ISTA 421/521 Introduction to Machine Learning

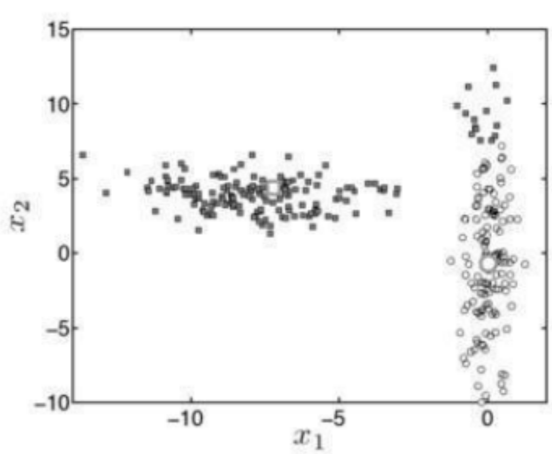
Lecture 24: Clustering
Gaussian Mixture Model
Expectation Maximization

Clay Morrison
clayton@sista.arizona.edu
Gould-Simpson 819
Phone 621-6609

20 November 2014

Mixture Models

- Some similarities to *K*-means, but much richer representations of the data (rather than points / centroids)

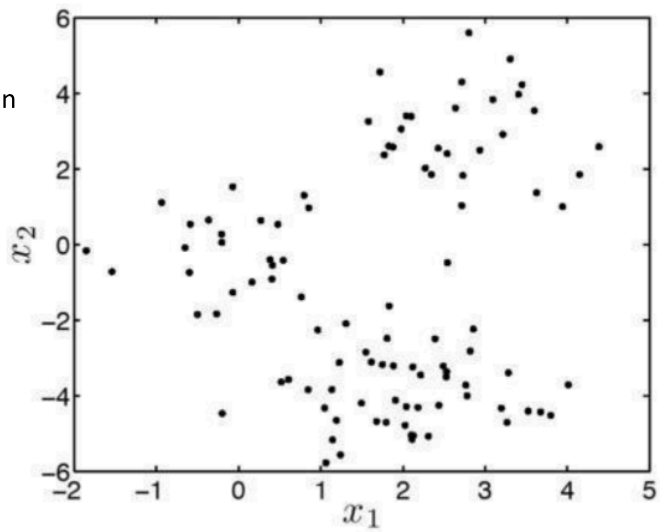


Centroids model of clusters is too simple to capture the structure here

The Generative Picture (again)

- How could we ***generate*** this data?

For each $\mathbf{x}_n$:
(1) Select one of three Gaussians
probability π_k for each Gaussian
(where $\Sigma_k \pi_k = 1$)
(2) Sample $\mathbf{x}_n$ from that Gaussian

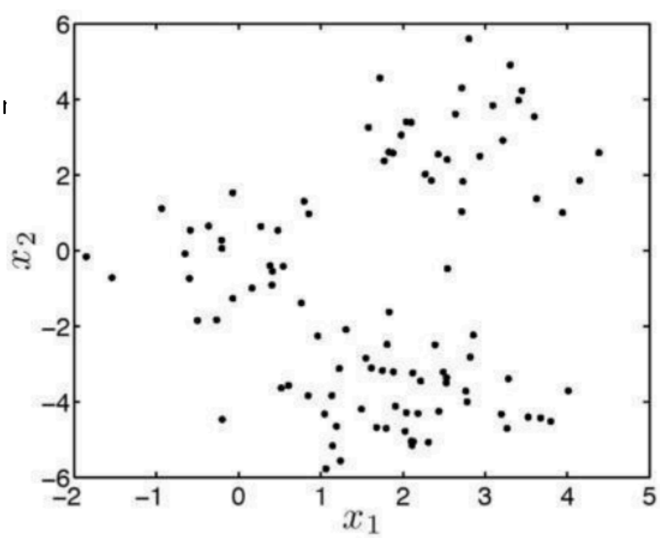


The Generative Picture (again)

- How could we ***generate*** this data?

For each $\mathbf{x}_n$:
(1) Select one of three Gaussians
probability π_k for each Gaussia
(where $\Sigma_k \pi_k = 1$)
(2) Sample $\mathbf{x}_n$ from that Gaussian

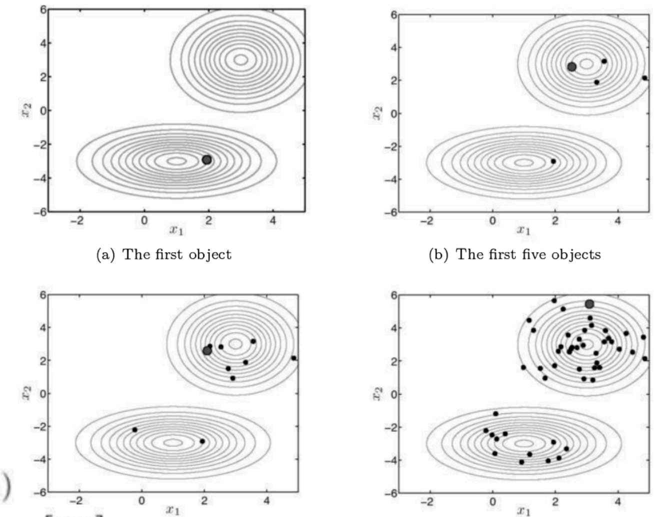
- Use $z_{nk} = 1$ to mean individual n was "sampled from" generator k ($z_{nj} = 0$ for all other $j \neq k$)
- Each Guassian is modeled with mean and covariance $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$



The Generative Picture (again)

- How could we ***generate*** this data?

For each $\mathbf{x}_n$:
(1) Select one of three Gaussians
probability π_k for each Gaussian
(where $\Sigma_k \pi_k = 1$)
(2) Sample $\mathbf{x}_n$ from that Gaussian



(a) The first object (b) The first five objects
The first 10 objects (d) The first 50 objects

- Use $z_{nk} = 1$ to mean individual n was "sampled from" generator k ($z_{nj} = 0$ for all other $j \neq k$)
- Each Guassian is modeled with mean and covariance $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$

$$p(\mathbf{x}_n|z_{nk}=1,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)=\mathcal{N}(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

$$\boldsymbol{\mu}_1=[3,3]^{\mathsf{T}},\ \boldsymbol{\Sigma}_1=\begin{bmatrix}1&0\\0&2\end{bmatrix} \qquad \boldsymbol{\mu}_2=[1,-3]^{\mathsf{T}},\ \boldsymbol{\Sigma}_2=\begin{bmatrix}2&0\\0&1\end{bmatrix}$$

$$\pi_1 = 0.7,\ \pi_2 = 0.3$$

Note axis scale; x_2 is being squashed

The EM Algorithm

- Our learning task: infer, from observed data, the component
  - **parameters** $(\boldsymbol{\mu}_k,\ \boldsymbol{\Sigma}_k)$ and π_k , ***and***
  - **assignments** z_{nk} of objects to components.
- Similar to *K*-means problem:
  - component **parameters** depend on **assignment**,
  - and **assignments** depend on **parameters**.
- In this probabilistic framework, we also have a two-step algorithm that alternates between steps until convergence; but now the steps are defined in terms of calculating
  - ***expectations*** (to update data-to-cluster assignments) and then
  - making adjustments to the parameters that ***maximize*** the likelihood (update cluster definitions: parameters)
- The **Expectation Maximization** (**EM**) algorithm.

Derive: the (general) Mixture Model Likelihood

Likelihood of getting individual $\mathbf{x}_n$ assuming it was generated from cluster k: $p(\mathbf{x}_n|z_{nk}=1,\Delta_k)$

Δ_k Represents the parameters of the kth density (In a moment, we'll use a Gaussian distribution, but could be any suitable density)

$\Delta = \{\Delta_1, \ldots, \Delta_K\}$ Collection of parameters for all of the mixture components

$\pi = \{\pi_1, \ldots, \pi_K\}$ Collection of all of the probabilities of the mixture components

Or goal: find the likelihood of the data object $\mathbf{x}_n$ under the whole model: $p(\mathbf{x}_n \mid \Delta, \pi)$

Start with likelihood for one cluster:

$$p(\mathbf{x}_n|z_{nk}=1,\Delta)$$

Need to "get rid" of z_{nk} (i.e., over all mixtures)

Multiply both sides by the probability that object n is in cluster k:

$$p(\mathbf{x}_n|z_{nk}=1,\Delta)p(z_{nk}=1) = p(\mathbf{x}_n|\Delta_k)p(z_{nk}=1)$$

The probability chain rule: P(a|b,c)xP(b) = P(a,b|c)

By definition: $p(z_{nk}=1)=\pi_k$

$$p(\mathbf{x}_n, z_{nk}=1|\Delta,\boldsymbol{\pi}) = p(\mathbf{x}_n|\Delta_k)\pi_k$$

Marginalize over all of the individual components:

$$\sum_{k=1}^{K} p(\mathbf{x}_n, z_{nk}=1|\Delta,\boldsymbol{\pi}) = \sum_{k=1}^{K} p(\mathbf{x}_n|\Delta_k)\pi_k$$

$$p(\mathbf{x}_n|\Delta,\boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\Delta_k)$$

Make standard assumption that all data points are independent (given their mixture):

This is the likelihood of all N data points

$$p(\mathbf{X}|\Delta,\boldsymbol{\pi}) = \prod_{n=1}^{N}\sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\Delta_k)$$

$$p(\mathbf{X}|\Delta,\boldsymbol{\pi}) = \prod_{n=1}^{N}\sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\Delta_k)$$

Maximizing the Mixture Model Likelihood

- Now we'll look at an instance of the **EM** algorithm for Gaussian mixtures: a **G**aussian **M**ixture **M**odel.
- We'll want to do maximization, so easier to work with the logarithm of the likelihood

$$L = \log p(\mathbf{X}|\Delta,\boldsymbol{\pi}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- **Immediate problem!**: the summation inside the log makes finding the optimal parameters challenging

  We want to take partial derivatives w.r.t. $\boldsymbol{\mu}_k,\ \boldsymbol{\Sigma}_k,\ \pi_k$

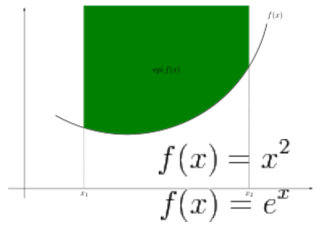
- **Trick**: derive a lower-bound on L and maximize ***that***

SISTA

Jensen's Inequality

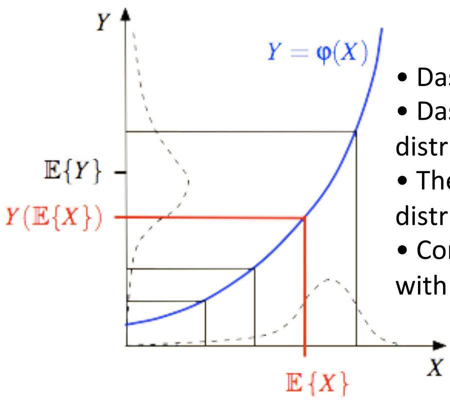
- A very general result that has wide application in convex optimization and probability theory
- Generally: relates the value of a convex function of an integral to the integral of the convex function
- Simplest probabilistic form: convex transformation of a mean is less than or equal to the mean after convex transformation
- If φ is a convex function: $\boxed{\varphi\left(\mathbb{E}[X]\right) \leq \mathbb{E}\left[\varphi(X)\right]}$.

Convex function



A visual proof (for probabilistic case):

(Concave functions, e.g., log(x), just reverse the inequality: $\varphi(\mathbb{E}[X]) \geq \mathbb{E}[\varphi(X)]$)



- Dashed curve along X axis is the hypothetical distribution of X,
- Dashed curve along Y axis is corresponding convex-mapped distribution of Y values ($Y = \varphi(X)$).
- The convex mapping $Y(X)$ increasingly "stretches" the distribution for increasing values of X.
- Consequently, the expectation of Y will always shift upwards with respect to the position of $\varphi(E\{X\})$, thus:

$$\mathbb{E}\{Y\} = \mathbb{E}\{\varphi(X)\} \geq \varphi(\mathbb{E}\{X\}),$$

$$L = \log p(\mathbf{X}|\Delta, \boldsymbol{\pi}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$ Original log-likelihood

$$\log \mathbf{E}_{p(z)}\left\{f(z)\right\} \geq \mathbf{E}_{p(z)}\left\{\log f(z)\right\}$$ Jensen's inequality ('concave' form)

- Need to make right-hand side of log-likelihood look like the log of an *expectation*.
  - (Note: it sort of does now with π_k, except that we want to keep π_k around in order to maximize w.r.t. it!)
- Multiply *and* divide the expression in the summation over k by a new variable, q_{nk}

$$L = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \frac{q_{nk}}{q_{nk}}$$

- Restrict q_{nk} to be positive and sum to 1 over k
  - I.e., q_{nk} is some probability distribution over the K components for the nth object (like a "soft" version of the z_{nk} membership function)
- Now rewrite the above as an expectation w.r.t. q_{nk}:

$$L = \sum_{n=1}^{N} \log \sum_{k=1}^{K} q_{nk} \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{q_{nk}} \quad = \sum_{n=1}^{N} \log \mathbf{E}_{q_{nk}} \left\{ \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{q_{nk}} \right\}$$

- Now apply Jensen's inequality to get the lower bound we desire: $\mathcal{B}$

$$L = \sum_{n=1}^{N} \log \mathbf{E}_{q_{nk}} \left\{ \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{q_{nk}} \right\} \geq \boxed{\sum_{n=1}^{N} \mathbf{E}_{q_{nk}} \left\{ \log \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{q_{nk}} \right\}}$$

More Algebra to make it nicer…

$$\mathcal{B} = \sum_{n=1}^{N} \mathbf{E}_{q_{nk}} \left\{ \log \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{q_{nk}} \right\}$$
$$= \sum_{n=1}^{N} \sum_{k=1}^{K} q_{nk} \log \left(\frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{q_{nk}} \right)$$
$$= \sum_{n=1}^{N} \sum_{k=1}^{K} q_{nk} \log \pi_k + \sum_{n=1}^{N} \sum_{k=1}^{K} q_{nk} \log p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) - \sum_{n=1}^{N} \sum_{k=1}^{K} q_{nk} \log q_{nk}.$$

(We just used the expectation form in order to make use of Jensen's inequality…)

The parameters we now want to adjust in order to (locally) maximize *B*, which in turn corresponds to local maxima of *L*: $q_{nk},\ \boldsymbol{\pi},\ \boldsymbol{\mu}_k,\ \boldsymbol{\Sigma}_k$

Now to get down to business with the actual EM algorithm…

Where we are headed: we will find equations that are updates to the params that maximize the likelihood (via *B*). – so maximize *B* w.r.t. the parameters.

But, $\pi_k,\ \boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ will turn out to be in terms of q_{nk}

So make EM an iterative algorithm (analogous to *K*-means):

- Update $\pi_k,\ \boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ with q_{nk} (i.e., the assignments) fixed (**M** step: maximize)
- Update q_{nk} (by taking the expectation w.r.t. the unknown q_{nk} assignments – **E** step)

Maximizing *B*

$$\mathcal{B} = \boxed{\sum_{n=1}^{N} \sum_{k=1}^{K} q_{nk} \log \pi_k} + \sum_{n=1}^{N} \sum_{k=1}^{K} q_{nk} \log p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) - \sum_{n=1}^{N} \sum_{k=1}^{K} q_{nk} \log q_{nk}$$

Find values of $q_{nk}, \underline{\pi_k}, \mu_k, \Sigma_k$ that correspond to a local maxima of

Update for π_k

The "mixture" probability: $\sum_k \pi_k = 1$

Thus, optimization w.r.t. π_k is constrained

Use Lagrange terms to capture constraint! Add it to B (still only involving π_k)

$$\mathcal{B} = \sum_{n=1}^{N} \sum_{k=1}^{K} q_{nk} \log \pi_k - \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) + \ldots$$

Take partial derivative w.r.t. π_k and set to 0

$$\frac{\partial \mathcal{B}}{\partial \pi_k} = \frac{\sum_{n=1}^{N} q_{nk}}{\pi_k} - \lambda = 0$$
$$\sum_{n=1}^{N} q_{nk} = \lambda \pi_k.$$

Now have a λ – solve for it:
Sum both sides over *k*, b/c that eliminates π_k and q_{nk}

$$\sum_{k=1}^{K} \sum_{n=1}^{N} q_{nk} = \lambda \sum_{k=1}^{K} \pi_k$$
$$\sum_{n=1}^{N} 1 = \lambda$$
$$\lambda = N$$

Plug in N for λ – now have π_k:

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} q_{nk}$$

Yay! One down.

Maximizing *B*

$$\mathcal{B} = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log\pi_k + \boxed{\sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} - \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log q_{nk}$$

Find values of $q_{nk}, \underline{\pi_k}, \underline{\mu_k}, \Sigma_k$ that correspond to a local maxima of

Update for μ_k

$p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ in this case will be a multivariate Gaussian, so rewrite *B*

$$\mathcal{B} = \ \dots + \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log\left(\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_k|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\right)\right) + \dots$$

$$= \ \dots - \cancel{\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log\left((2\pi)^d|\boldsymbol{\Sigma}_k|\right)} - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k) + \dots$$

Take partial derivative w.r.t. μ_k and set to 0

Use this linear algebra derivative fact:

$$f(\mathbf{w}) = \mathbf{w}^{\mathsf{T}}\mathbf{C}\mathbf{w}, \quad \frac{\partial f(\mathbf{w})}{\partial \mathbf{w}} = 2\mathbf{C}\mathbf{w}$$

… and chain rule of derivatives:

$$(f \circ g)'(t) = f'(g(t))g'(t).$$

$$\frac{\partial \mathcal{B}}{\partial \boldsymbol{\mu}_k} = -\frac{1}{2}\sum_{n=1}^{N} q_{nk} \times \frac{\partial\ (\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)}{\partial\ (\mathbf{x}_n - \boldsymbol{\mu}_k)} \times \frac{\partial\ (\mathbf{x}_n - \boldsymbol{\mu}_k)}{\partial\ \boldsymbol{\mu}_k}$$

$$= \sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k).$$

Set to 0 and solve for μ_k

$$\sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0$$

$$\sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n = \sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k$$

$$\sum_{n=1}^{N} q_{nk}\mathbf{x}_n = \boldsymbol{\mu}_k\sum_{n=1}^{N} q_{nk}$$

Two down.

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} q_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} q_{nk}}.$$

Maximizing *B*

$$\mathcal{B} = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log\pi_k + \boxed{\sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} - \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log q_{nk}$$

Find values of $q_{nk}, \underline{\pi_k}, \underline{\mu_k}, \underline{\Sigma_k}$ that correspond to a local maxima of

Update for Σ_k

… also only shows up in the term with $p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

$$\mathcal{B} = \ \dots - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log\left(\cancel{(2\pi)^d}|\boldsymbol{\Sigma}_k|\right) - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k) + \dots$$

Take partial derivative w.r.t. Σ_k and set to 0

Use these two linear algebra derivative facts:

$$\frac{\partial \log|\mathbf{C}|}{\partial \mathbf{C}} = (\mathbf{C}^{\mathsf{T}})^{-1} \qquad \frac{\partial \mathbf{a}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{b}}{\partial \mathbf{C}} = -(\mathbf{C}^{\mathsf{T}})^{-1}\mathbf{a}\mathbf{b}^{\mathsf{T}}(\mathbf{C}^{\mathsf{T}})^{-1}$$

$$\frac{\partial \mathcal{B}}{\partial \boldsymbol{\Sigma}_k} = -\frac{1}{2}\sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1} + \frac{1}{2}\sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1}$$

Note:

$$\boldsymbol{\Sigma}_k^{\mathsf{T}} = \boldsymbol{\Sigma}_k$$

Set to 0 and solve for Σ_k

$$-\frac{1}{2}\sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1} + \frac{1}{2}\sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1} = 0$$

$$\frac{1}{2}\sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1} = \frac{1}{2}\sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1}$$

Pre and post multiply by Σ_k removes all Σ_k^{-1}

$$\boldsymbol{\Sigma}_k\sum_{n=1}^{N} q_{nk}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_k^{-1}\sum_{n=1}^{N} q_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_k$$

$$\boldsymbol{\Sigma}_k\sum_{n=1}^{N} q_{nk} = \sum_{n=1}^{N} q_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}$$

Three down.

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} q_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}}{\sum_{n=1}^{N} q_{nk}}.$$

Maximizing $\mathcal{B}$

$$\mathcal{B} = \boxed{\sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log \pi_k + \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log p(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) - \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log q_{nk}}$$

Find values of $q_{nk}, \pi_k, \mu_k, \boldsymbol{\Sigma}_k$ that correspond to a local maxima of

Update for q_{nk}

Shows up in all three terms! And has this constraint: $\sum_{k=1}^{K} q_{nk} = 1$

Lagrangian term for constraint

$$\mathcal{B} = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log \pi_k + \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log p(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_K) - \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log q_{nk} - \lambda\left(\sum_{k=1}^{K} q_{nk} - 1\right)$$

Take partial derivative w.r.t. q_{nk} and set to 0

Need the derivative product rule: for $f(a) = g(a)h(a)$

$$\frac{\partial f(a)}{\partial a} = g(a)\frac{\partial h(a)}{\partial a} + \frac{\partial g(a)}{\partial a}h(a)$$

$$\frac{\partial \mathcal{B}}{\partial q_{nk}} = \log \pi_k + \log p(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) - (1 + \log q_{nk}) - \lambda.$$

Set to 0, rearrange, exponentiate, and solve for q_{nk}

$$\begin{aligned} 1 + \log q_{nk} + \lambda &= \log \pi_k + \log p(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \\ \exp(\log q_{nk} + (\lambda + 1)) &= \exp(\log \pi_k + \log p(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)) \\ q_{nk}\exp(\lambda+1) &= \pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k). \end{aligned}$$

Same trick we used for μ_k: sum over k on both sides makes q_{nk} go to 1 on the left side:

$$\begin{aligned} \exp(\lambda+1)\sum_{k=1}^{K} q_{nk} &= \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \\ \exp(\lambda+1) &= \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k). \end{aligned}$$

Can substitute this here and solve for q_{nk}:

$$q_{nk} = \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}$$

Done!

Expectation Maximization (EM) for GMM

$$\mathcal{B} = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log \pi_k + \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log p(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) - \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\log q_{nk}$$

The final expressions:

The mean value of q_{nk} for a particular cluster k :
Average of all posterior probabilities of belonging to cluster k
i.e., the expected proportion of the data belonging to cluster k

$$\pi_k = \frac{1}{N}\sum_{n=1}^{N} q_{nk}$$

Average of the data objects weighted by q_{nk} :
When all cluster membership is such that posterior prob's are 0 or 1, then this is just the proportion of the data assigned to component k

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} q_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} q_{nk}} \qquad \boldsymbol{\mu}_k = \frac{\sum_n z_{nk}\mathbf{x}_n}{\sum_n z_{nk}}$$

weighted covariance :

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} q_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}}{\sum_{n=1}^{N} q_{nk}}$$

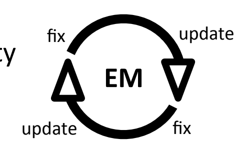
Looks like Bayes' Rule! :

$$q_{nk} = \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}$$

Posterior probability of object n belonging to cluster k

$$p(z_{nk} = 1|\mathbf{x}_n, \boldsymbol{\pi}, \Delta) = \frac{p(z_{nk} = 1|\pi_k)p(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} p(z_{nj} = 1|\pi_j)p(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} = q_{nk}.$$

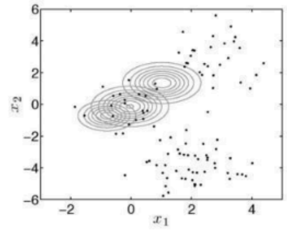
Cluster membership probability
q_{nk}
E step
(expected value of unknown assignments z_{nk})



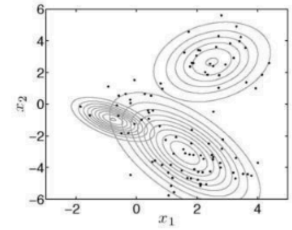
All about the model
$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\pi}$
M step
(maximizing w.r.t cluster membership)

Running EM with a GMM

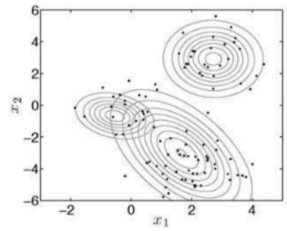
- Initialize the parameters:
  - Mixture paramters: random means and covariances
  - Mixture priors: could choose uniform $\pi_k = 1/K$



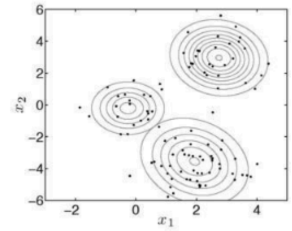
(a) The three randomly initialised Gaussian mixture components



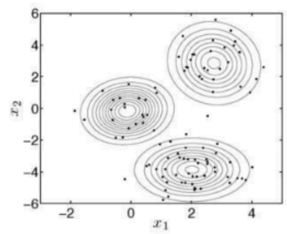
(b) The three components after one iteration of the EM algorithm



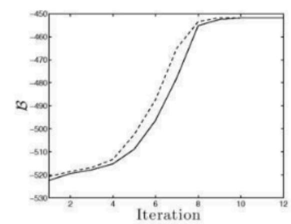
(c) The three components after five iterations of the EM algorithm



(d) The three components after seven iterations of the EM algorithm



(e) The three components at convergence of the EM algorithm



(f) The evolution of the bound $\mathcal{B}$ (solid line, Equation 6.8) and log likelihood L (dashed line, Equation 6.5)

To find cluster membership:

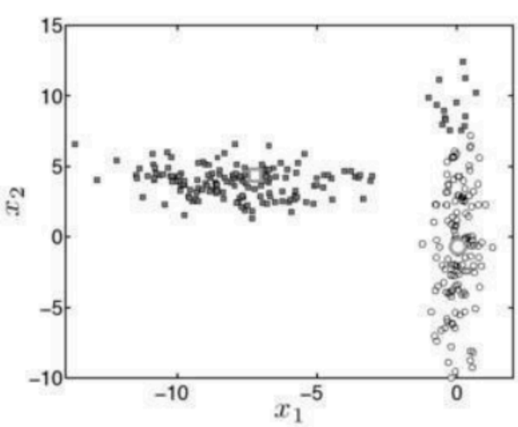
q_{nk} = posterior prob of n belonging to k

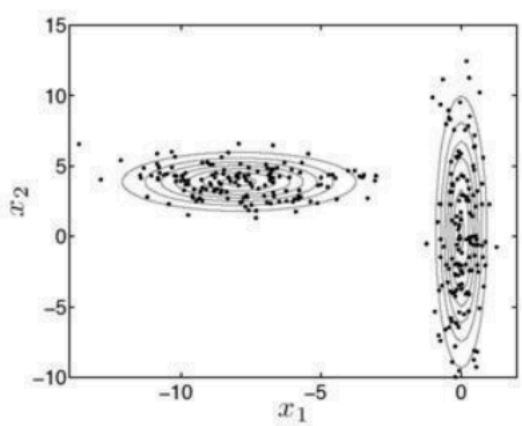
Hard assignment: for each indiv. n, choose the k with the highest q_{nk}

Or, consider the distribution – reveals interesting relationships of individual to more than one cluster; e.g.,

$$q_{n1} = 0.53, \quad q_{n2} = 0.45, \quad q_{n3} = 0.02$$

GMM does well on problem hard for *K*-means

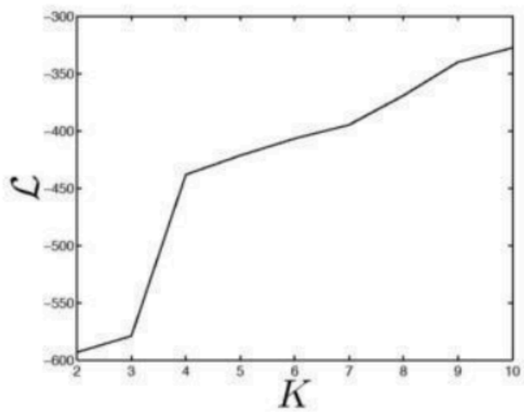




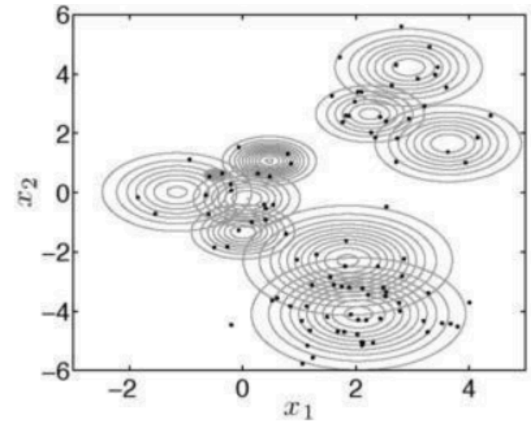
GMM with K=2

Choosing K

- Similar to *K*-means, where we can't just choose the *K* that minimizes the total distances (*K*=*N* → *D*=0), we can't just choose the *K* that maximizes the log likelihood *L* (or bound *B*); it increases with more mixtures:



(a) The increase in model likelihood as the number of components increases



(b) An example of the model at convergence for $K = 10$

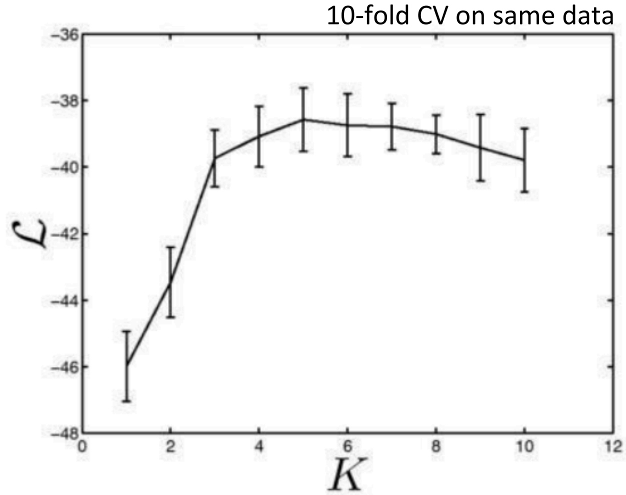
The power of a generative model

- **However**: because mixture models are generative models, we can run cross-validation:
  - For each potential *K*: Hold out data, fit mixtures, then measure likelihood of held-out data

It's not perfect (somewhere between 3 and 8)

But with non-generative K-means, we have no direct basis for choosing K

10-fold CV on same data



Other Mixtures besides Gaussians

- Could be any probability density

$$p(\mathbf{X}|\Delta, \boldsymbol{\pi}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\Delta_k)$$

- For D=10 dimensional binary data, e.g.,

$$\mathbf{x}_n = [0, 1, 0, 1, 1, 1, 0, 0, 0, 1]$$

  – Represent as product of (i.e., independent) Bernoulli distributions:

$$p(\mathbf{x}_n|\mathbf{p}_k) = \prod_{d=1}^{D} p_{kd}^{x_{nd}} (1 - p_{kd})^{1-x_{nd}}$$

Mixture probabilities

$$\mathbf{p}_k = [p_{k1}, \ldots, p_{kD}]^{\mathsf{T}}$$

EM is General!

A general pattern:

$\mathbf{X}$ Observed data (may be discrete or continuous; possibly vector-valued for each observation)

$\mathbf{Z}$ Unobserved (latent, missing) data/values (one per observed data point; discrete indicator fn in *hard* EM, continuous probability of indicator in *soft* EM)

$\boldsymbol{\theta}$ Unknown parameters (continuous)

$L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ A likelihood function

The maximum likelihood estimate (MLE) of the unknown parameters is determined by the marginal likelihood of the observed data: $L(\boldsymbol{\theta}; \mathbf{X}) = p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$

Typically this is intractable (***Z*** may be exponential or worse in size)

EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying the following two steps:

(E) Expectation step: Calculate the expected value of the log likelihood function with respect to the conditional distribution of ***Z*** given ***X*** under the current estimate of the parameters $\boldsymbol{\theta}^{(t)}$ $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(t)}} [\log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})]$ (What typically happens here is updating ***Z*** under fixed $\boldsymbol{\theta}^{(t)}$ and data ***X***.)

(M) Maximization step: Find the parameter(s), $\boldsymbol{\theta}^{(t+1)}$, that maximize(s) this quantity

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$

Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society*, Series **B** 39 (1): 1–38.