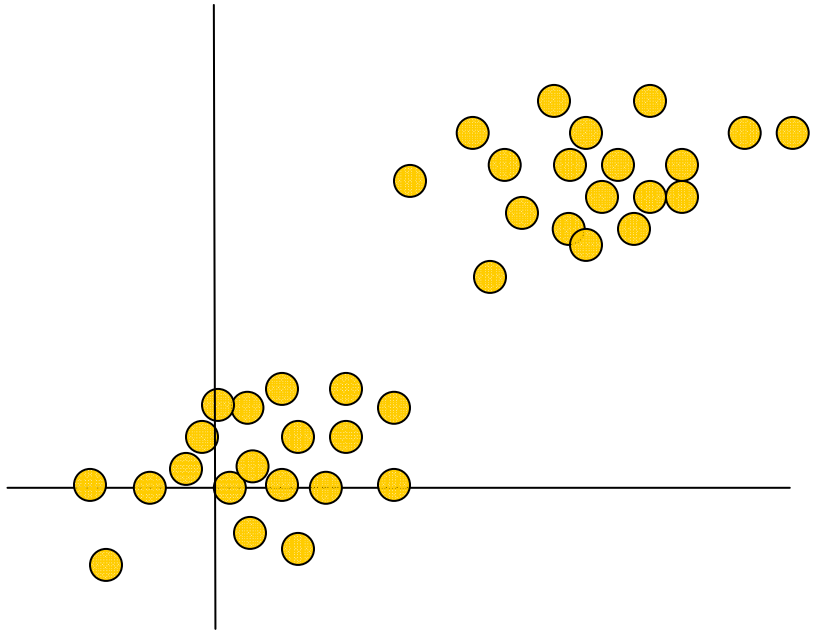
Hidden Variables?
What do we mean by this?

Hidden vs. Observed Variables

Assuming our observed data is in R2



How many observed variables?



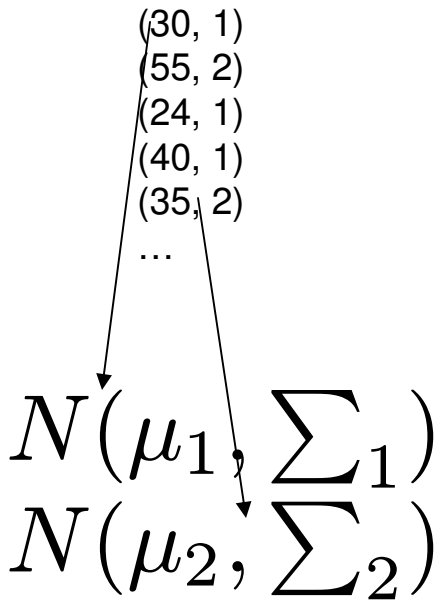
How many observed variables?

How many hidden variables?

Clustering

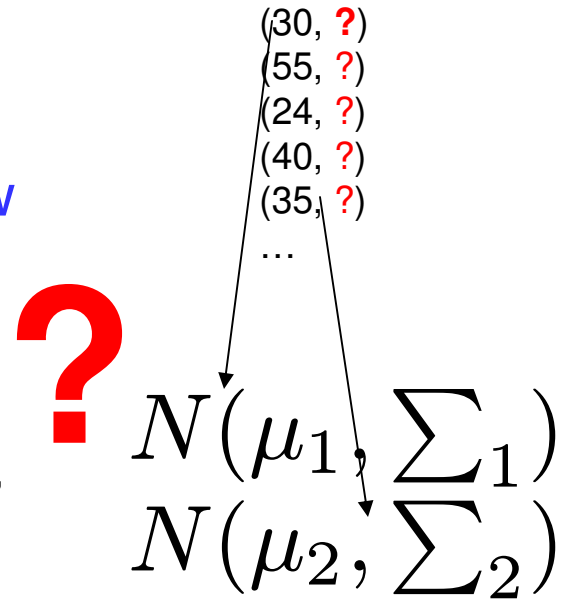
- If we have data with labels

Find out μ_i and $\sum_i$ from data for both classes

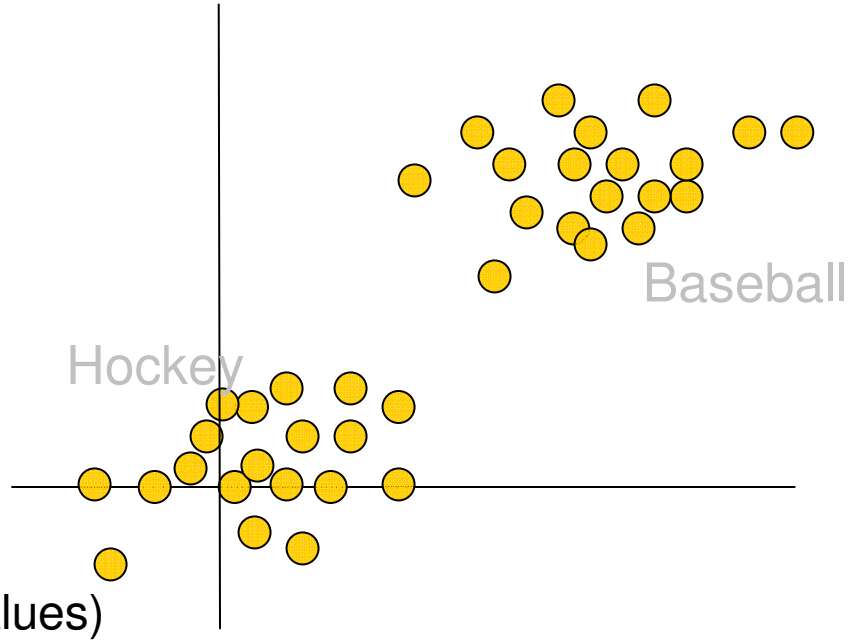


- If we have data with NO labels but know data comes from 2 classes

Find out μ_i and $\sum_i$ from data for both classes



K-Means in Words



- Parameters to estimate for K classes
- Let us assume we can model this data
- with mixture of two Gaussians
- Start with 2 Gaussians (initialize mu values)
- Compute distance of each point to the mu of 2 Gaussians and assign it to the closest Gaussian (class label (Ck))
- Use the assigned points to recompute mu for 2 Gaussians

K-Means Clustering

Let us define Dataset in D dimension$\{x_1, x_2, ..., x_N\}$

We want to cluster the data in Kclusters

Let μ_k be D dimension vector representing cluster$K$

Let us define r_{nk} for each x_n such that
$r_{nk} \in \{0, 1\}$ where $k = 1, ..., K$ and
$r_{nk} = 1$ if x_n is assigned to cluster k

Distortion Measure

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2$$

Represents sum of squares of distances to mu_k from each data point

We want to minimize J

Estimating Parameters

- We can estimate parameters by doing 2 step iterative process
  - Minimize J with respect to r_{nk}
    - Keep μ_k fixed

Step 1

  - Minimize J with respect to μ_k
    - Keep r_{nk} fixed

Step 2

- Minimize J with respect to r_{nk}
  - Keep μ_k fixed

Step 1

- Optimize for each n separately by choosing r_{nk} for k that gives minimum $||x_n - r_{nk}||^2$

$$r_{nk} = 1 \text{ if } k = argmin_j ||x_n - \mu_j||^2$$
$$= 0 \text{ otherwise}$$

- Assign each data point to the cluster that is the closest
- Hard decision to cluster assignment

- Minimize J with respect to μ_k
  - Keep r_{nk} fixed

Step 2

- J is quadratic in μ_k. Minimize by setting derivative w.rt. μ_k to zero

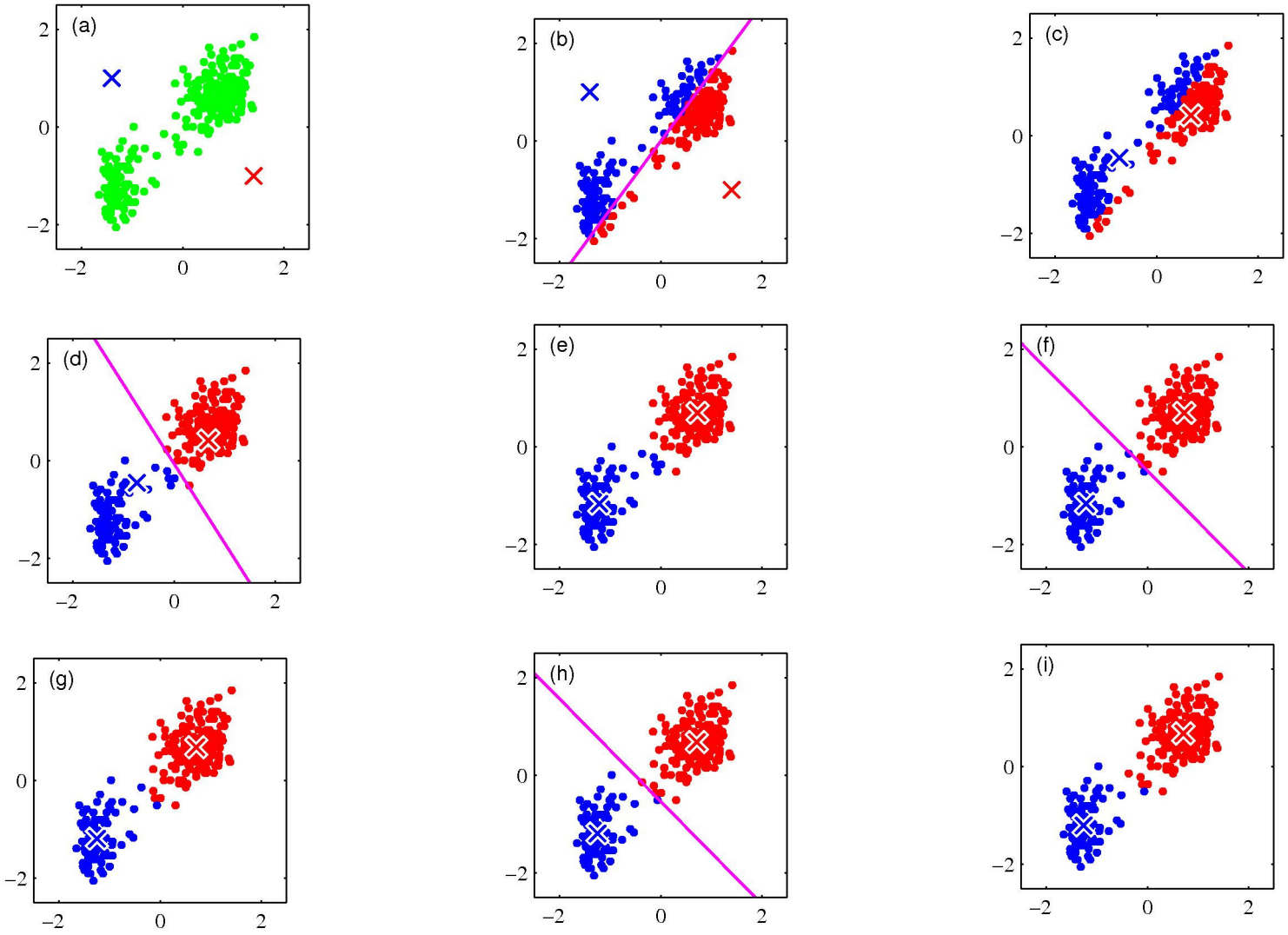
$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$

- Take all the points assigned to cluster K and re-estimate the mean for cluster K

Document Clustering with K-means

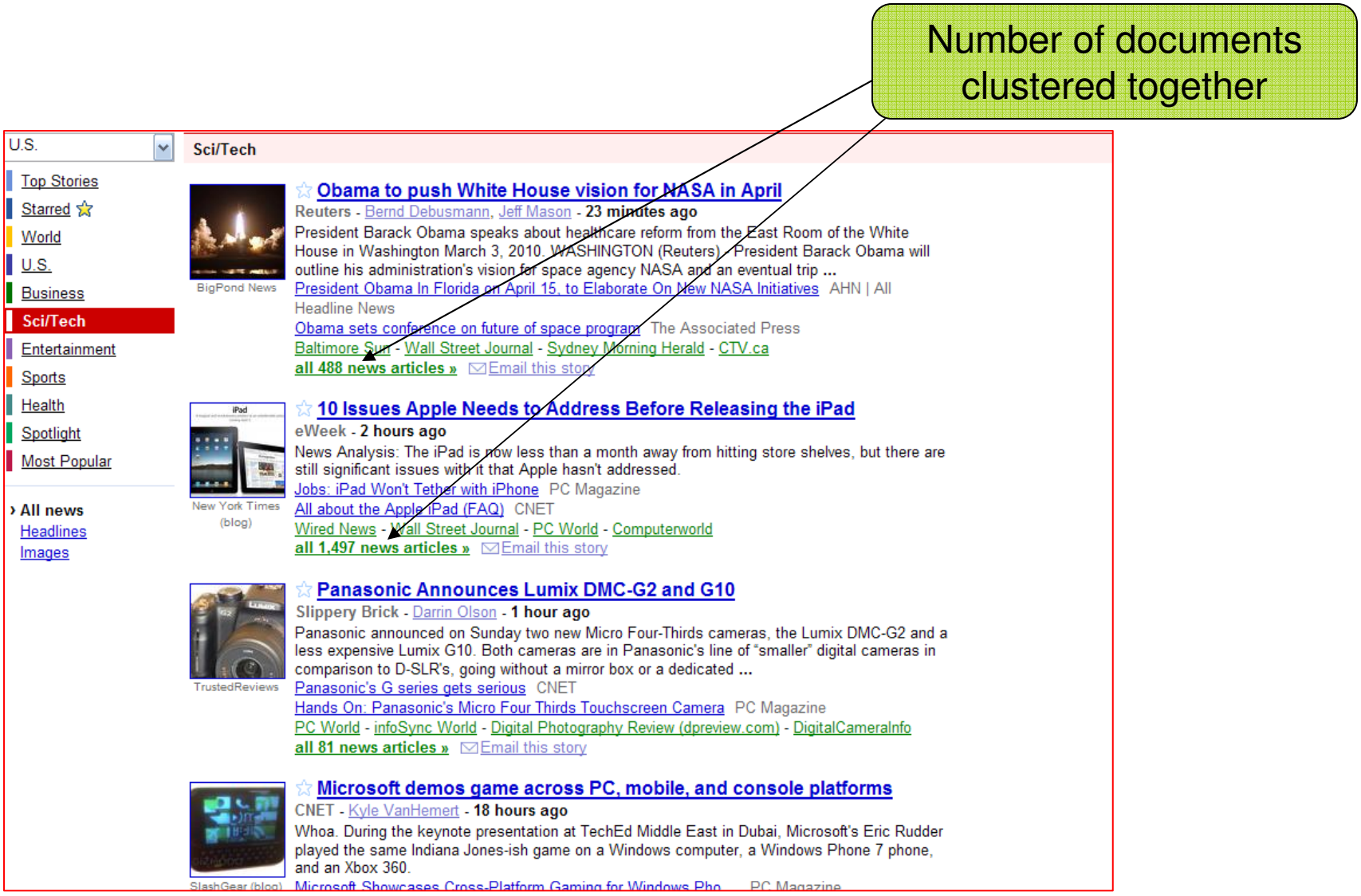
- Assuming we have data with no labels for Hockey and Baseball data
- We want to be able to categorize a new document into one of the 2 classes (K=2)
- We can extract represent document as feature vectors
  - Features can be word id or other NLP features such as POS tags, word context etc (D=total dimension of Feature vectors)
  - N documents are available
- Randomly initialize 2 class means
- Compute square distance of each point (x_n)(D dimension) to class means (μ_k)
- Assign the point to K for which μ_k is lowest
- Re-compute μ_k and re-iterate

K-Means Example



K-means algorithm Illustration [1]

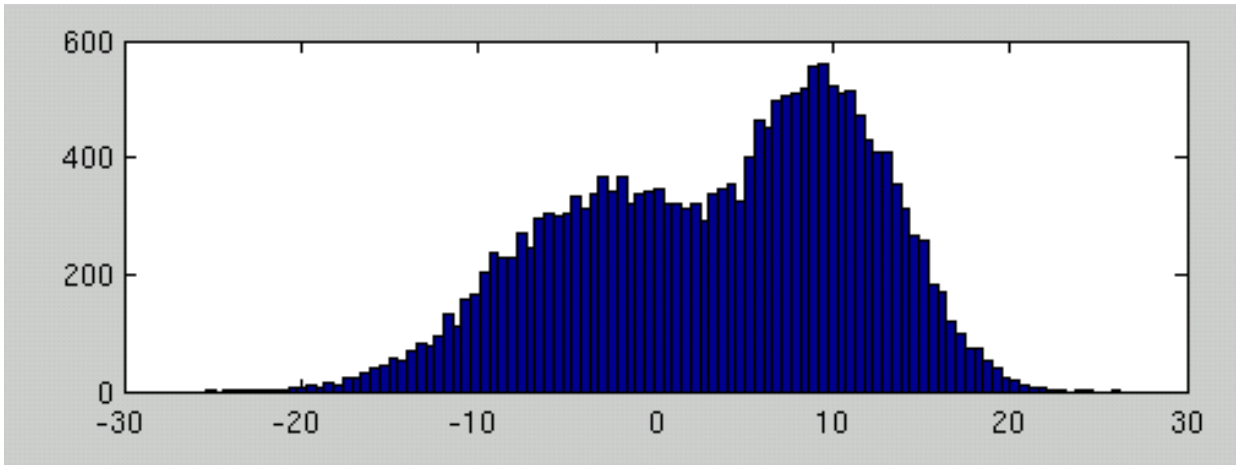
Clusters

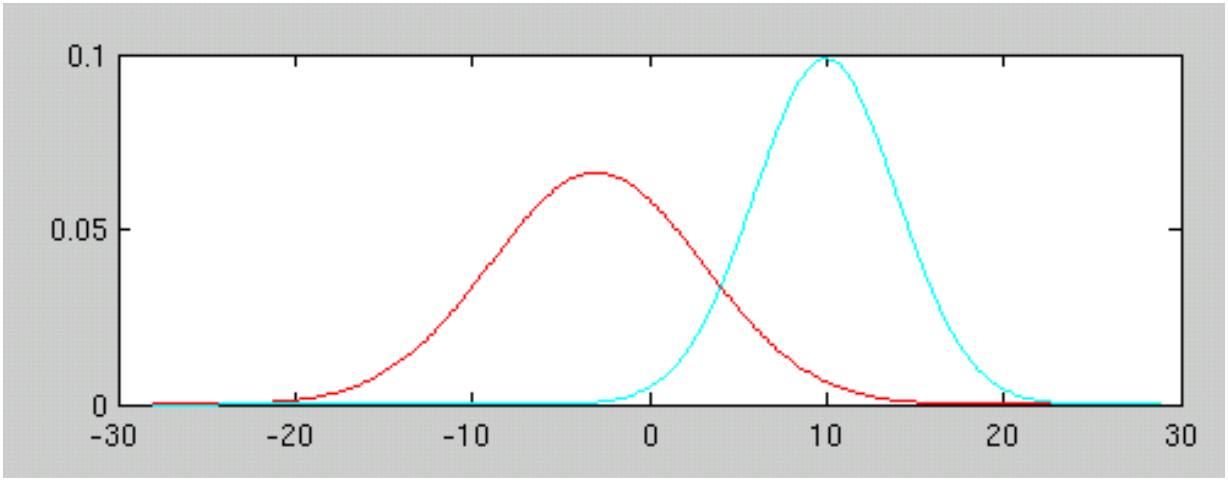

Number of documents clustered together
U.S.
Top Stories
Starred
World
U.S.
Business
Sci/Tech
Entertainment
Sports
Health
Spotlight
Most Popular
› All news
Headlines
Images
Sci/Tech
Obama to push White House vision for NASA in April
Reuters - Bernd Debusmann, Jeff Mason - 23 minutes ago
President Barack Obama speaks about healthcare reform from the East Room of the White House in Washington March 3, 2010. WASHINGTON (Reuters) - President Barack Obama will outline his administration's vision for space agency NASA and an eventual trip ...
President Obama In Florida on April 15, to Elaborate On New NASA Initiatives AHN | All Headline News
Obama sets conference on future of space program The Associated Press
Baltimore Sun - Wall Street Journal - Sydney Morning Herald - CTV.ca
all 488 news articles » Email this story
BigPond News
10 Issues Apple Needs to Address Before Releasing the iPad
eWeek - 2 hours ago
News Analysis: The iPad is now less than a month away from hitting store shelves, but there are still significant issues with it that Apple hasn't addressed.
Jobs: iPad Won't Tether with iPhone PC Magazine
All about the Apple iPad (FAQ) CNET
Wired News - Wall Street Journal - PC World - Computerworld
all 1,497 news articles » Email this story
New York Times (blog)
Panasonic Announces Lumix DMC-G2 and G10
Slippery Brick - Darrin Olson - 1 hour ago
Panasonic announced on Sunday two new Micro Four-Thirds cameras, the Lumix DMC-G2 and a less expensive Lumix G10. Both cameras are in Panasonic's line of "smaller" digital cameras in comparison to D-SLR's, going without a mirror box or a dedicated ...
Panasonic's G series gets serious CNET
Hands On: Panasonic's Micro Four Thirds Touchscreen Camera PC Magazine
PC World - infoSync World - Digital Photography Review (dpreview.com) - DigitalCameraInfo
all 81 news articles » Email this story
TrustedReviews
Microsoft demos game across PC, mobile, and console platforms
CNET - Kyle VanHemert - 18 hours ago
Whoa. During the keynote presentation at TechEd Middle East in Dubai, Microsoft's Eric Rudder played the same Indiana Jones-ish game on a Windows computer, a Windows Phone 7 phone, and an Xbox 360.
SlashGear (blog)

Hard Assignment to Clusters

- K-means algorithm assigns each point to the closest cluster
  - Hard decision
  - Each data point affects the mean computation equally
- How does the points almost equidistant from 2 clusters affect the algorithm?
- Soft decision?
  - Fractional counts?

Gaussian Mixture Models (GMMs)

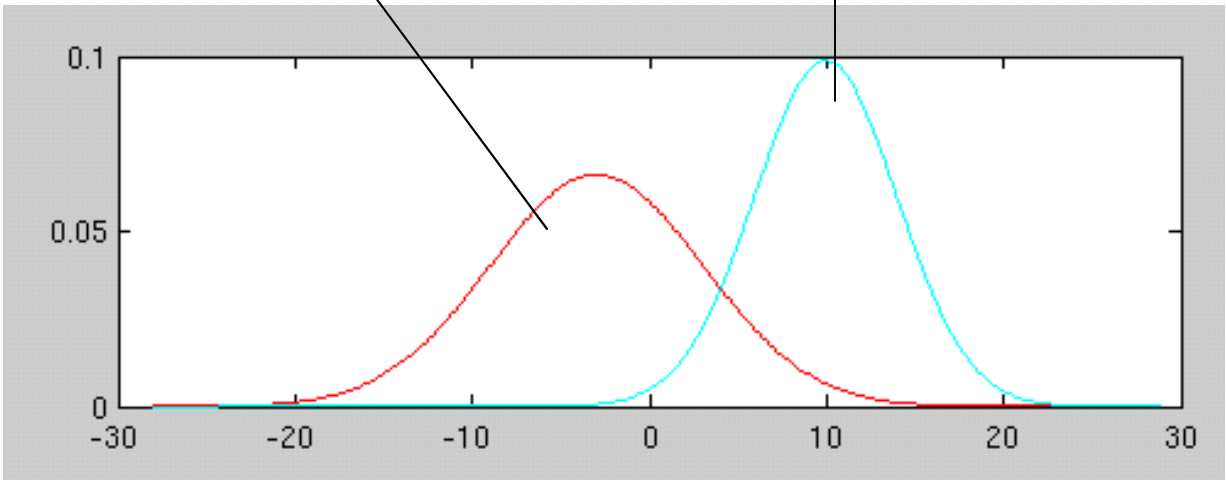




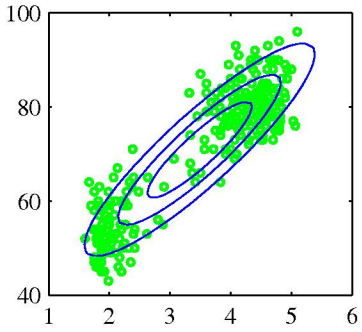
Mixtures of 2 Gaussians

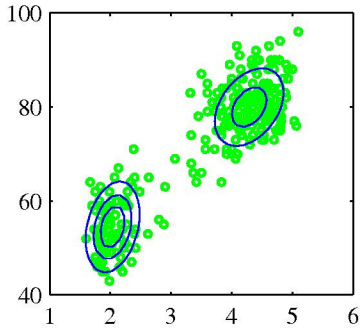
$$\mathrm{P(x)} = \pi\mathcal{N}(x|\mu_1, \sum_1) + (1-\pi)\mathcal{N}(x|\mu_2, \sum_2)$$

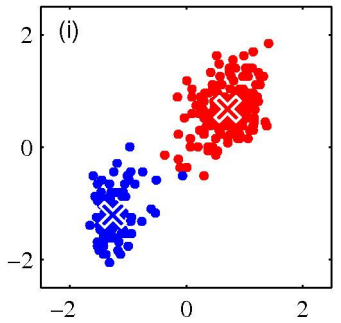
GMM with 2 gaussians



Mixture Models





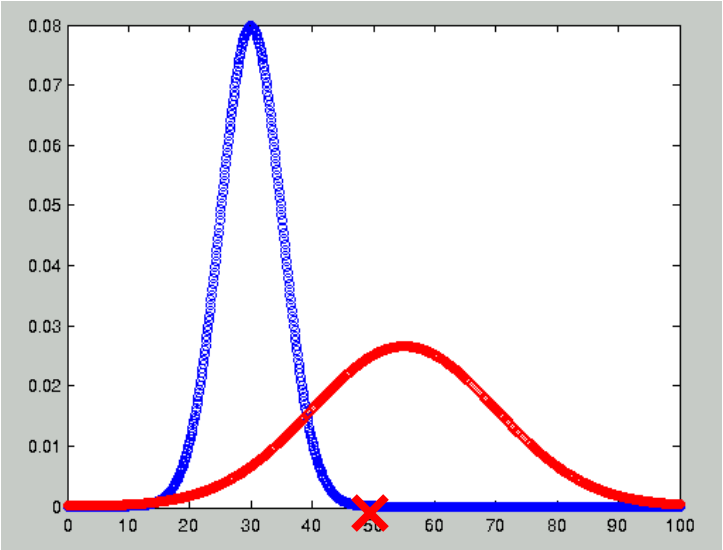


Mixture of Gaussians [1]

- 1 Gaussian may not fit the data
- 2 Gaussians may fit the data better
- Each Gaussian can be a class category
- When labeled data not available we can treat class category as hidden variable

Mixture Model Classifier

- Given a new data point find out posterior probability from each class



$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$p(y=1|x) \propto \mathcal{N}(x|\mu_1, \sum_1)p(y=1)$$

Cluster ID/Class Label as Hidden Variables

$$p(x) = \sum_z p(x, z) = \sum_z p(z)p(x|z)$$

- We can treat class category as hidden variable z
- Z is K-dimensional binary random variable in which z_k = 1 and 0 for other elements

$$z = [00100...]$$

$$\mathrm{p(z)} = \prod_{k=1}^{K} \pi_k^{z_k}$$

- Also, sum of priors sum to 1 $\sum_{k=1}^{K} \pi_k = 1$

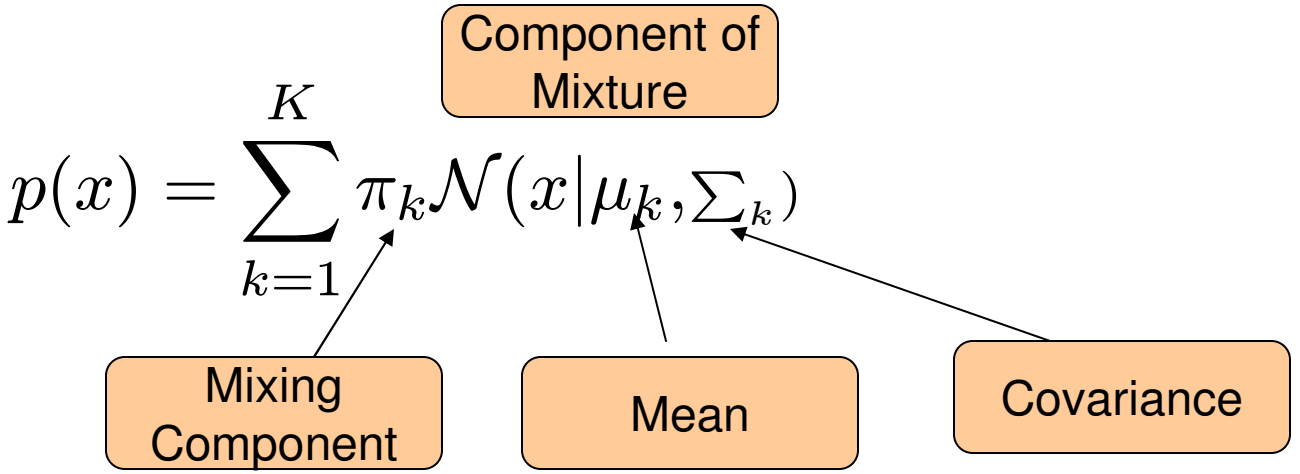
- Conditional distribution of x given a particular z can be written as

$$P(x|z_k = 1) = \mathcal{N}(x|\mu_k, \Sigma_k)$$

$$P(x|\bar{z}) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}$$

Mixture of Gaussians with Hidden Variables

$$p(x) = \sum_z p(x, z) = \sum_z p(z)p(x|z)$$



$$p(x) = \sum_{k=1}^{K} \pi_k \frac{1}{(2\pi)^{D/2}\sqrt{(|\Sigma_k|}} exp(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k))$$

- Mixture models can be linear combinations of other distributions as well
- Mixture of binomial distribution for example

Conditional Probability of Label Given Data

- Mixture model with parameters mu, sigma and prior can represent the parameter
- We can maximize the data given the model parameters to find the best parameters
- If we know the best parameters we can estimate

$$\circ(z_k) \acute{} \ \mathrm{p}(\mathrm{z}_k = 1|x) = \frac{p(z_k{=}1)p(x|z_k{=}1)}{\sum_{j=1}^{K} p(z_j{=}1)p(x|z_j{=}1)}$$

$$= \frac{\pi_k \mathcal{N}(x|\mu_k, \sum_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \sum_j)}$$

This essentially gives us probability of class given the data i.e label for the given data point

Maximizing Likelihood

- If we had labeled data we could maximize likelihood simply by counting and normalizing to get mean and variance of Gaussians for the given classes

$$l = \sum_{n=1}^{N} log\ p(x_n, y_n | \pi, \mu, \sum)$$

$$l = \sum_{n=1}^{N} log\ \pi_{y_n} \mathcal{N}(x_n | \mu_{y_n}, \sum_{y_n})$$

- If we have two classes C1 and C2
  - Let's say we have a feature x
    - x = number of words 'field'
  - And class label (y)
    - y = 1 hockey or 2 baseball documents

(30, 1)
(55, 2)
(24, 1)
(40, 1)
(35, 2)
…

$N(\mu_1, \sum_1)$
$N(\mu_2, \sum_2)$

Find out μ_i and $\sum_i$ from data for both classes

Maximizing Likelihood for Mixture Model with Hidden Variables

- For a mixture model with a hidden variable representing 2 classes, log likelihood is

$$l = \sum_{n=1}^{N} logp(x_n|\pi,\mu,\Sigma)$$

$$l = \sum_{n=1}^{N} log \sum_{y=0}^{1} \mathcal{N}(x_n, y|\pi,\mu,\Sigma)$$

$$= \sum_{n=1}^{N} log\ (\pi_0 \mathcal{N}(x_n|\mu_0,\Sigma_0) + \pi_1 \mathcal{N}(x_n|\mu_1,\Sigma_1))$$

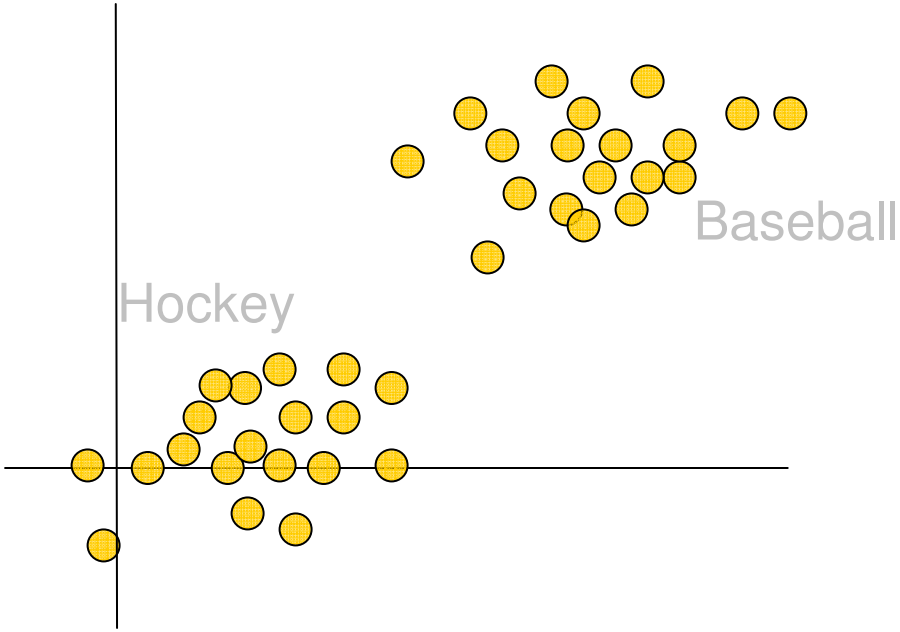
Log-likelihood for Mixture of Gaussians

$$log\ p(X|\pi,\mu,\sum) = \sum_{n=1}^{N} log\ (\sum_{k=1}^{k} \pi_k \mathcal{N}(x|\mu_k,\sum_k))$$

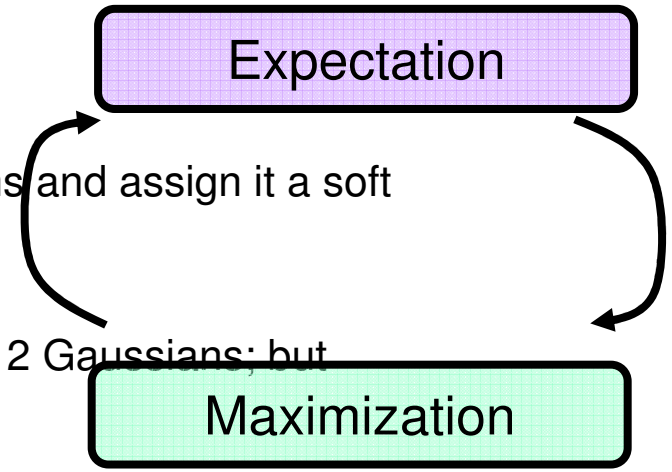
- We want to find maximum likelihood of the above log-likelihood function to find the best parameters that maximize the data given the model
- We can again do iterative process for estimating the log-likelihood of the above function
  - This 2-step iterative process is called Expectation-Maximization

Explaining Expectation Maximization

- EM is like fuzzy K-means
- Parameters to estimate for K classes
- Let us assume we can model this data with mixture of two Gaussians (K=2)



- Start with 2 Gaussians (initialize mu and sigma values)
- Compute distance of each point to the mu of 2 Gaussians and assign it a soft class label (Ck)
- Use the assigned points to recompute mu and sigma for 2 Gaussians; but weight the updates with soft labels



Expectation Maximization

An expectation-maximization (EM) algorithm is used in statistics for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved hidden variables.

EM alternates between performing an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and a maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the E step. The parameters found on the M step are then used to begin another E step, and the process is repeated.

The EM algorithm was explained and given its name in a classic 1977 paper by A. Dempster and D. Rubin in the Journal of the Royal Statistical Society.

Estimating Parameters

$$\circ(z_{nk}) = E(z_{nk}|x_n) = p(z_k = 1|x_n)$$

- E-Step

$$\circ(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \sum_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \sum_j)}$$

Estimating Parameters

- M-step

$$\mu'_k = \frac{1}{N_k} \sum_{n=1}^{N} \circ(z_{nk}) x_n$$

$$\sum'_k = \frac{1}{N_k} \sum_{n=1}^{N} \circ(z_{nk})(x_n - \mu'_k)(x_n - \mu'_k)^T$$

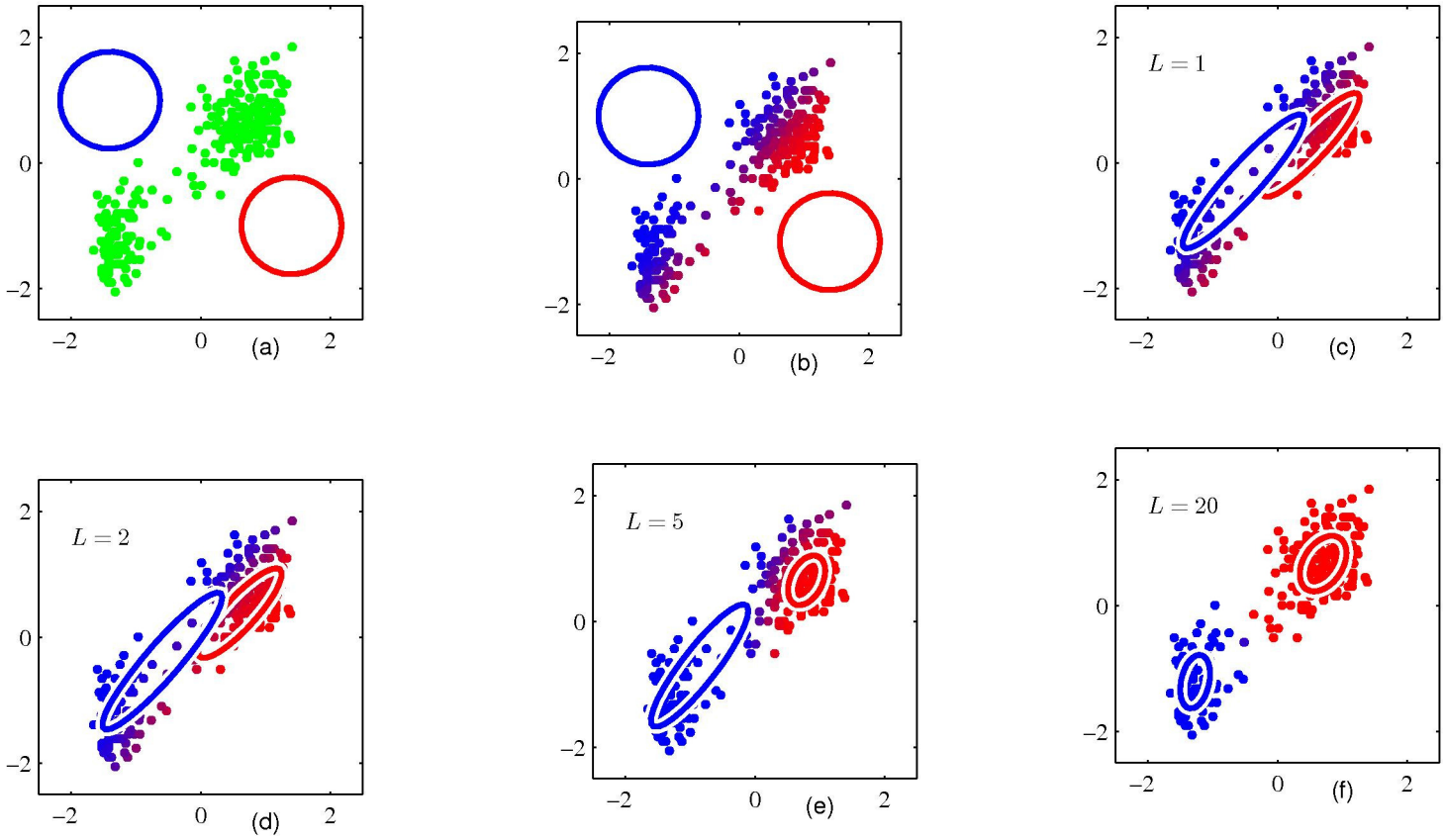
$$\pi'_k = \frac{N_k}{N}$$

$$\text{where } N_k = \sum_{n=1}^{N} \circ(z_{nk})$$

- Iterate until convergence of log likelihood

$$log\ p(X|\pi, \mu, \sum) = \sum_{n=1}^{N} log\ (\sum_{k=1}^{k} \mathcal{N}(x|\mu_k, \sum_k))$$

EM Iterations



EM iterations [1]

Clustering Documents with EM

- Clustering documents requires representation of documents in a set of features
  - Set of features can be bag of words model
  - Features such as POS, word similarity, number of sentences, etc
- Can we use mixture of Gaussians for any kind of features?
- How about mixture of multinomial for document clustering?
- How do we get EM algorithm for mixture of multinomial?

Clustering Algorithms

- We just described two kinds of clustering algorithms
  - K-means
  - Expectation Maximization
- Expectation-Maximization is a general way to maximize log likelihood for distributions with hidden variables
  - For example, EM for HMM, state sequences were hidden
- For document clustering other kinds of clustering algorithm exists

Hierarchical Clustering

- Build a binary tree that groups similar data in iterative manner
- K-means
  - distance of data point to center of the gaussian
- EM
  - Posterior of data point w.r.t to the gaussian
- Hierarchical
  - Similarity : ?
    - Similarity across **groups** of data

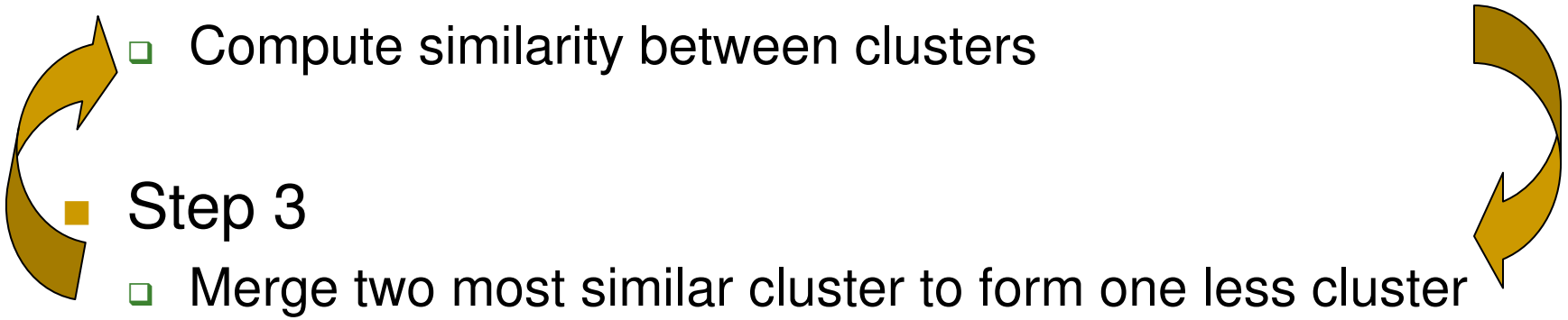
Types of Hierarchical Clustering

- **Agglomerative (bottom-up):**
  - Assign each data point as one cluster
  - Iteratively combine sub-clusters
  - Eventually, all data points is a part of 1 cluster
- **Divisive (top-down):**
  - Assign all data points to the same cluster.
  - Eventually each data point forms its own cluster

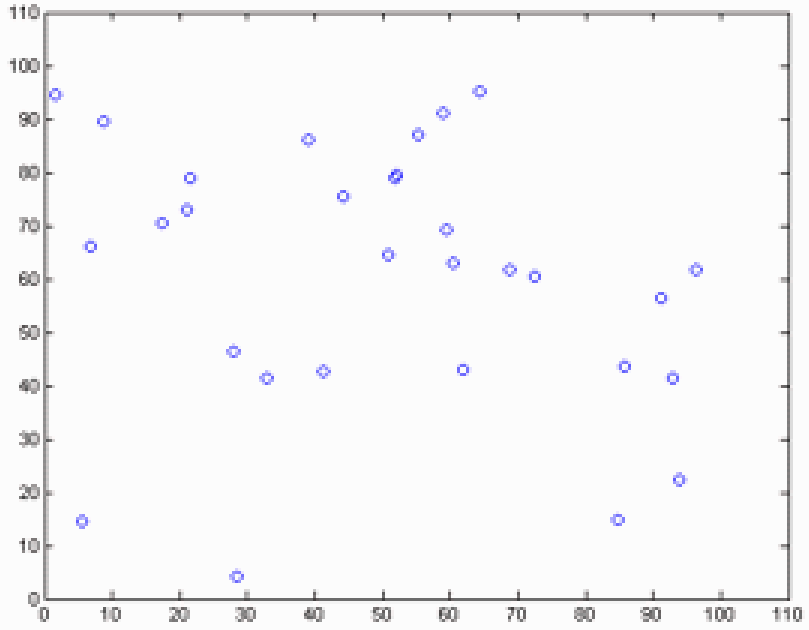
One advantage :
Do not need to define K, number of clusters before we begin clustering

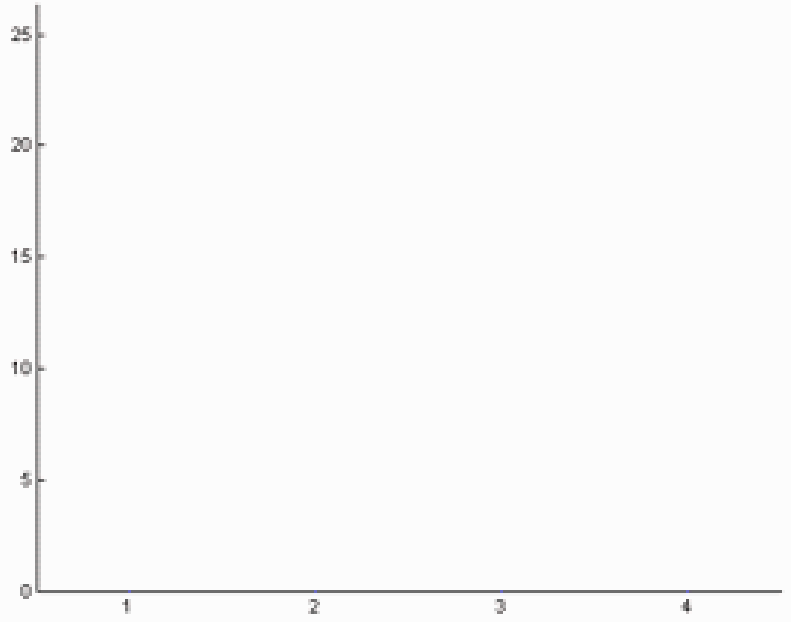
Hierarchical Clustering Algorithm

- Step 1
  - Assign each data point to its own cluster
- Step 2
  - Compute similarity between clusters
- Step 3
  - Merge two most similar cluster to form one less cluster



Hierarchical Clustering Demo



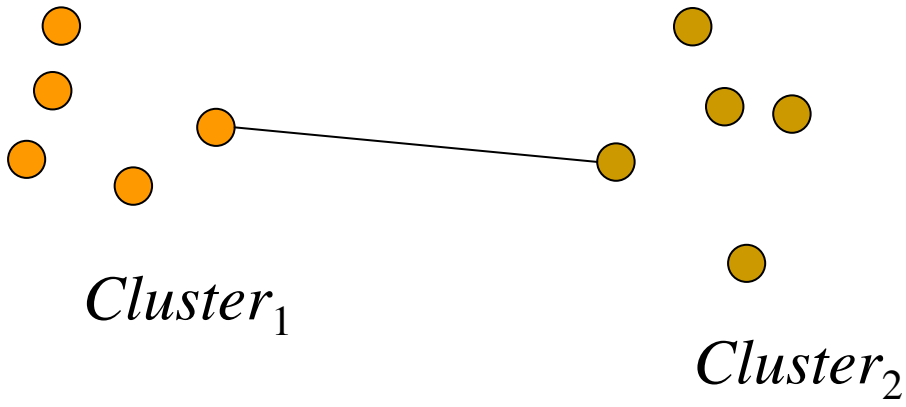


Animation source [4]

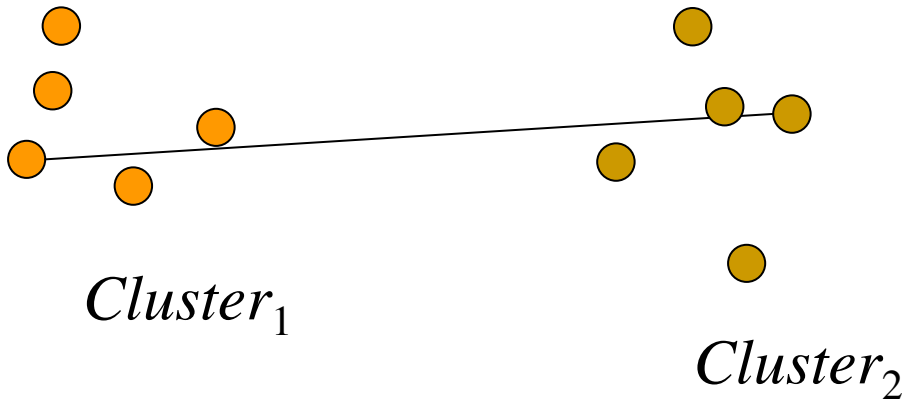
Similar Clusters?

- How do we compute similar clusters?
  - Distance between 2 points in the clusters?
  - Distance from means of two clusters?
  - Distance between two closest points in the clusters?
- Different similarity metric could produce different types of cluster
- Common similarity metric used
  - Single Linkage
  - Complete Linkage
  - Average Group Linkage

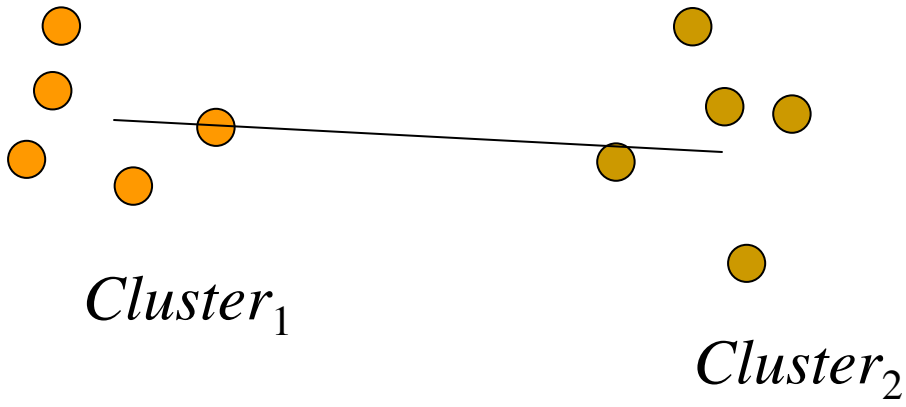
Single Linkage



Complete Linkage



Average Group Linkage



Hierarchical Cluster for Documents

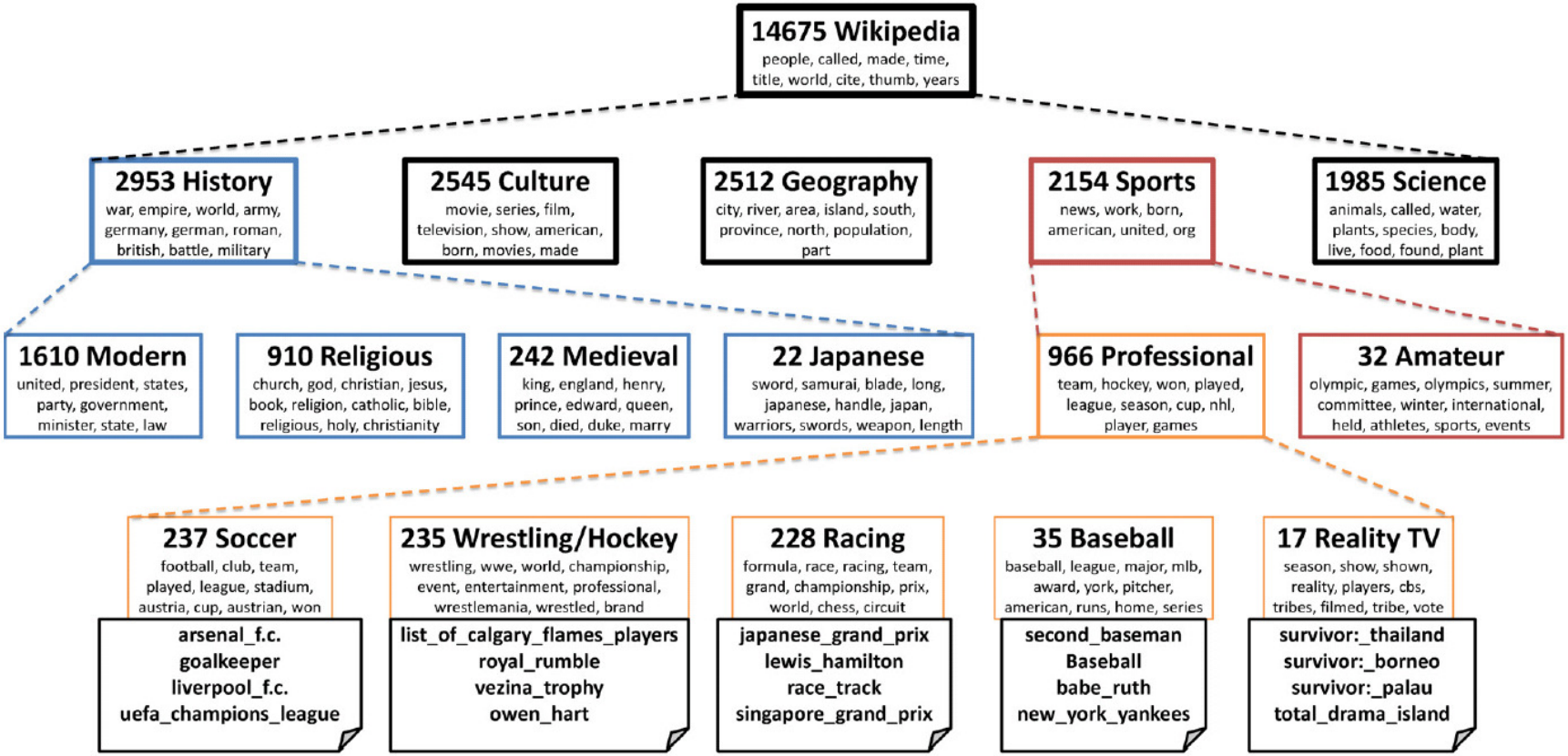


Figure : [Ho, Qirong, et. al]

Hierarchical Document Clusters

- Highlevel multi view of the corpus
- Taxonomy useful for various purposes
  - Q&A related to a subtopic
  - Finding broadly important topics
  - Recursive drill down on topics
  - Filter irrelevant topics

Summary

- Unsupervised clustering algorithms
  - K-means
  - Expectation Maximization
  - Hierarchical clustering
- EM is a general algorithm that can be used to estimate maximum likelihood of functions with hidden variables
- Similarity Metric is important when clustering segments of text

References

- [1] Christopher Bishop, “Pattern Recognition and Machine Learning,” 2006
- [2] http://www.smartmoney.com/map-of-the-market/
- [3] Ho, Qirong, et. al, Document Hierarchies from Text and Links, 2012