

Statistical Methods for NLP

Document and Topic Clustering, K-Means,
Mixture Models, Expectation-Maximization

Sameer Maskey

Week 8, March 2010

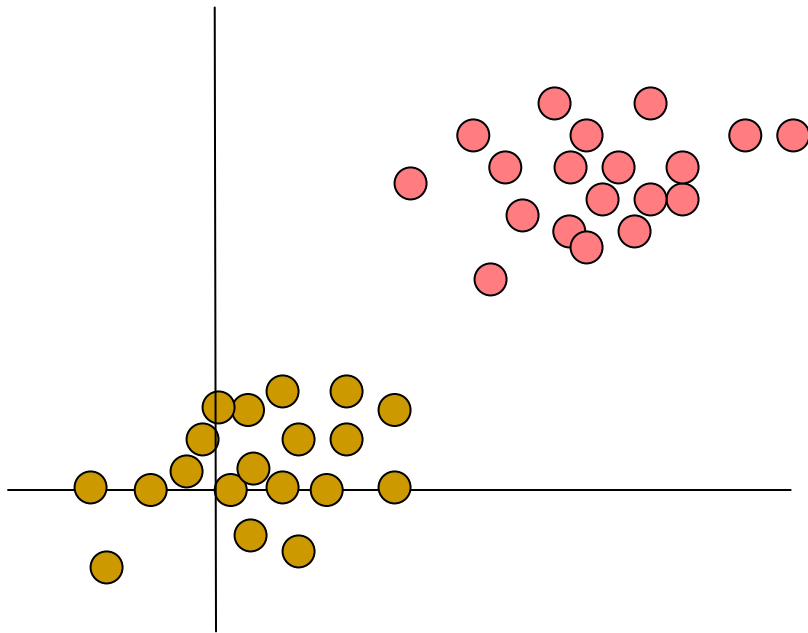
Topics for Today

- Document, Topic Clustering
- K-Means
- Mixture Models
- Expectation Maximization

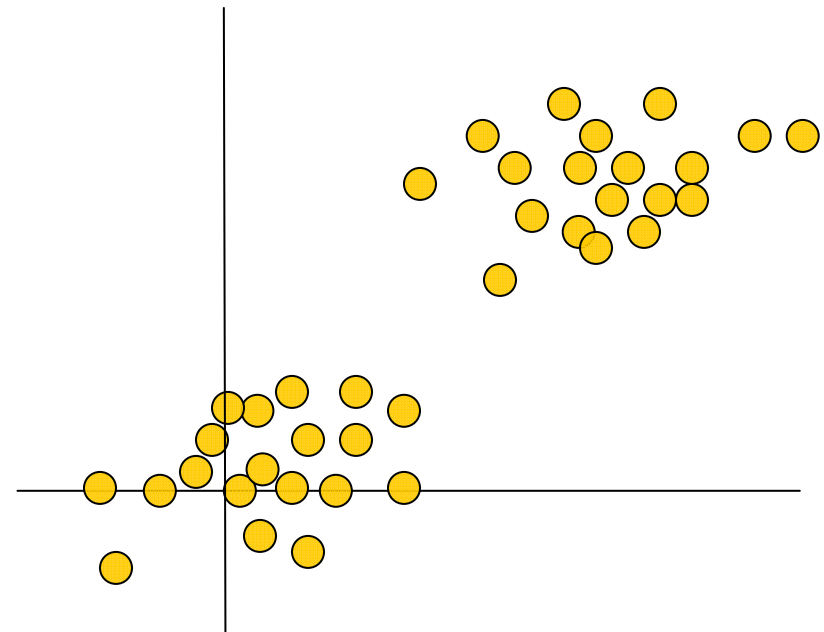
Document Clustering

- Previously we classified Documents into Two Classes
 - Hockey (Class1) and Baseball (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

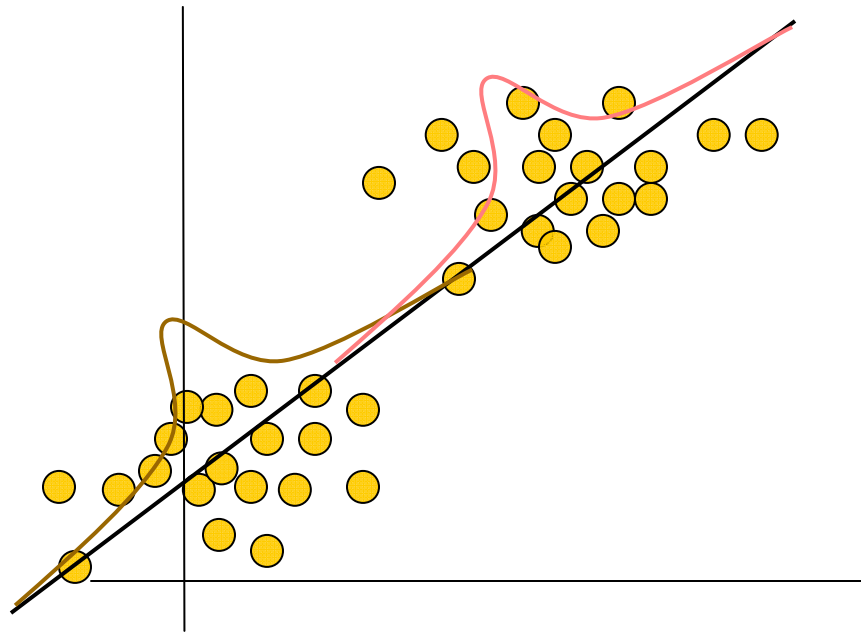


Supervised Training
of Classification Algorithm



Unsupervised Training
of Clustering Algorithm

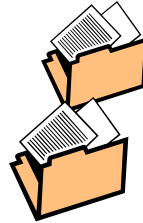
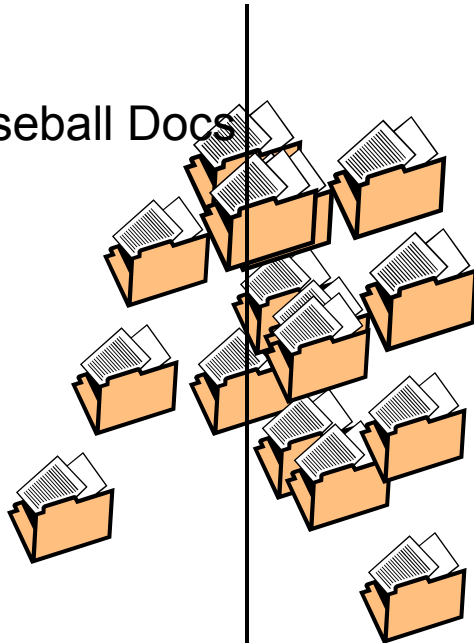
Clusters for Classification



Automatically Found Clusters
can be used for Classification

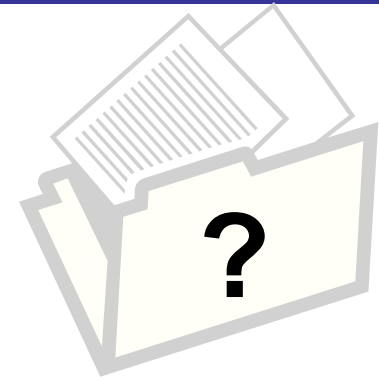
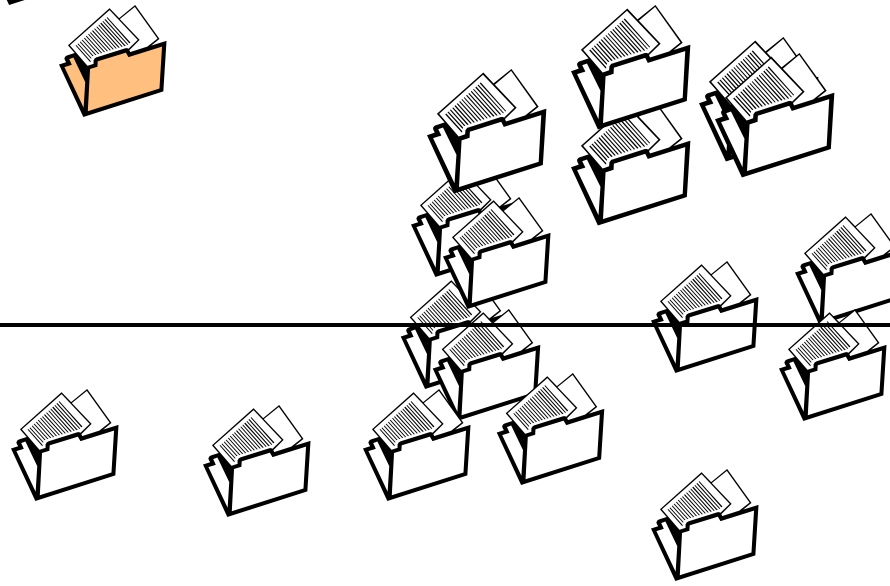
Document Clustering

Baseball Docs



Which cluster does the new document belong to?

Hockey Docs



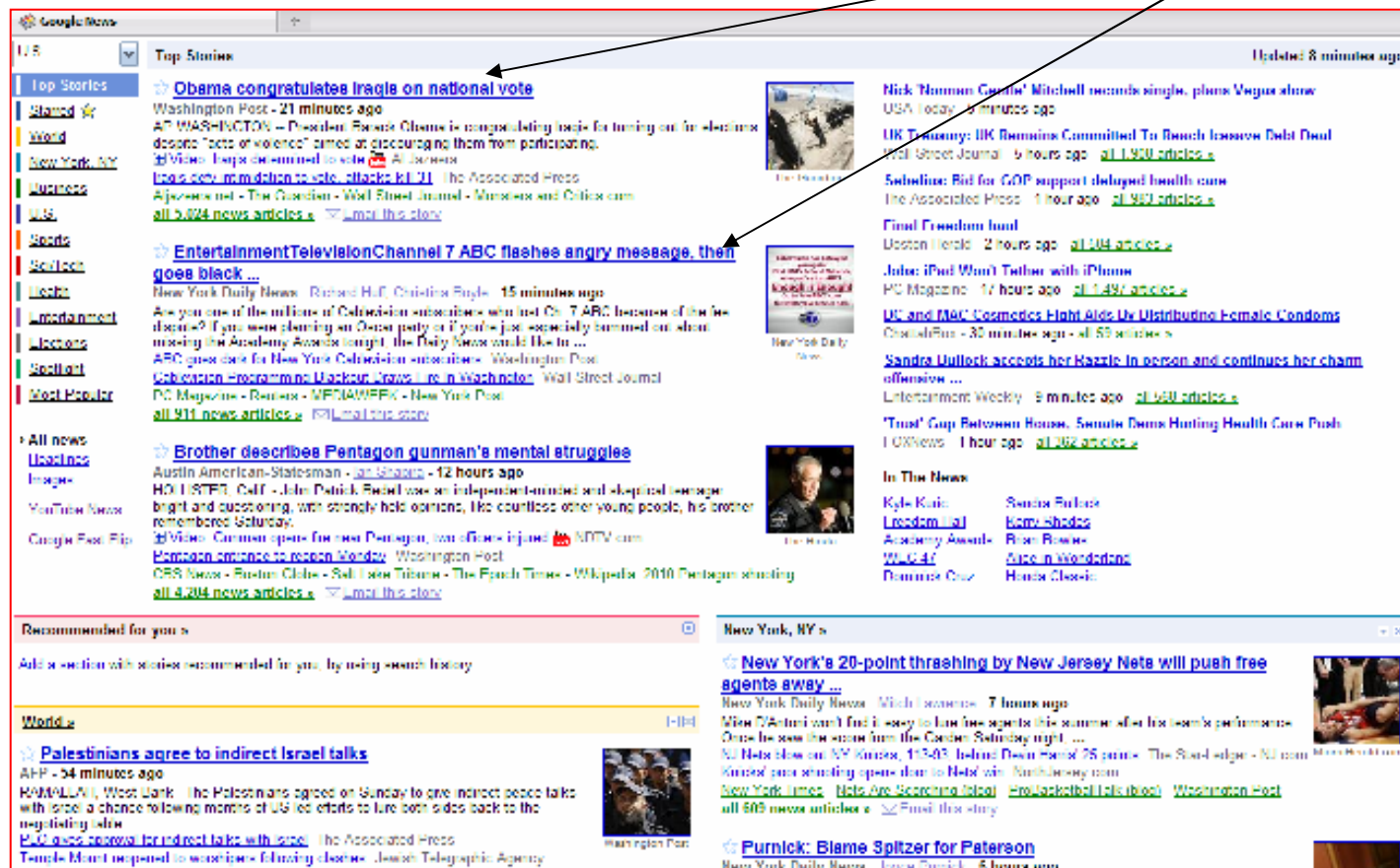
Document Clustering

- 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

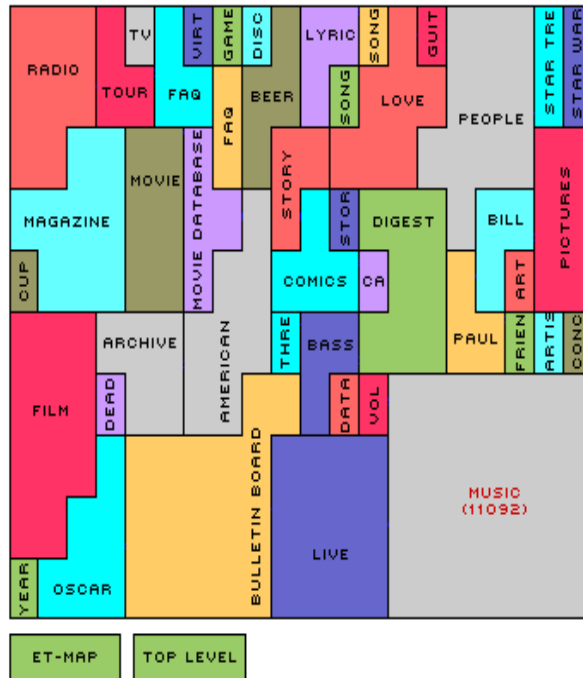
Document Clustering Application

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

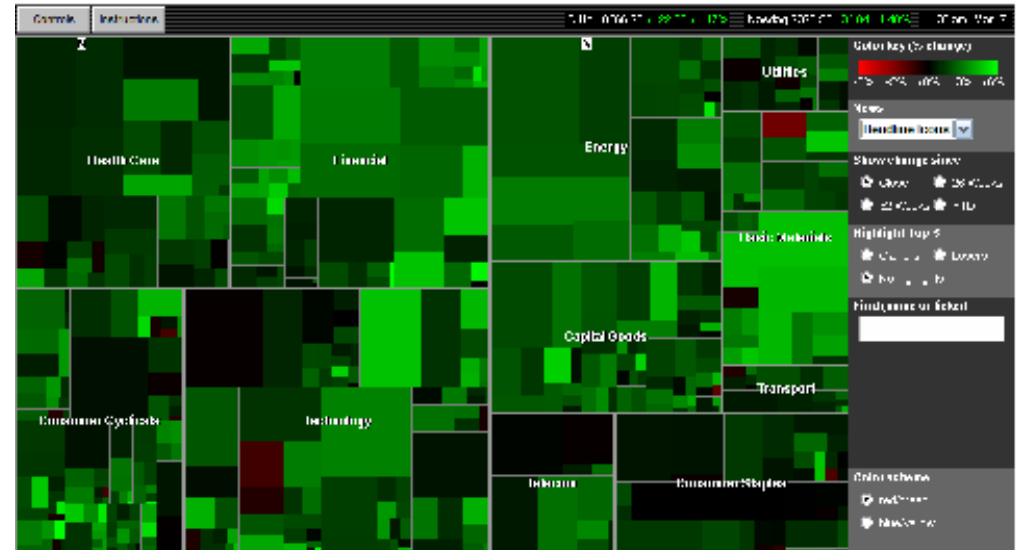
News Clusters



Document Clustering Application



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



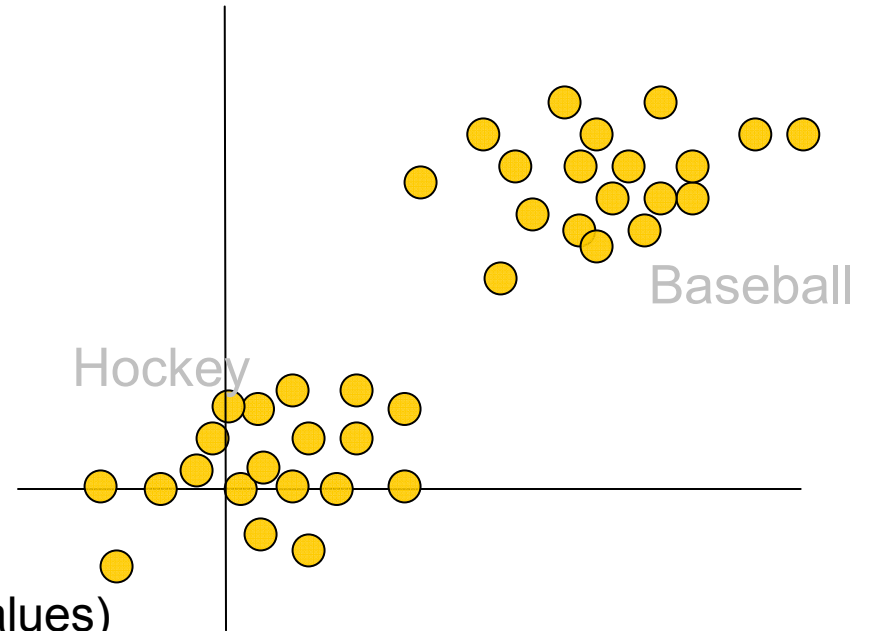
Map of the Market with Headlines Smartmoney [2]

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

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 (C_k))
- Use the assigned points to recompute mu for 2 Gaussians



K-Means Clustering

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

We want to cluster the data in K clusters

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 $k = 1, \dots, K$ 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 μ_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 $||x_n - r_{nk}||^2$

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

- 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.r.t. μ_k to zero

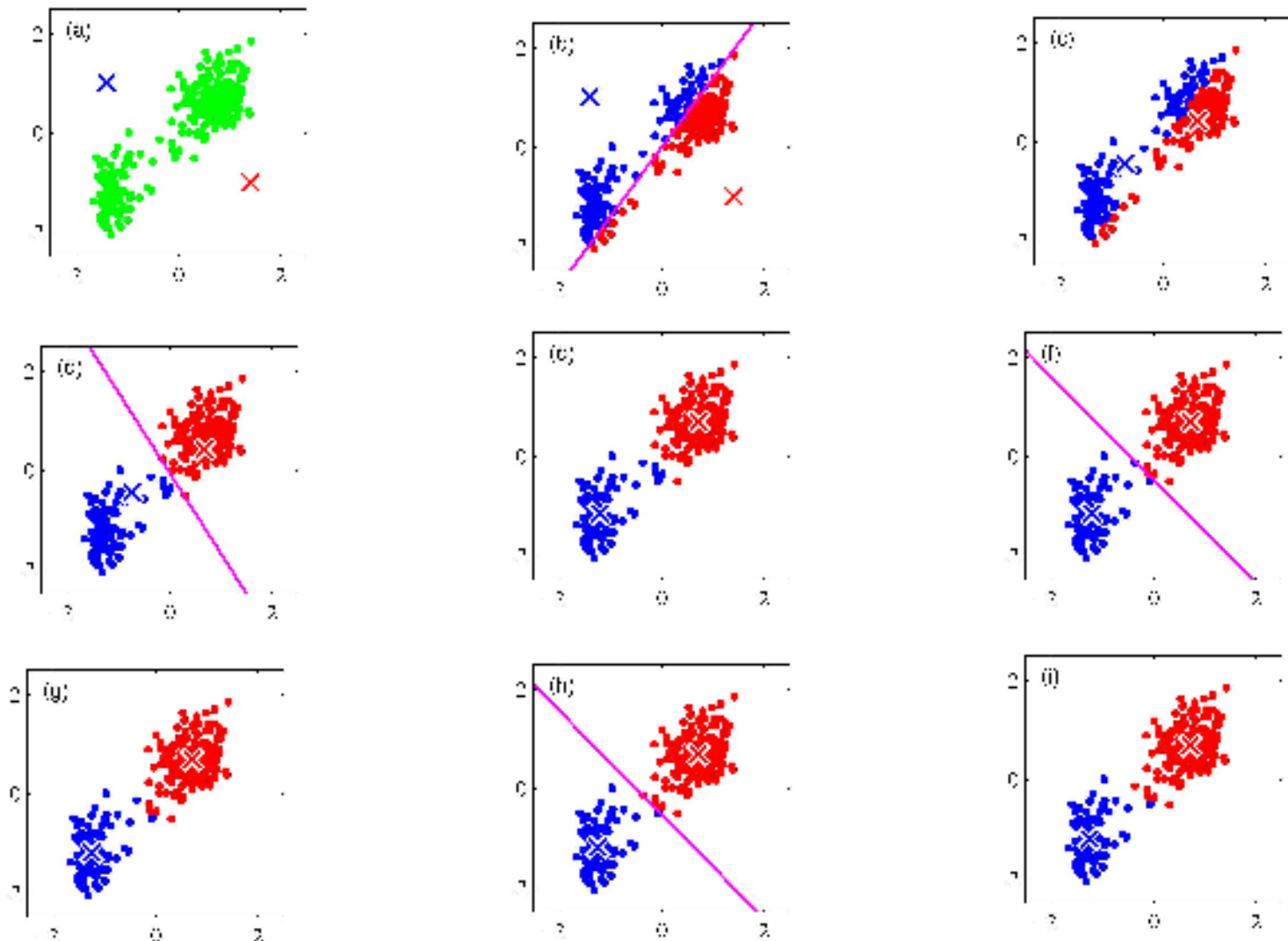
$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$

- 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 from Homework 1 but 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. ▾ Sci/Tech

Top Stories
Starred ☆
World
U.S.
Business
Sci/Tech
Entertainment
Sports
Health
Spotlight
Most Popular


› All news
[Headlines](#)
[Images](#)

 BigPond News


☆ [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](#)

 New York Times (blog)

☆ [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](#)

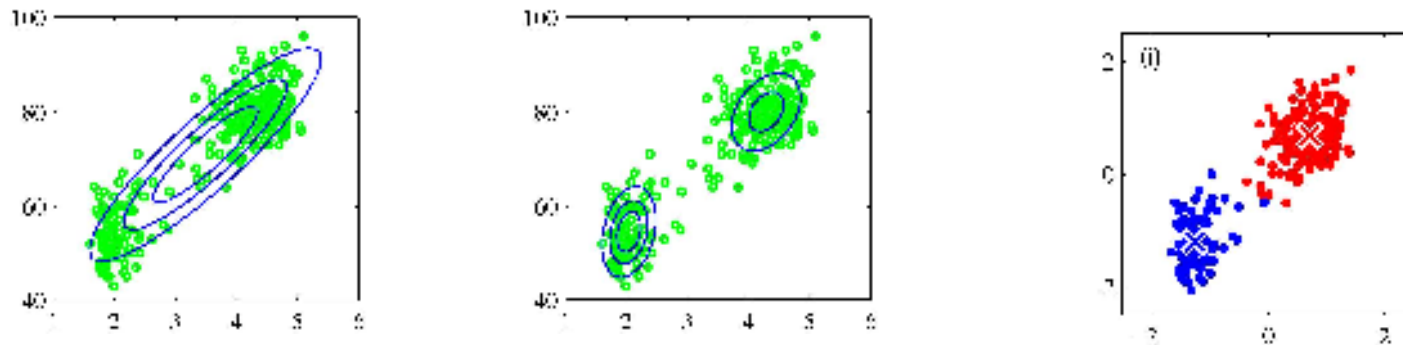
 TrustedReviews

☆ [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](#)

 SlashGear (blog)

☆ [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.
[Microsoft Showcases Cross-Platform Gaming for Windows Pho](#) PC Magazine

Mixture Models

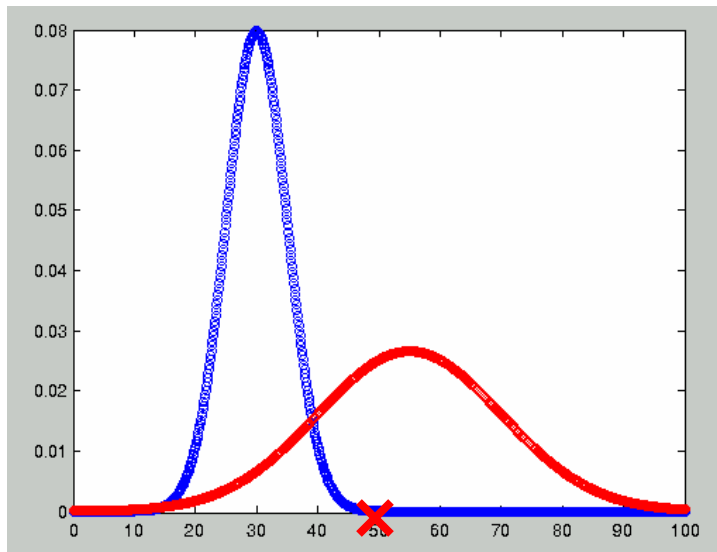


Mixture of Gaussians [1]

- 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



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

$$p(y = 1|x) \propto \mathcal{N}(x|\mu_1, \Sigma_1)p(y = 1)$$

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

$$z = [00100\dots]$$

$$\sum_{i=1}^K z^i = 1$$

- 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

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

Mixture of Gaussians with Hidden Variables

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

The diagram shows the equation $p(x) = \sum_{k=1}^K \pi_k \mathcal{N}(x | \mu_k, \Sigma_k)$ with three orange boxes and arrows pointing to its parts:

- Component of Mixture**: Points to the summation index k .
- Mixing Component**: Points to the weight π_k .
- Mean**: Points to the mean parameter μ_k .
- Covariance**: Points to the covariance parameter Σ_k .

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

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

- 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 μ , σ 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

$$\begin{aligned} p(z_k = 1|x) &= \frac{p(z_k=1)p(x|z_k=1)}{\sum_{j=1}^K p(z_j=1)p(x|z_j=1)} \\ &= \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)} \end{aligned}$$

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

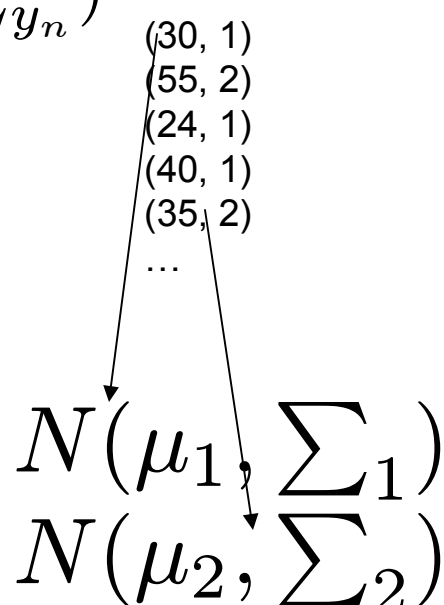
$$l = \sum_{n=1}^N \log p(x_n, y_n | \pi, \mu, \Sigma)$$

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

- 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

Find out μ_i and Σ_i from data for both classes

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



The diagram illustrates the process of mapping labeled data points to Gaussian distributions for two classes. On the right, a vertical list of data points is shown: (30, 1), (55, 2), (24, 1), (40, 1), (35, 2), and an ellipsis. Two arrows originate from this list. One arrow points from the first three points (30, 1), (55, 2), and (24, 1) to the Gaussian distribution $N(\mu_1, \Sigma_1)$. The other arrow points from the remaining points (40, 1), (35, 2), and the ellipsis to the Gaussian distribution $N(\mu_2, \Sigma_2)$. This represents the grouping of data points by class label (y) to estimate the parameters of the Gaussian distribution for each class.

$$N(\mu_1, \Sigma_1)$$
$$N(\mu_2, \Sigma_2)$$

Maximizing Likelihood for Mixture Model with Hidden Variables

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

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

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

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

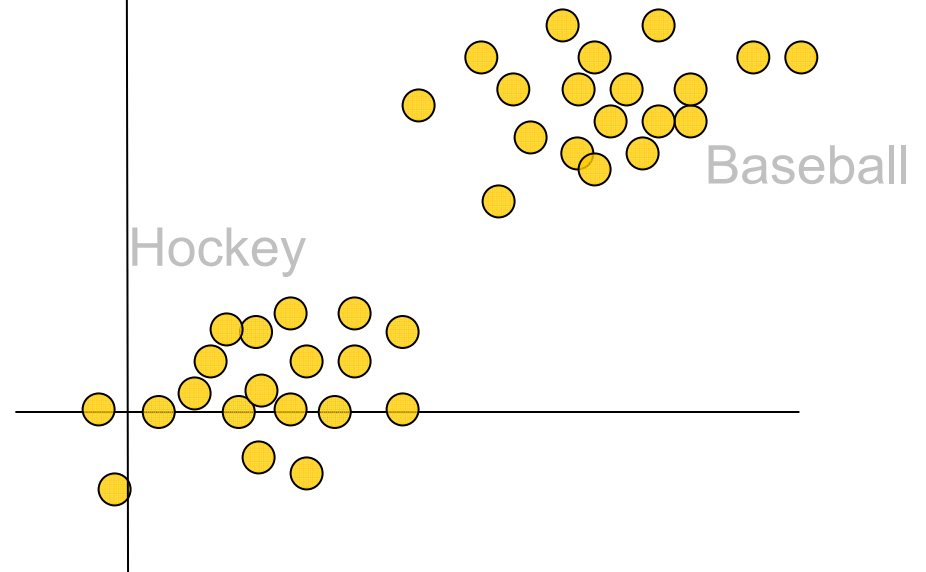
Log-likelihood for Mixture of Gaussians

$$\log p(X|\pi, \mu, \Sigma) = \sum_{n=1}^N \log \left(\sum_{k=1}^K \mathcal{N}(x|\mu_k, \Sigma_k) \right)$$

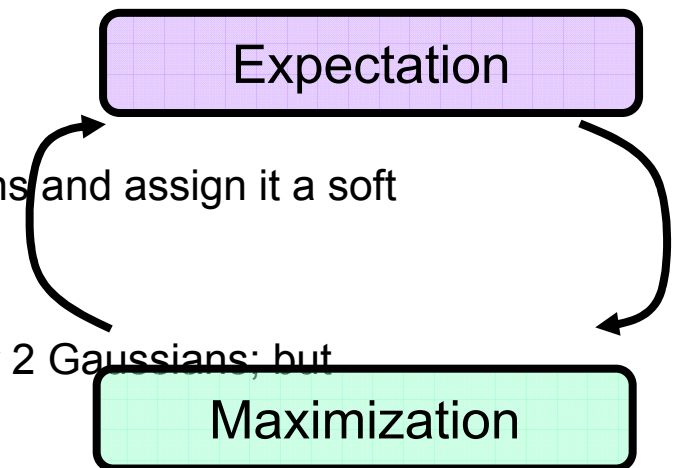
- We want to find maximum likelihood of the above log-likelihood function to find the best parameters that maximize the data given the model
- We can again do iterative process for estimating the log-likelihood of the above function
 - This 2-step iterative process is called Expectation-Maximization

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 μ and σ values)
- Compute distance of each point to the μ of 2 Gaussians and assign it a soft class label (C_k)
- Use the assigned points to recompute μ and σ 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

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

■ E-Step

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

Estimating Parameters

- M-step

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

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

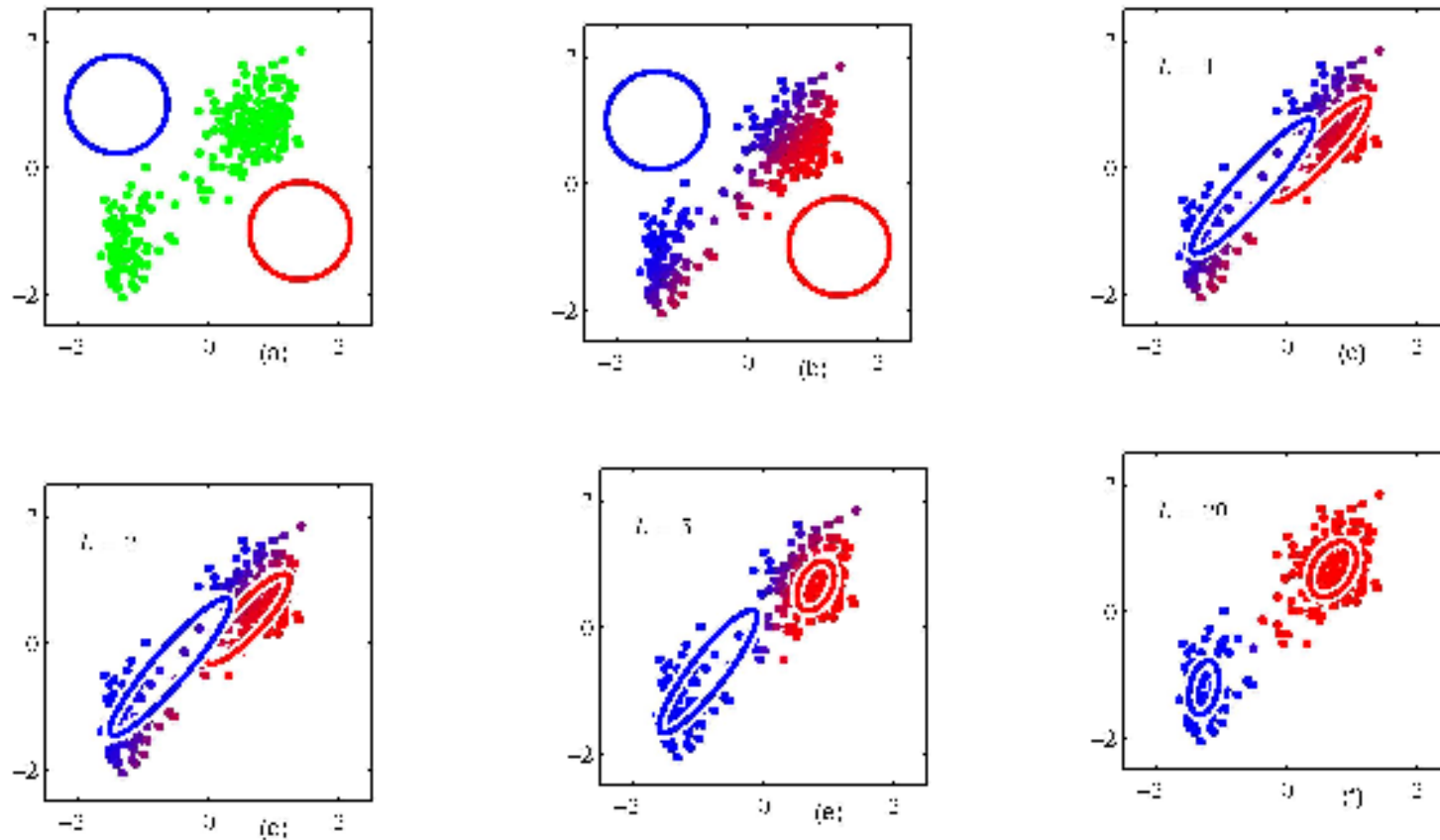
$$\pi'_k = \frac{N_k}{N}$$

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

- Iterate until convergence of log likelihood

$$\log p(X|\pi, \mu, \Sigma) = \sum_{n=1}^N \log \left(\sum_{k=1}^K \mathcal{N}(x|\mu_k, \Sigma_k) \right)$$

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?

EM Algorithm in General

- We want to find maximum likelihood solution for the model with latent variables
 - For document clustering latent variable may represent class tags
- The method of expectation-maximization can be used for maximizing many flavors of functions with latent variables
- Let us look at general representation of EM algorithm

General EM Algorithm

- ~ We want to maximize likelihood function $p(X|\theta)$
- ~ Let the latent variables be Z
- ~ Joint distribution over observed and hidden/latent variables is $p(X, Z|\theta)$

$$p(X|\theta) = \sum_z p(X, Z|\theta)$$

$$\log p(X|\theta) = \log(\sum_z p(X, Z|\theta))$$

↑
We want to maximize the log likelihood
but
Log likelihood not concave so cannot just derivative to zero

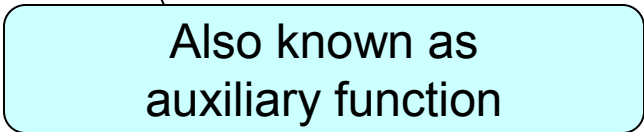
General EM Algorithm

- If we were given the class labels (values for hidden variables Z) we will have $\{X, Z\}$ which is complete data set
 - maximization of complete-data log-likelihood would be simpler
- Even though we may not have real class labels we can get expected Z using posterior distribution

$$p(Z|X, \theta)$$

- We can then use this posterior distribution and find expectation of the complete-data log likelihood evaluated for some parameter θ denoted by

$$Q(\theta, \theta^{old}) = \sum_z p(Z|X, \theta^{old}) \log p(X, Z|\theta)$$



Also known as
auxiliary function

General EM Algorithm

- E-Step:

$$p(Z|X, \theta^{old})$$

We are using expected values of hidden parameters to maximize the log likelihood in M step, thus finding better parameters in each iteration

- M-Step:

$$\theta^{new} = \operatorname{argmax}_{\theta} Q(\theta, \theta^{old})$$

where

$$Q(\theta, \theta^{old}) = \sum_z p(Z|X, \theta^{old}) \log p(X, Z|\theta)$$

EM as Bound Maximization

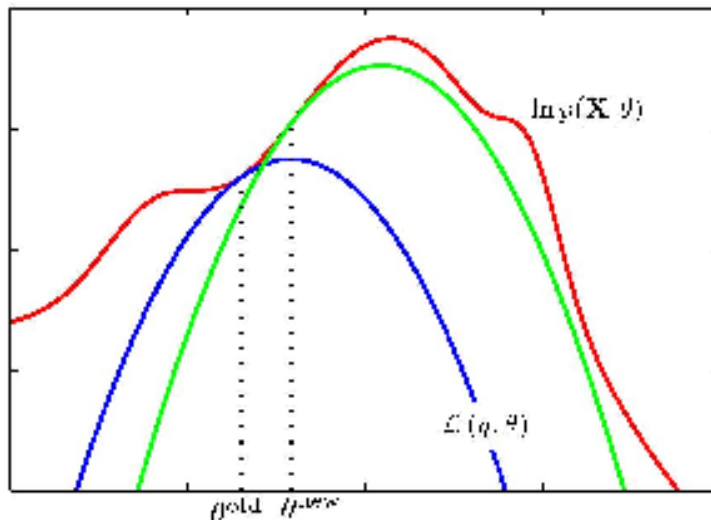
- If we cannot maximize a log-likelihood function directly maximize it's lower bound
- Lower bound takes the form

Entropy of q distribution, independent of θ

$$\mathcal{L}(q, \theta) = Q(\theta, \theta^{old}) - \text{const}$$

- Maximizing auxiliary function we showed before

$$Q(\theta, \theta^{old}) = \sum_z p(Z|X, \theta^{old}) \log p(X, Z|\theta)$$



Maximizing auxiliary function [1]

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

Similarity

- While clustering documents we are essentially finding 'similar' documents
- How we compute similarity makes a difference in the performance of clustering algorithm
- Some similarity metrics
 - Euclidean distance
 - Cross Entropy
 - Cosine Similarity
- Which similarity metric to use?

Similarity for Words

- Edit distance
 - Insertion, deletion, substitution
 - Dynamic programming algorithm
- Longest Common Subsequence
- Bigram overlap of characters
- Similarity based on meaning
 - WordNet synonyms
- Similarity based on collocation

Similarity of Text : Surface, Syntax and Semantics

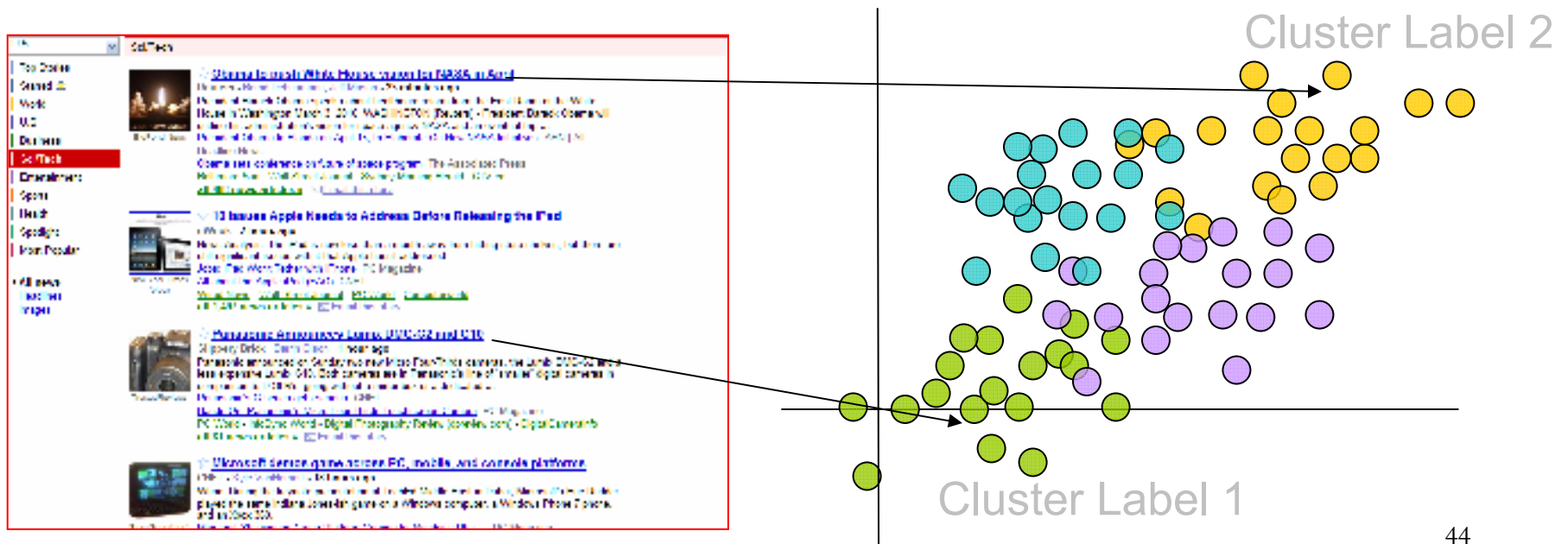
- Cosine Similarity
 - Binary Vectors
 - Multinomial Vectors
- Edit distance
 - Insertion, deletion, substitution
- Semantic similarity
 - Look beyond surface forms
 - WordNet, semantic classes
- Syntactic similarity
 - Syntactic structure
 - Tree Kernels
- Many ways to look at similarity and choice of the metric is important for the type of clustering algorithm we are using

Clustering Documents

- Represent documents as feature vectors
- Decide on Similarity Metric for computing similarity across feature vectors
- Use Iterative algorithm that maximize the log-likelihood of the function with hidden variables that represent the cluster IDs

Automatic Labeling of Clusters

- How do you automatically label the clusters
- For example, how do you find the headline that represent the news pieces in given topic
 - ❑ One possible way is to find the most similar sentence to the centroid of the cluster



Clustering Sentences by Topic

- We can cluster documents, sentences or any segment of text
- Similarity across text segments can take account of topic similarity
- We can still use our unsupervised clustering algorithm based on K-means or EM
 - Similarity needs to be computed at the sentence level
- Useful for summarization, question answering, text categorization

Summary

- Unsupervised clustering algorithms
 - K-means
 - Expectation Maximization
- 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/>