Clustering / Unsupervised Methods
Lecture 9

Jason Corso, Albert Chen

SUNY at Buffalo

April 2010

Introduction

- Until now, we've assumed our training samples are "labeled" by their category membership.
- Methods that use labeled samples are said to be *supervised*; otherwise, they're said to be *unsupervised*.
- However:
  - Why would one even be interested in learning with unlabeled samples?
  - Is it even possible in principle to learn anything of value from unlabeled samples?

Why Unsupervised Learning?

1. Collecting and labeling a large set of sample patterns can be surprisingly costly.
    - E.g., videos are virtually free, but accurately *labeling* the video pixels is expensive and time consuming.

Why Unsupervised Learning?

1. Collecting and labeling a large set of sample patterns can be surprisingly costly.
   - E.g., videos are virtually free, but accurately *labeling* the video pixels is expensive and time consuming.
2. Extend to a larger training set by using *semi-supervised learning*.
   - Train a classifier on a small set of samples, then tune it up to make it run without supervision on a large, unlabeled set.
   - Or, in the reverse direction, let a large set of unlabeled data group automatically, then label the groupings found.

Why Unsupervised Learning?

1. Collecting and labeling a large set of sample patterns can be surprisingly costly.
    - E.g., videos are virtually free, but accurately *labeling* the video pixels is expensive and time consuming.
2. Extend to a larger training set by using *semi-supervised learning*.
    - Train a classifier on a small set of samples, then tune it up to make it run without supervision on a large, unlabeled set.
    - Or, in the reverse direction, let a large set of unlabeled data group automatically, then label the groupings found.
3. To detect the gradual change of pattern over time.

Why Unsupervised Learning?

1. Collecting and labeling a large set of sample patterns can be surprisingly costly.
   - E.g., videos are virtually free, but accurately *labeling* the video pixels is expensive and time consuming.
2. Extend to a larger training set by using *semi-supervised learning*.
   - Train a classifier on a small set of samples, then tune it up to make it run without supervision on a large, unlabeled set.
   - Or, in the reverse direction, let a large set of unlabeled data group automatically, then label the groupings found.
3. To detect the gradual change of pattern over time.
4. To find features that will then be useful for categorization.

Why Unsupervised Learning?

1. Collecting and labeling a large set of sample patterns can be surprisingly costly.
   - E.g., videos are virtually free, but accurately *labeling* the video pixels is expensive and time consuming.
2. Extend to a larger training set by using *semi-supervised learning*.
   - Train a classifier on a small set of samples, then tune it up to make it run without supervision on a large, unlabeled set.
   - Or, in the reverse direction, let a large set of unlabeled data group automatically, then label the groupings found.
3. To detect the gradual change of pattern over time.
4. To find features that will then be useful for categorization.
5. To gain insight into the nature or structure of the data during the early stages of an investigation.

Data Clustering

Source: A. K. Jain and R. C. Dubes. Alg. for Clustering Data, Prentiice Hall, 1988.

- What is data clustering?
  - Grouping of objects into meaningful categories
  - Given a **representation** of N objects, find k clusters based on a measure of **similarity**.

Data Clustering

Source: A. K. Jain and R. C. Dubes. Alg. for Clustering Data, Prentiice Hall, 1988.

- What is data clustering?
  - Grouping of objects into meaningful categories
  - Given a **representation** of N objects, find k clusters based on a measure of **similarity**.
- Why data clustering?
  - Natural Classification: degree of similarity among forms.
  - Data exploration: discover underlying structure, generate hypotheses, detect anomalies.
  - Compression: for organizing data.
  - Applications: can be used by any scientific field that collects data!

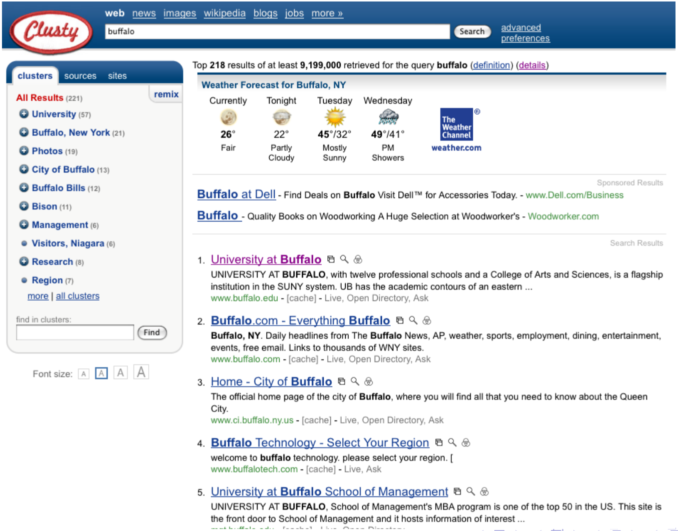
Data Clustering

Source: A. K. Jain and R. C. Dubes. Alg. for Clustering Data, Prentiice Hall, 1988.

- What is data clustering?
  - Grouping of objects into meaningful categories
  - Given a **representation** of N objects, find k clusters based on a measure of **similarity**.
- Why data clustering?
  - Natural Classification: degree of similarity among forms.
  - Data exploration: discover underlying structure, generate hypotheses, detect anomalies.
  - Compression: for organizing data.
  - Applications: can be used by any scientific field that collects data!
- Google Scholar: 1500 clustering papers in 2007 alone!

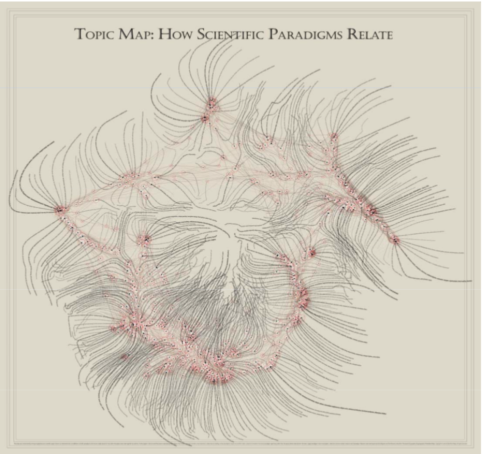
E.g.: Structure Discovering via Clustering

Source: http://clusty.com



E.g.: Topic Discovery
Source: Map of Science, Nature, 2006

- 800,000 scientific papers clustered into 776 topics based on how often the papers were cited together by authors of other papers



Data Clustering - Formal Definition

- Given a set of N unlabeled examples $D = x_1, x_2, ..., x_N$ in a d-dimensional feature space, D is partitioned into a number of disjoint subsets D_j's:

$$D = \cup_{j=1}^{k} D_j \quad \text{where } D_i \cup D_j = \emptyset, i \neq j \ , \tag{1}$$

  where the points in each subset are similar to each other according to a given criterion Φ.

Data Clustering - Formal Definition

- Given a set of N unlabeled examples $D = x_1, x_2, ..., x_N$ in a d-dimensional feature space, D is partitioned into a number of disjoint subsets D_j's:

$$D = \cup_{j=1}^{k} D_j \quad \text{where } D_i \cup D_j = \emptyset, i \neq j \ , \tag{1}$$

where the points in each subset are similar to each other according to a given criterion Φ.

- A partition is denoted by

$$\pi = (D_1, D_2, ..., D_k) \tag{2}$$

and the problem of data clustering is thus formulated as

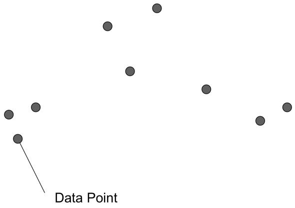
$$\pi^* = \underset{\pi}{\text{argmin}} \ f(\pi) \ , \tag{3}$$

where $f(\cdot)$ is formulated according to Φ.

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

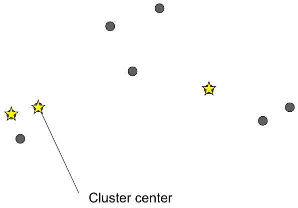
- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i



k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i

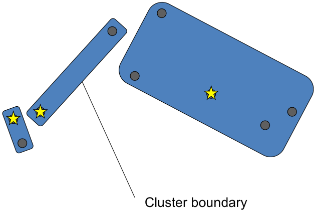


First choose *k* arbitrary centers

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i

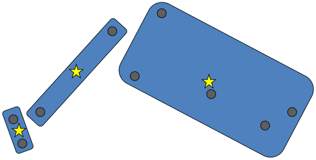


Assign points to closest centers

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i

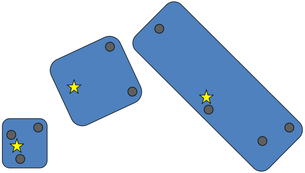


Recompute centers

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i

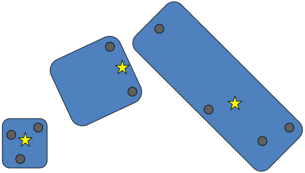


Assign points to closest centers

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i

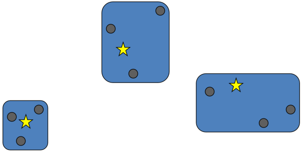


Recompute centers

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i

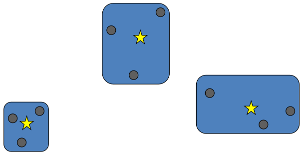


Assign points to closest centers

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i

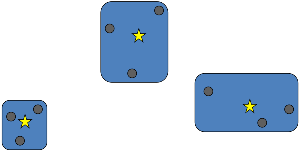


Recompute centers

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i

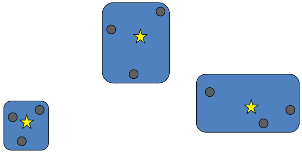


Points already assigned to nearest

k-Means Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Randomly initialize $\mu_1, \mu_2, ..., \mu_c$
- Repeat until no change in μ_i:
    - Classify N samples according to nearest μ_i
    - Recompute μ_i



Points already assigned to nearest centers; Algorithm ends

k-Means++ Clustering

Source: D. Aurthor and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding

- Choose starting centers iteratively.
- Let $D(x)$ be the distance from x to the nearest existing center, take x as new center with probability $\propto D(x)^2$.
- Repeat until no change in μ_i:
  - Classify N samples according to nearest μ_i
  - Recompute μ_i
- (refer to the slides by D. Author and S. Vassolvitskii for details)

User's Dilemma

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

1. What is a cluster?
2. How to define pair-wise similarity?

User's Dilemma

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

1. What is a cluster?
2. How to define pair-wise similarity?
3. Which features and normalization scheme?

User's Dilemma

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

1. What is a cluster?
2. How to define pair-wise similarity?
3. Which features and normalization scheme?
4. How many clusters?

User's Dilemma

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

1. What is a cluster?
2. How to define pair-wise similarity?
3. Which features and normalization scheme?
4. How many clusters?
5. Which clustering method?
6. Are the discovered clusters and partition valid?

User's Dilemma

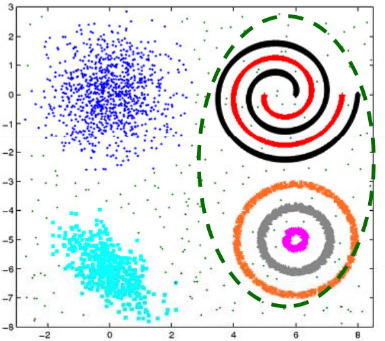
Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

1. What is a cluster?
2. How to define pair-wise similarity?
3. Which features and normalization scheme?
4. How many clusters?
5. Which clustering method?
6. Are the discovered clusters and partition valid?
7. Does the data have any clustering tendency?

Cluster Similarity?

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

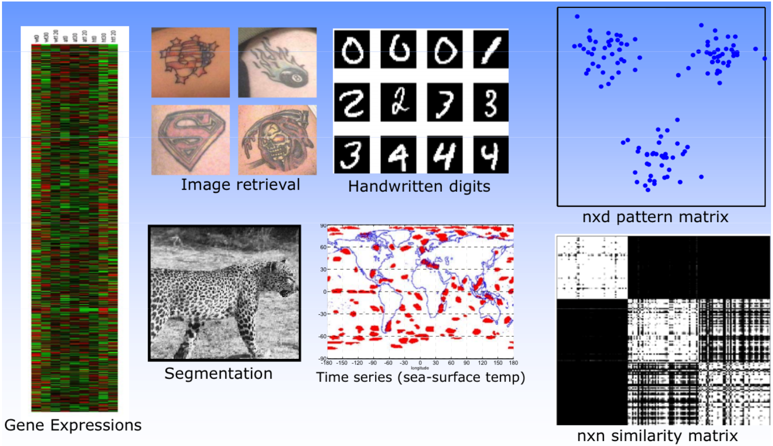
- Compact Clusters
  - Within-cluster **distance** $<$ between-cluster connectivity
- Connected Clusters
  - Within-cluster **connectivity** $>$ between-cluster connectivity
- Ideal cluster: **compact** and **isolated**.



Representation (features)?

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

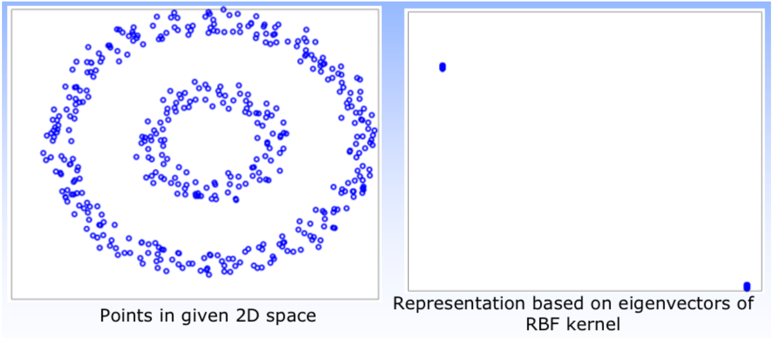
- There's no universal representation; they're domain dependent.



Good Representation

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

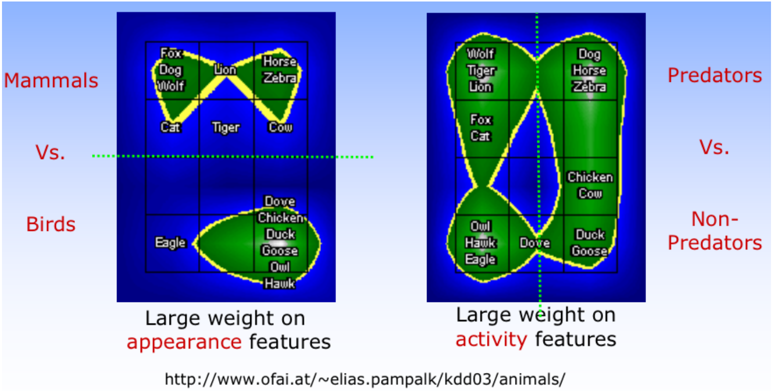
- A good representation leads to compact and isolated clusters.



How do we weigh the features?

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

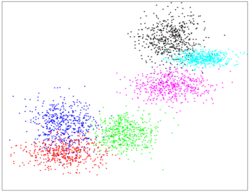
- Two different meaningful groupings produced by different weighting schemes.



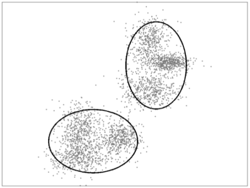
How do we decide the Number of Clusters?

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

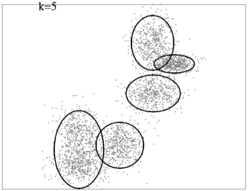
- The samples are generated by 6 independent classes, yet:



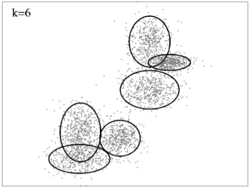
ground truth



$k = 2$



$k = 5$

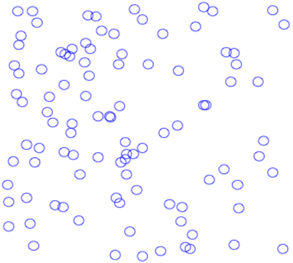


$k = 6$

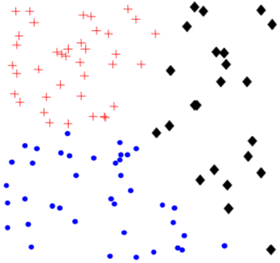
Cluster Validity

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

- Clustering algorithms find clusters, even if there are no **natural** clusters in the data.



100 2D uniform data points

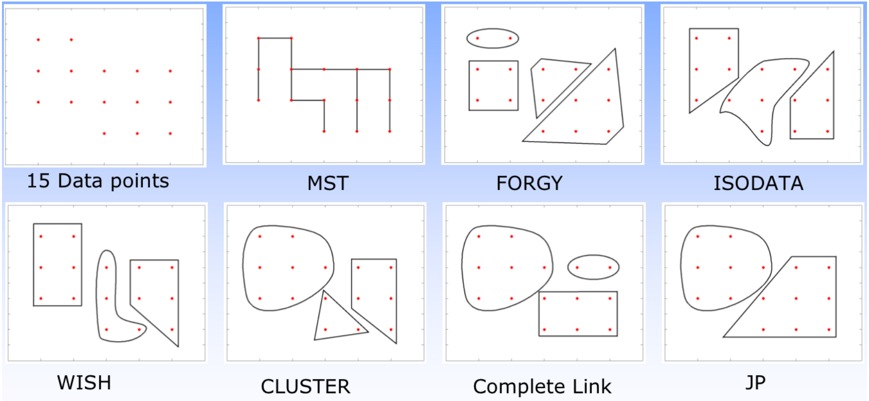


k-Means with k=3

Comparing Clustering Methods

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

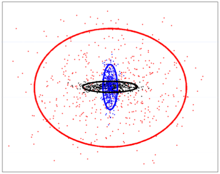
- Which clustering algorithm is the best?



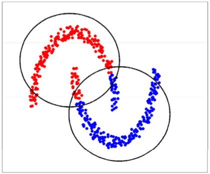
There's no best Clustering Algorithm!

Source: R. Dubes and A. K. Jain, Clustering Techniques: User's Dilemma, PR 1976

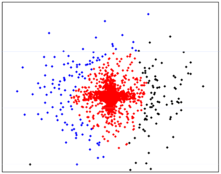
- Each algorithm imposes a structure on data.
- Good fit between model and data $\Rightarrow$ success.



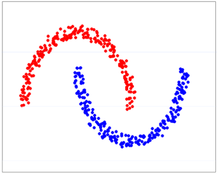
GMM; k=3



GMM; k=2



Spectral; k=3



Spectral; k=2

Gaussian Mixture Models

- Recall the Gaussian distribution:

$$\mathcal{N}(x|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \quad (4)$$

Gaussian Mixture Models

- Recall the Gaussian distribution:

$$\mathcal{N}(x|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{4}$$

- It forms the basis for the important Mixture of Gaussians density.

Gaussian Mixture Models

- Recall the Gaussian distribution:

$$\mathcal{N}(x|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{4}$$

- It forms the basis for the important Mixture of Gaussians density.
- The Gaussian mixture is a **linear superposition of Gaussians** in the form:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad . \tag{5}$$

Gaussian Mixture Models

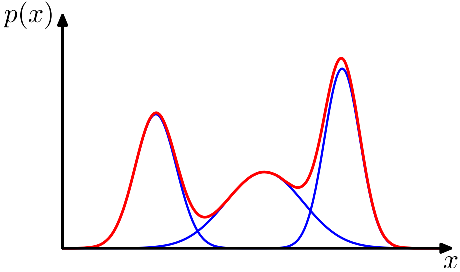
- Recall the Gaussian distribution:

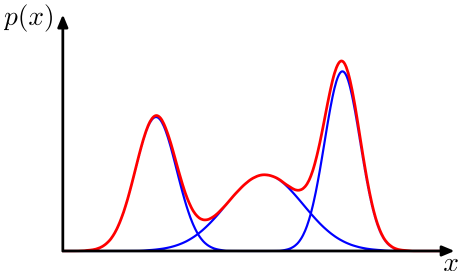
$$\mathcal{N}(x|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{4}$$

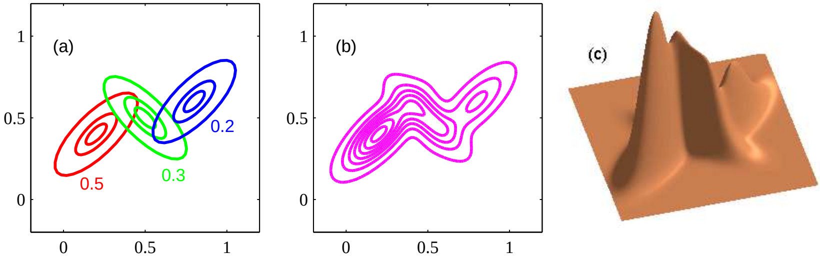
- It forms the basis for the important Mixture of Gaussians density.
- The Gaussian mixture is a **linear superposition of Gaussians** in the form:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \ . \tag{5}$$

- The π_k are non-negative scalars called **mixing coefficients** and they govern the relative importance between the various Gaussians in the mixture density. $\sum_k \pi_k = 1$.


$p(x)$
x


$p(x)$
x


(a)
1
0.5
0
0
0.5
1
0.5
0.3
0.2
(b)
1
0.5
0
0
0.5
1
(c)

Introducing Latent Variables

- Define a K-dimensional binary random variable $\mathbf{z}$.

Introducing Latent Variables

- Define a K-dimensional binary random variable $\mathbf{z}$.
- $\mathbf{z}$ has a 1-of-K representation such that a particular element z_k is 1 and all of the others are zero. Hence:

$$z_k \in \{0, 1\} \tag{6}$$

$$\sum_k z_k = 1 \tag{7}$$

Introducing Latent Variables

- Define a K-dimensional binary random variable $\mathbf{z}$.
- $\mathbf{z}$ has a 1-of-K representation such that a particular element z_k is 1 and all of the others are zero. Hence:

$$z_k \in \{0, 1\} \tag{6}$$

$$\sum_k z_k = 1 \tag{7}$$

- The marginal distribution over $\mathbf{z}$ is specified in terms of the mixing coefficients:

$$p(z_k = 1) = \pi_k \tag{8}$$

And, recall, $0 \leq \pi_k \leq 1$ and $\sum_k \pi_k = 1$.

- Since $\mathbf{z}$ has a 1-of-K representation, we can also write this distribution as

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{9}$$

- Since $\mathbf{z}$ has a 1-of-K representation, we can also write this distribution as

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{9}$$

- The conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ is a Gaussian:

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{10}$$

or

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \tag{11}$$

- We are interested in the marginal distribution of $\mathbf{x}$:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) \tag{12}$$

$$= \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) \tag{13}$$

$$= \sum_{\mathbf{z}} \prod_{k=1}^{K} \pi_k^{z_k} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \tag{14}$$

$$= \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{15}$$

- We are interested in the marginal distribution of $\mathbf{x}$:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) \tag{12}$$

$$= \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) \tag{13}$$

$$= \sum_{\mathbf{z}} \prod_{k=1}^{K} \pi_k^{z_k} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \tag{14}$$

$$= \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{15}$$

- So, given our latent variable $\mathbf{z}$, the marginal distribution of $\mathbf{x}$ is a Gaussian mixture.

- We are interested in the marginal distribution of $\mathbf{x}$:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) \tag{12}$$
$$= \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) \tag{13}$$
$$= \sum_{\mathbf{z}} \prod_{k=1}^{K} \pi_k^{z_k} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \tag{14}$$
$$= \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{15}$$

- So, given our latent variable $\mathbf{z}$, the marginal distribution of $\mathbf{x}$ is a Gaussian mixture.
- If we have N observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$, then because of our chosen representation, it follows that we have a latent variable $\mathbf{z}_n$ for each observed data point $\mathbf{x}_n$.

Component Responsibility Term

- We need to also express the conditional probability of $\mathbf{z}$ given $\mathbf{x}$.

Component Responsibility Term

- We need to also express the conditional probability of $\mathbf{z}$ given $\mathbf{x}$.
- Denote this conditional $p(z_k = 1|\mathbf{x})$ as $\gamma(z_k)$.

Component Responsibility Term

- We need to also express the conditional probability of $\mathbf{z}$ given $\mathbf{x}$.
- Denote this conditional $p(z_k = 1|\mathbf{x})$ as $\gamma(z_k)$.
- We can derive this value with Bayes' theorem:

$$\gamma(z_k) \doteq p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \tag{16}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{17}$$

Component Responsibility Term

- We need to also express the conditional probability of $\mathbf{z}$ given $\mathbf{x}$.
- Denote this conditional $p(z_k = 1|\mathbf{x})$ as $\gamma(z_k)$.
- We can derive this value with Bayes' theorem:

$$\gamma(z_k) \doteq p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \tag{16}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{17}$$

- View π_k as the prior probability of $z_k = 1$ and the quantity $\gamma(z_k)$ as the corresponding posterior probability once we have observed $\mathbf{x}$.

Component Responsibility Term

- We need to also express the conditional probability of $\mathbf{z}$ given $\mathbf{x}$.
- Denote this conditional $p(z_k = 1|\mathbf{x})$ as $\gamma(z_k)$.
- We can derive this value with Bayes' theorem:

$$\gamma(z_k) \doteq p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \tag{16}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{17}$$

- View π_k as the prior probability of $z_k = 1$ and the quantity $\gamma(z_k)$ as the corresponding posterior probability once we have observed $\mathbf{x}$.
- $\gamma(z_k)$ can also be viewed as the responsibility that component k takes for explaining the observation $\mathbf{x}$.

Sampling from the GMM

- To sample from the GMM, we can first generate a value for $\mathbf{z}$ from the marginal distribution $p(\mathbf{z})$. Denote this sample $\hat{\mathbf{z}}$.

Sampling from the GMM

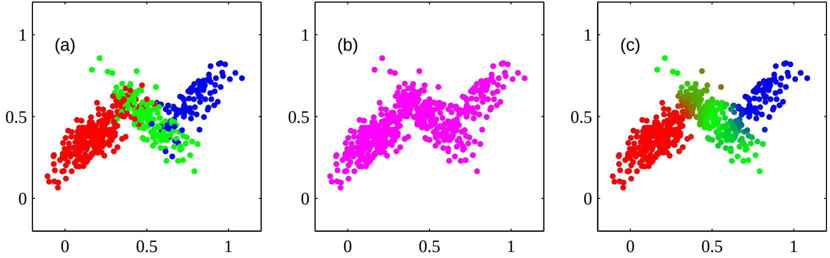
- To sample from the GMM, we can first generate a value for $\mathbf{z}$ from the marginal distribution $p(\mathbf{z})$. Denote this sample $\hat{\mathbf{z}}$.
- Then, sample from the conditional distribution $p(\mathbf{x}|\hat{\mathbf{z}})$.

Sampling from the GMM

- To sample from the GMM, we can first generate a value for $\mathbf{z}$ from the marginal distribution $p(\mathbf{z})$. Denote this sample $\hat{\mathbf{z}}$.
- Then, sample from the conditional distribution $p(\mathbf{x}|\hat{\mathbf{z}})$.
- The figure below-left shows samples from a three-mixture and colors the samples based on their $\mathbf{z}$. The figure below-middle shows samples from the marginal $p(\mathbf{x})$ and ignores $\mathbf{z}$. On the right, we show the $\gamma(z_k)$ for each sampled point, colored accordingly.

Sampling from the GMM

- To sample from the GMM, we can first generate a value for $\mathbf{z}$ from the marginal distribution $p(\mathbf{z})$. Denote this sample $\hat{\mathbf{z}}$.
- Then, sample from the conditional distribution $p(\mathbf{x}|\hat{\mathbf{z}})$.
- The figure below-left shows samples from a three-mixture and colors the samples based on their $\mathbf{z}$. The figure below-middle shows samples from the marginal $p(\mathbf{x})$ and ignores $\mathbf{z}$. On the right, we show the $\gamma(z_k)$ for each sampled point, colored accordingly.



Maximum-Likelihood

- Suppose we have a set of N observations $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ that we wish to model with a GMM.

Maximum-Likelihood

- Suppose we have a set of N observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ that we wish to model with a GMM.
- Consider this data set as an $N \times d$ matrix $\mathbf{X}$ in which the n^{th} row is given by $\mathbf{x}_n^{\mathsf{T}}$.

Maximum-Likelihood

- Suppose we have a set of N observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ that we wish to model with a GMM.
- Consider this data set as an $N \times d$ matrix $\mathbf{X}$ in which the n^{th} row is given by $\mathbf{x}_n^{\mathsf{T}}$.
- Similarly, the corresponding latent variables define an $N \times K$ matrix $\mathbf{Z}$ with rows $\mathbf{z}_n^{\mathsf{T}}$.

Maximum-Likelihood

- Suppose we have a set of N observations $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ that we wish to model with a GMM.
- Consider this data set as an $N \times d$ matrix $\mathbf{X}$ in which the n^{th} row is given by $\mathbf{x}_n^\mathsf{T}$.
- Similarly, the corresponding latent variables define an $N \times K$ matrix $\mathbf{Z}$ with rows $\mathbf{z}_n^\mathsf{T}$.
- The log-likelihood of the corresponding GMM is given by

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right] \quad . \tag{18}$$

Maximum-Likelihood

- Suppose we have a set of N observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ that we wish to model with a GMM.
- Consider this data set as an $N \times d$ matrix $\mathbf{X}$ in which the n^{th} row is given by $\mathbf{x}_n^\mathsf{T}$.
- Similarly, the corresponding latent variables define an $N \times K$ matrix $\mathbf{Z}$ with rows $\mathbf{z}_n^\mathsf{T}$.
- The log-likelihood of the corresponding GMM is given by

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right] \quad . \tag{18}$$

- Ultimately, we want to find the values of the parameters $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ that maximize this function.

- However, maximizing the log-likelihood terms for GMMs is much more complicated than for the case of a single Gaussian. Why?

- However, maximizing the log-likelihood terms for GMMs is much more complicated than for the case of a single Gaussian. Why?
- The difficulty arises from the sum over k inside of the log-term. The log function no longer acts directly on the Gaussian, and no closed-form solution is available.

Singularities

- There is a significant problem when we apply MLE to estimate GMM parameters.

Singularities

- There is a significant problem when we apply MLE to estimate GMM parameters.
- Consider simply covariances defined by $\mathbf{\Sigma}_k = \sigma_k^2 \mathbf{I}$.

Singularities

- There is a significant problem when we apply MLE to estimate GMM parameters.
- Consider simply covariances defined by $\mathbf{\Sigma}_k = \sigma_k^2 \mathbf{I}$.
- Suppose that one of the components of the mixture model, j, has its mean $\boldsymbol{\mu}_j$ exactly equal to one of the data points so that $\boldsymbol{\mu}_j = \mathbf{x}_n$ for some n.

Singularities

- There is a significant problem when we apply MLE to estimate GMM parameters.
- Consider simply covariances defined by $\mathbf{\Sigma}_k = \sigma_k^2 \mathbf{I}$.
- Suppose that one of the components of the mixture model, j, has its mean $\boldsymbol{\mu}_j$ exactly equal to one of the data points so that $\boldsymbol{\mu}_j = \mathbf{x}_n$ for some n.
- This term contributes

$$\mathcal{N}(\mathbf{x}_n | \mathbf{x}_n, \sigma_j^2 \mathbf{I}) = \frac{1}{(2\pi)^{(1/2)} \sigma_j} \tag{19}$$

Singularities

- There is a significant problem when we apply MLE to estimate GMM parameters.
- Consider simply covariances defined by $\mathbf{\Sigma}_k = \sigma_k^2 \mathbf{I}$.
- Suppose that one of the components of the mixture model, j, has its mean $\boldsymbol{\mu}_j$ exactly equal to one of the data points so that $\boldsymbol{\mu}_j = \mathbf{x}_n$ for some n.
- This term contributes

$$\mathcal{N}(\mathbf{x}_n | \mathbf{x}_n, \sigma_j^2 \mathbf{I}) = \frac{1}{(2\pi)^{(1/2)} \sigma_j} \tag{19}$$

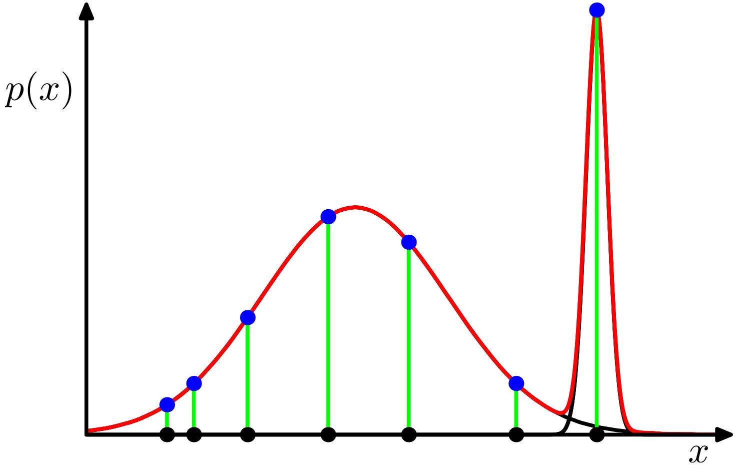
- Consider the limit $\sigma_j \to 0$ to see that this term goes to infinity and hence the log-likelihood will also go to infinity.

Singularities

- There is a significant problem when we apply MLE to estimate GMM parameters.
- Consider simply covariances defined by $\boldsymbol{\Sigma}_k = \sigma_k^2 \mathbf{I}$.
- Suppose that one of the components of the mixture model, j, has its mean $\boldsymbol{\mu}_j$ exactly equal to one of the data points so that $\boldsymbol{\mu}_j = \mathbf{x}_n$ for some n.
- This term contributes

$$\mathcal{N}(\mathbf{x}_n | \mathbf{x}_n, \sigma_j^2 \mathbf{I}) = \frac{1}{(2\pi)^{(1/2)} \sigma_j} \tag{19}$$

- Consider the limit $\sigma_j \rightarrow 0$ to see that this term goes to infinity and hence the log-likelihood will also go to infinity.
- **Thus, the maximization of the log-likelihood function is not a well posed problem because such a singularity will occur whenever one of the components collapses to a single, specific data point.**


$p(x)$
x

Expectation-Maximization for GMMs

- **Expectation-Maximization** or EM is an elegant and powerful method for finding MLE solutions in the case of missing data such as the latent variables $\mathbf{z}$ indicating the mixture component.

Expectation-Maximization for GMMs

- **Expectation-Maximization** or EM is an elegant and powerful method for finding MLE solutions in the case of missing data such as the latent variables $\mathbf{z}$ indicating the mixture component.
- Recall the conditions that must be satisfied at a maximum of the likelihood function.

Expectation-Maximization for GMMs

- **Expectation-Maximization** or EM is an elegant and powerful method for finding MLE solutions in the case of missing data such as the latent variables $\mathbf{z}$ indicating the mixture component.
- Recall the conditions that must be satisfied at a maximum of the likelihood function.
- For the mean $\boldsymbol{\mu}_k$, setting the derivatives of $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ w.r.t. $\boldsymbol{\mu}_k$ to zero yields

$$0 = -\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \boldsymbol{\Sigma}_k(\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{20}$$

$$= -\sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{\Sigma}_k(\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{21}$$

Expectation-Maximization for GMMs

- **Expectation-Maximization** or EM is an elegant and powerful method for finding MLE solutions in the case of missing data such as the latent variables $\mathbf{z}$ indicating the mixture component.
- Recall the conditions that must be satisfied at a maximum of the likelihood function.
- For the mean $\boldsymbol{\mu}_k$, setting the derivatives of $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ w.r.t. $\boldsymbol{\mu}_k$ to zero yields

$$0 = -\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \boldsymbol{\Sigma}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{20}$$

$$= -\sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{\Sigma}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{21}$$

- Note the natural appearance of the responsibility terms on the RHS.

- Multiplying by $\mathbf{\Sigma}_k^{-1}$, which we assume is non-singular, gives

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{22}$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{23}$$

- Multiplying by $\mathbf{\Sigma}_k^{-1}$, which we assume is non-singular, gives

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{22}$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{23}$$

- We see the k^{th} mean is the weighted mean over all of the points in the dataset.

- Multiplying by $\boldsymbol{\Sigma}_k^{-1}$, which we assume is non-singular, gives

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{22}$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{23}$$

- We see the k^{th} mean is the weighted mean over all of the points in the dataset.
- Interpret N_k as the number of points assigned to component k.

- Multiplying by $\boldsymbol{\Sigma}_k^{-1}$, which we assume is non-singular, gives

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{22}$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{23}$$

- We see the k^{th} mean is the weighted mean over all of the points in the dataset.
- Interpret N_k as the number of points assigned to component k.
- We find a similar result for the covariance matrix:

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \boldsymbol{\mu}_k)(x_n - \boldsymbol{\mu}_k)^{\mathsf{T}} \ . \tag{24}$$

- We also need to maximize $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to the mixing coefficients π_k.

- We also need to maximize $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to the mixing coefficients π_k.
- Introduce a Lagrange multiplier to enforce the constraint $\sum_k \pi_k = 1$.

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{25}$$

- We also need to maximize $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to the mixing coefficients π_k.
- Introduce a Lagrange multiplier to enforce the constraint $\sum_k \pi_k = 1$.

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{25}$$

- Maximizing it yields:

$$0 = \frac{1}{N_k} \sum_{n=1} \gamma(z_{nk}) + \lambda \tag{26}$$

- We also need to maximize $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to the mixing coefficients π_k.
- Introduce a Lagrange multiplier to enforce the constraint $\sum_k \pi_k = 1$.

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{25}$$

- Maximizing it yields:

$$0 = \frac{1}{N_k} \sum_{n=1} \gamma(z_{nk}) + \lambda \tag{26}$$

- After multiplying both sides by π and summing over k, we get

$$\lambda = -N \tag{27}$$

- We also need to maximize $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to the mixing coefficients π_k.
- Introduce a Lagrange multiplier to enforce the constraint $\sum_k \pi_k = 1$.

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{25}$$

- Maximizing it yields:

$$0 = \frac{1}{N_k} \sum_{n=1} \gamma(z_{nk}) + \lambda \tag{26}$$

- After multiplying both sides by π and summing over k, we get

$$\lambda = -N \tag{27}$$

- Eliminate λ and rearrange to obtain:

$$\pi_k = \frac{N_k}{N} \tag{28}$$

Solved...right?

- So, we're done, right? We've computed the maximum likelihood solutions for each of the unknown parameters.

Solved...right?

- So, we're done, right? We've computed the maximum likelihood solutions for each of the unknown parameters.
- Wrong!

Solved...right?

- So, we're done, right? We've computed the maximum likelihood solutions for each of the unknown parameters.
- Wrong!
- The responsibility terms depend on these parameters in an intricate way:

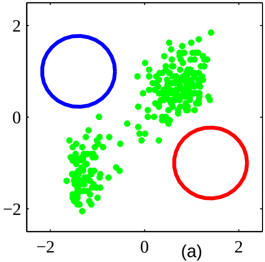
$$\gamma(z_k) \doteq p(z_k = 1|\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Solved...right?

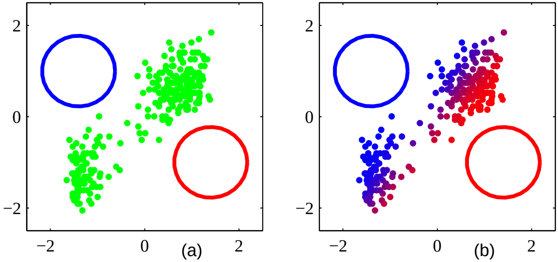
- So, we're done, right? We've computed the maximum likelihood solutions for each of the unknown parameters.
- Wrong!
- The responsibility terms depend on these parameters in an intricate way:

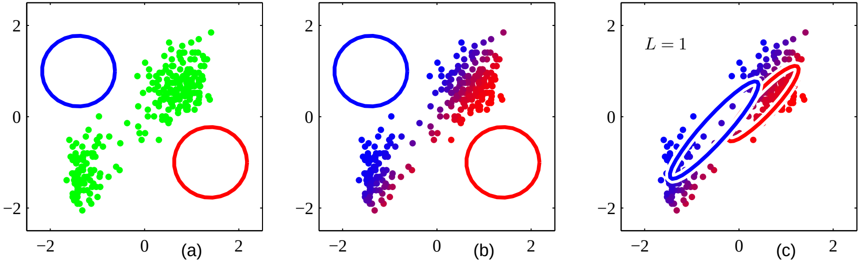
$$\gamma(z_k) \doteq p(z_k = 1|\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

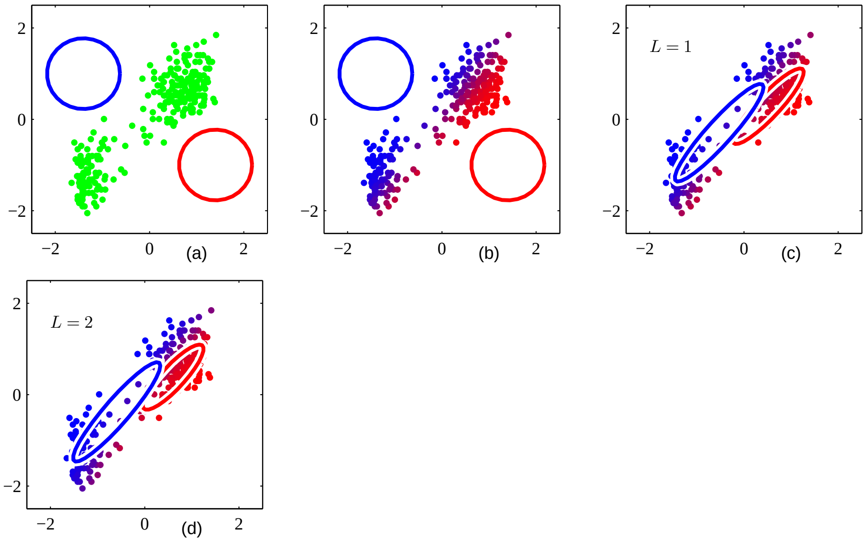
- But, these results do suggest an iterative scheme for finding a solution to the maximum likelihood problem.
  1. Chooce some initial values for the parameters, $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$.
  2. Use the current parameters estimates to compute the posteriors on the latent terms, i.e., the responsibilities.
  3. Use the responsibilities to update the estimates of the parameters.
  4. Repeat 2 and 3 until convergence.

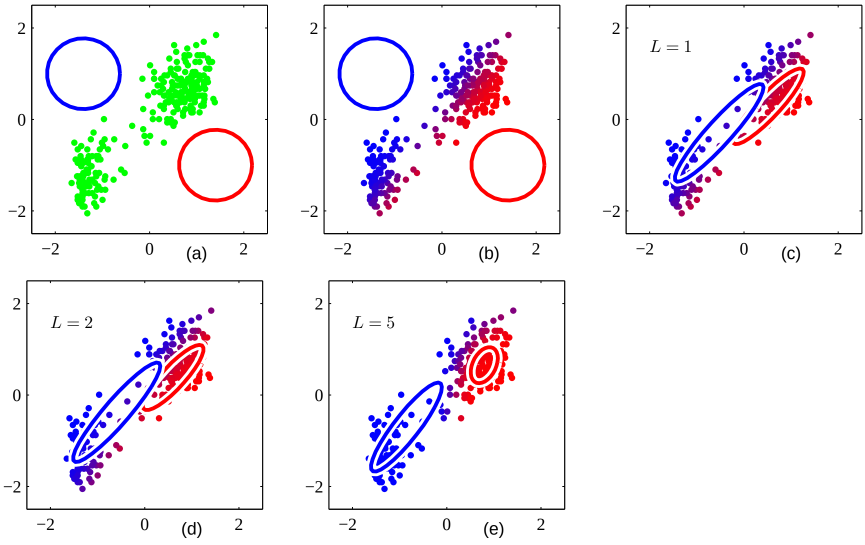


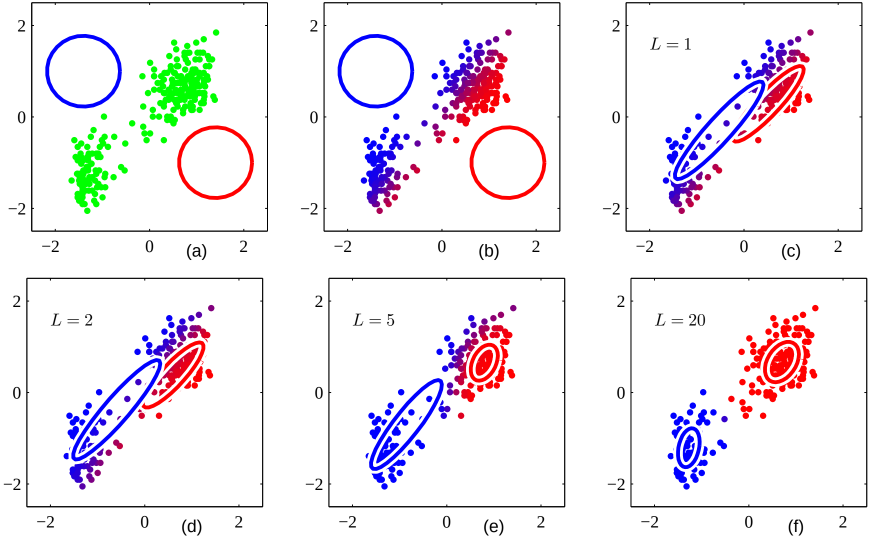
(a)


2
0
−2
−2
0
(a)
2
2
0
−2
−2
0
(b)
2


2
0
−2
−2
0
2
(a)
(b)
(c)
L = 1


2
0
−2
−2
0
2
(a)
(b)
(c)
(d)
$L = 1$
$L = 2$


2
0
−2
−2
0
2
(a)
(b)
(c)
(d)
(e)
$L = 1$
$L = 2$
$L = 5$


L = 1
L = 2
L = 5
L = 20
(a)
(b)
(c)
(d)
(e)
(f)

Some Quick, Early Notes on EM

- EM generally tends to take more steps than the K-Means clustering algorithm.

Some Quick, Early Notes on EM

- EM generally tends to take more steps than the K-Means clustering algorithm.
- Each step is more computationally intense than with K-Means too.

Some Quick, Early Notes on EM

- EM generally tends to take more steps than the K-Means clustering algorithm.
- Each step is more computationally intense than with K-Means too.
- So, one commonly computes K-Means first and then initializes EM from the resulting clusters.

Some Quick, Early Notes on EM

- EM generally tends to take more steps than the K-Means clustering algorithm.
- Each step is more computationally intense than with K-Means too.
- So, one commonly computes K-Means first and then initializes EM from the resulting clusters.
- Care must be taken to avoid singularities in the MLE solution.

Some Quick, Early Notes on EM

- EM generally tends to take more steps than the K-Means clustering algorithm.
- Each step is more computationally intense than with K-Means too.
- So, one commonly computes K-Means first and then initializes EM from the resulting clusters.
- Care must be taken to avoid singularities in the MLE solution.
- There will generally be multiple local maxima of the likelihood function and EM is not guaranteed to find the largest of these.

Given a GMM, the goal is to maximize the likelihood function with respect to the parameters (the means, the covarianes, and the mixing coefficients).

1. Initialize the means, $\boldsymbol{\mu}_k$, the covariances, $\boldsymbol{\Sigma}_k$, and mixing coefficients, $\boldsymbol{\pi}_k$. Evaluate the initial value of the log-likelihood.

Given a GMM, the goal is to maximize the likelihood function with respect to the parameters (the means, the covarianes, and the mixing coefficients).

1. Initialize the means, $\boldsymbol{\mu}_k$, the covariances, $\boldsymbol{\Sigma}_k$, and mixing coefficients, $\boldsymbol{\pi}_k$. Evaluate the initial value of the log-likelihood.
2. **E-Step** Evaluate the responsibilities using the current parameter values:

$$\gamma(z_k) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Given a GMM, the goal is to maximize the likelihood function with respect to the parameters (the means, the covarianes, and the mixing coefficients).

1. Initialize the means, $\boldsymbol{\mu}_k$, the covariances, $\boldsymbol{\Sigma}_k$, and mixing coefficients, $\boldsymbol{\pi}_k$. Evaluate the initial value of the log-likelihood.
2. **E-Step** Evaluate the responsibilities using the current parameter values:
$$\gamma(z_k) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$
3. **M-Step** Update the parameters using the current responsibilities
$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{29}$$
$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})^{\mathsf{T}} \tag{30}$$
$$\pi_k^{\text{new}} = \frac{N_k}{N} \tag{31}$$
where
$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{32}$$

4. Evaluate the log-likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}^{\text{new}}, \boldsymbol{\Sigma}^{\text{new}}, \boldsymbol{\pi}^{\text{new}}) = \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k^{\text{new}} \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k^{\text{new}}, \boldsymbol{\Sigma}_k^{\text{new}}\right) \right] \quad (33)$$

4. Evaluate the log-likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}^{\text{new}}, \boldsymbol{\Sigma}^{\text{new}}, \boldsymbol{\pi}^{\text{new}}) = \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k^{\text{new}} \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k^{\text{new}}, \boldsymbol{\Sigma}_k^{\text{new}}\right) \right] \tag{33}$$

5. Check for convergence of either the parameters of the log-likelihood. If the convergence is not satisfied, set the parameters:

$$\boldsymbol{\mu} = \boldsymbol{\mu}^{\text{new}} \tag{34}$$

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\text{new}} \tag{35}$$

$$\boldsymbol{\pi} = \boldsymbol{\pi}^{\text{new}} \tag{36}$$

and goto step 2.

A More General View of EM

- The goal of EM is to find maximum likelihood solutions for models having latent variables.

A More General View of EM

- The goal of EM is to find maximum likelihood solutions for models having latent variables.
- Denote the set of all model parameters as $\boldsymbol{\theta}$, and so the log-likelihood function is

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left[\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right] \tag{37}$$

A More General View of EM

- The goal of EM is to find maximum likelihood solutions for models having latent variables.
- Denote the set of all model parameters as $\boldsymbol{\theta}$, and so the log-likelihood function is

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left[\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right] \tag{37}$$

- Note how the summation over the latent variables appears inside of the log.
  - Even if the joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ belongs to the exponential family, the marginal $p(\mathbf{X}|\boldsymbol{\theta})$ typically does not.

A More General View of EM

- The goal of EM is to find maximum likelihood solutions for models having latent variables.
- Denote the set of all model parameters as $\boldsymbol{\theta}$, and so the log-likelihood function is

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left[\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right] \tag{37}$$

- Note how the summation over the latent variables appears inside of the log.
  - Even if the joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ belongs to the exponential family, the marginal $p(\mathbf{X}|\boldsymbol{\theta})$ typically does not.
- If, for each sample $\mathbf{x}_n$ we were given the value of the latent variable $\mathbf{z}_n$, then we would have a **complete** data set, $\{\mathbf{X}, \mathbf{Z}\}$, with which maximizing this likelihood term would be straightforward.

- However, in practice, we are not given the latent variables values.

- However, in practice, we are not given the latent variables values.
- So, instead, we focus on the expectation of the log-likelihood under the posterior distribution of the latent variables.

- However, in practice, we are not given the latent variables values.
- So, instead, we focus on the expectation of the log-likelihood under the posterior distribution of the latent variables.
- In the E-Step, we use the current parameter values $\boldsymbol{\theta}^{\text{old}}$ to find the posterior distribution of the latent variables given by $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.

- However, in practice, we are not given the latent variables values.
- So, instead, we focus on the expectation of the log-likelihood under the posterior distribution of the latent variables.
- In the E-Step, we use the current parameter values $\boldsymbol{\theta}^{\text{old}}$ to find the posterior distribution of the latent variables given by $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
- This posterior is used to define the **expectation of the complete-data log-likelihood**, denoted $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$, which is given by

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \tag{38}$$

- However, in practice, we are not given the latent variables values.
- So, instead, we focus on the expectation of the log-likelihood under the posterior distribution of the latent variables.
- In the E-Step, we use the current parameter values $\boldsymbol{\theta}^{\text{old}}$ to find the posterior distribution of the latent variables given by $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
- This posterior is used to define the **expectation of the complete-data log-likelihood**, denoted $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$, which is given by

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \tag{38}$$

- Then, in the M-step, we revise the parameters to $\boldsymbol{\theta}^{\text{new}}$ by maximizing this function:

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \tag{39}$$

- However, in practice, we are not given the latent variables values.
- So, instead, we focus on the expectation of the log-likelihood under the posterior distribution of the latent variables.
- In the E-Step, we use the current parameter values $\boldsymbol{\theta}^{\text{old}}$ to find the posterior distribution of the latent variables given by $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
- This posterior is used to define the **expectation of the complete-data log-likelihood**, denoted $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$, which is given by

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \tag{38}$$

- Then, in the M-step, we revise the parameters to $\boldsymbol{\theta}^{\text{new}}$ by maximizing this function:

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \tag{39}$$

- Note that the log acts directly on the joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ and so the M-step maximization will likely be tractable.