



Lecture 12

Expectation Maximization and Clustering

Hannah Marienwald

Lecture based on Bishop, C. (2006). Pattern Recognition and Machine Learning. Ch. 9.

Most Figures are also taken from there.

Recap

- ▶ **Supervised Learning:**

Training data comprises of input data and their corresponding target label

Goal: find mapping between input and label

Examples: classification, regression, ...

- ▶ **Unsupervised Learning:**

Training data only consist of the input data without labels

Goal: find structure in the data

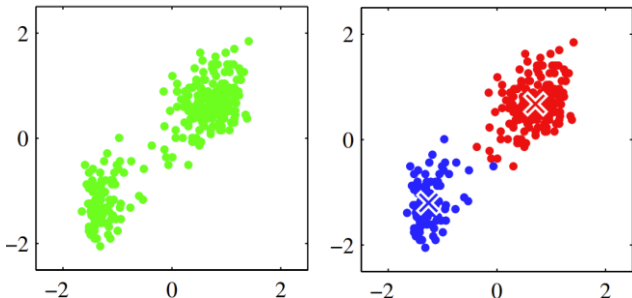
Examples: **clustering**, density estimation, PCA, ...

Outline

- ▶ K-means Clustering
- ▶ Gaussian Mixture Models (GMM)
- ▶ Expectation Maximization (EM)

Clustering

- ▶ **Input:** N data points in d -dimensions, $\{x_n\}_{n=1}^N \subseteq \mathbb{R}^d$
- ▶ **Goal:** partition the data into K clusters based on distances
- ▶ **Cluster:** set of data points whose inter-point distances are small compared to the distances to points outside of the cluster



K-Means

- ▶ **Input:** N data points in d -dimensions, $\{x_n\}_{n=1}^N \subseteq \mathbb{R}^d$
- ▶ **Goal:** partition the data into K clusters based on distances
- ▶ **Idea:** assign each data point to the cluster with the closest *cluster center* μ_k . Define the *assignments*

$$z_{n,k} = \begin{cases} 1, & \text{if } x_n \text{ belongs to cluster } k \\ 0, & \text{otherwise} \end{cases}$$

- ▶ **Objective:** minimize

$$J = \sum_{n=1}^N \sum_{k=1}^K z_{n,k} \cdot \|x_n - \mu_k\|^2$$

- ▶ K -Means finds $z_{n,k}$ and μ_k for all n, k by minimizing J

K-Means

Input: data points $\{x_n\}_{n=1}^N$, number of clusters K

Returns: cluster center $\{\mu_k\}_{k=1}^K$, cluster assignments

$\{z_{n,k}\}_{n,k=1}^{N,K}$

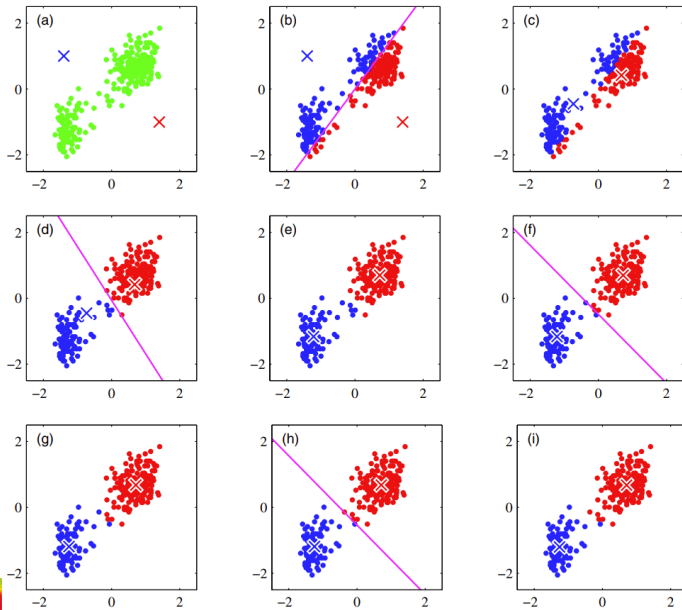
1. Randomly initialize μ_k for all $k \in \{1, \dots, K\}$
2. Loop until convergence:
 - 2.1. Minimize J wrt. $z_{n,k}$ and keep μ_k fixed:
assign x_n to the closest center

$$z_{n,k} = \begin{cases} 1, & \text{if } k = \underset{j}{\operatorname{argmin}} \|x_n - \mu_j\|^2 \\ 0, & \text{otherwise} \end{cases}$$

- 2.2. Minimize J wrt. μ_k and keep $z_{n,k}$ fixed:

$$\text{update the cluster centers } \mu_k = \frac{\sum_{n=1}^N z_{n,k} \cdot x_n}{\sum_{n=1}^N z_{n,k}}$$

K-Means



How to determine K ?

- ▶ By prior knowledge otherwise:
- ▶ Heuristic 1:
Plot J for different values of K . In general, $J \rightarrow 0$ as $K \rightarrow N$ but there might be a dip which indicates a good value.
- ▶ Heuristic 2:
For a suitable K the cluster centers and assignments are stable across different runs (started from different random initializations). Rerun K -means several times for different values of K and choose the one for which the results change the least.

K -Means

- ▶ Each iteration reduces value of J
- ▶ Guaranteed convergence
- ▶ Might only find local instead of global minimum
- ▶ Extension to general dissimilarity measures possible
- ▶ Hard assignment (every data points belongs to exactly one cluster)

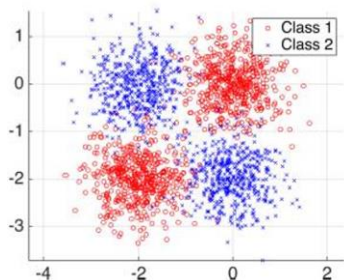
Gaussian Mixture Models (GMM)

Density Estimation

- Multivariate Gaussian distribution

$$\begin{aligned}\mathcal{N}(x|\mu, \Sigma) &= p(x|\mu, \Sigma) \\ &= \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \cdot \exp \left[-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]\end{aligned}$$

- Single Gaussian might not always be a good fit



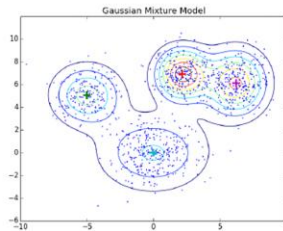
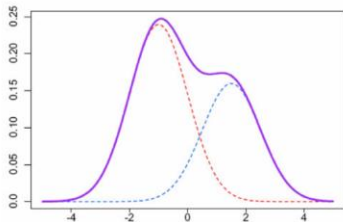
Mixture of Gaussians

- ▶ Linear superposition of K Gaussians

$$p(x|\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^K \tau_k \cdot \mathcal{N}(x|\mu_k, \Sigma_k)$$

- ▶ with prior probabilities

$$0 \leq \tau_k \leq 1 \text{ and } \sum_{k=1}^K \tau_k = 1$$



Figures from Eugene Weinstein, Yu Zhu

Maximum Likelihood for GMMs

- ▶ Find maximum likelihood estimators for $\theta = (\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ for data set $X = (x_1, \dots, x_N)$?
- ▶ Log-likelihood

$$\begin{aligned} L(\theta) &= \log p(X|\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \log \left[\prod_{n=1}^N \sum_{k=1}^K \tau_k \cdot \mathcal{N}(x_n | \mu_k, \Sigma_k) \right] \\ &= \sum_{n=1}^N \log \left[\sum_{k=1}^K \tau_k \cdot \mathcal{N}(x_n | \mu_k, \Sigma_k) \right] \\ &= \sum_{n=1}^N \log \left[\sum_{k=1}^K \tau_k \cdot \left(\frac{(2\pi)^{-\frac{d}{2}}}{|\Sigma_k|^{\frac{1}{2}}} \cdot \exp \left[-\frac{1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) \right] \right) \right] \end{aligned}$$

- ▶ Difficult to optimize and no analytic solution.

GMMs with Latent Variables

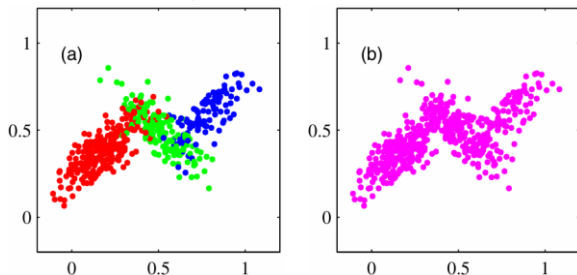
- **Trick:** introduce auxiliary variables indicating from which Gaussian each data point was sampled

$$z_{n,k} = \begin{cases} 1, & \text{if } x_n \text{ belongs to Gaussian } k \\ 0, & \text{otherwise} \end{cases}$$

$$p(z_{n,k} = 1) = \tau_k$$

$$p(x_n | z_{n,k} = 1) = \mathcal{N}(x_n | \mu_k, \Sigma_k)$$

- The assignments $z_{n,k}$ are *latent*, i.e., not observed



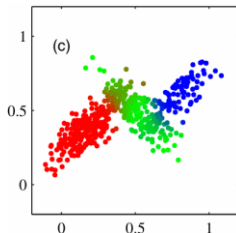
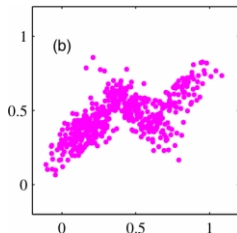
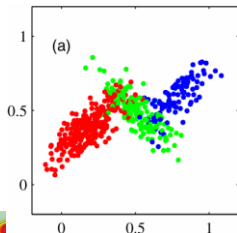
GMMs with Latent Variables

- ▶ Latent assignments

$$z_{n,k} = \begin{cases} 1, & x_n \text{ belongs to } k \\ 0, & \text{otherwise} \end{cases}, \quad p(z_{n,k} = 1) = \tau_k, \quad p(x_n | z_{n,k} = 1) = \mathcal{N}(x_n | \mu_k, \Sigma_k)$$

- ▶ Using Bayes rule

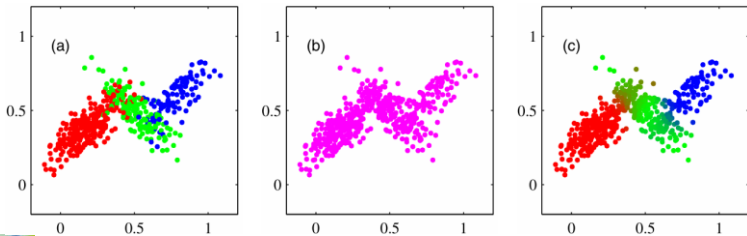
$$\begin{aligned} p(z_{n,k} = 1 | x_n) &= \frac{p(z_{n,k} = 1) \cdot p(x_n | z_{n,k} = 1)}{\sum_{j=1}^K p(z_{n,j} = 1) \cdot p(x_n | z_{n,j} = 1)} \\ &= \frac{\tau_k \mathcal{N}(x_n | \mu_k, \Sigma_k)}{\sum_j \tau_j \mathcal{N}(x_n | \mu_j, \Sigma_j)} \end{aligned}$$



GMMs for Clustering

- ▶ GMMs can be used to find clustering
- ▶ $p(z_{n,k} = 1|x_n)$ is *soft-assignment* to cluster k
- ▶ For clustering, we need to determine optimal values of $\theta = (\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $p(z_{n,k} = 1|x_n)$
- ▶ Maximum (log) likelihood hard to optimize

$$\log p(X|\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \log \left[\sum_{k=1}^K \tau_k \cdot \mathcal{N}(x_n|\mu_k, \Sigma_k) \right]$$



EM for GMMs



- ▶ No closed form for maximum (log) likelihood solution, because optimal $(\tau_k, \mu_k, \Sigma_k)$ depend on $p(z_{n,k} = 1 | x_n)$ which in turn depends on $(\tau_k, \mu_k, \Sigma_k)$
- ▶ **Expectation Maximization:** finds maximum likelihood solutions for models with latent variables
- ▶ Each iteration is guaranteed to increase the log likelihood
- ▶ Not guaranteed to find global optimum only local

EM for GMMs

Input: data points $\{x_n\}_{n=1}^N$, number of clusters K

Returns: GMM parameters $(\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, soft assignments

$$\{p(z_{n,k} = 1 | x_n)\}_{n,k=1}^{N,K}$$

1. Randomly initialize $(\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ for all $k \in \{1, \dots, K\}$
2. Loop until convergence:
 - 2.1. Update $p(z_{n,k} = 1 | x_n)$ and keep $(\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ fixed. 
 - 2.2. Update $(\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and keep $p(z_{n,k} = 1 | x_n)$ fixed. 

(formulas on next slide)

EM for GMMs

- Recall

$$p(z_{n,k} = 1 | x_n) = \frac{\tau_k \mathcal{N}(x_n | \mu_k, \Sigma_k)}{\sum_j \tau_j \mathcal{N}(x_n | \mu_j, \Sigma_j)}$$

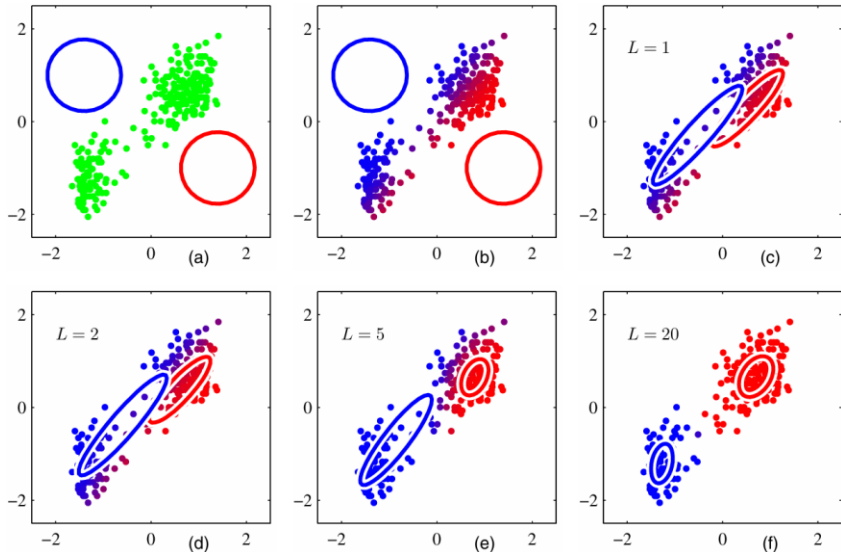
and set

$$N_k = \sum_{n=1}^N p(z_{n,k} = 1 | x_n)$$

- Update formulas are obtained by setting the derivative of $\log p(X | \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ to 0:

$$\begin{aligned}\mu_k &= \frac{1}{N_k} \sum_{n=1}^N p(z_{n,k} = 1 | x_n) \cdot x_n \\ \Sigma_k &= \frac{1}{N_k} \sum_{n=1}^N p(z_{n,k} = 1 | x_n) \cdot (x_n - \mu_k)(x_n - \mu_k)^T \\ \tau_k &= \frac{N_k}{N} \quad (\text{constrained optimization})\end{aligned}$$

EM for GMMs



K-Means

- ▶ K -means is a limit of Gaussian mixture models with EM where for all k , $\Sigma_k = \sigma^2 I$ are fixed (not optimized) and $\sigma^2 \rightarrow 0$. This yields *hard* assignments and K -means.

Input: data points $\{x_n\}_{n=1}^N$, number of clusters K

Returns: cluster center $\{\mu_k\}_{k=1}^K$, cluster assignments $\{z_{n,k}\}_{n,k=1}^{N,K}$

1. Randomly initialize μ_k for all $k \in \{1, \dots, K\}$
2. Loop until convergence:

- 2.1. Minimize J wrt. $z_{n,k}$ and keep μ_k fixed:
assign x_n to the closest center

$$z_{n,k} = \begin{cases} 1, & \text{if } k = \underset{j}{\operatorname{argmin}} \|x_n - \mu_j\|^2 \\ 0, & \text{otherwise} \end{cases}$$



- 2.2. Minimize J wrt. μ_k and keep $z_{n,k}$ fixed:

update the cluster centers $\mu_k = \frac{\sum_{n=1}^N z_{n,k} \cdot x_n}{\sum_{n=1}^N z_{n,k}}$

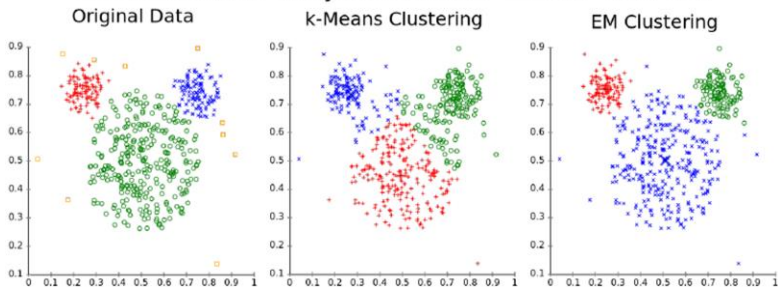


Comparison

In contrast to K -means, GMM allows for

- ▶ unequal cluster variances
- ▶ unequal cluster probabilities
- ▶ non-spherical clusters
- ▶ soft cluster

Different cluster analysis results on "mouse" data set:



(Figure taken from Wikipedia)

Expectation Maximization (EM)

Expectation Maximization

- ▶ **Goal:** find maximum likelihood solutions for models having latent variables
- ▶ **Notation:**
 - X : observed variables $\{X, Z\}$: complete data set
 - Z : latent variables $\{X\}$: incomplete data set
 - θ : model parameters
- ▶ **Difficulty:** optimizing the incomplete-data log likelihood $\log p(X|\theta) = \log[\sum_Z p(X, Z|\theta)]$ is hard because of sum inside the log, but maximizing the complete-data log likelihood $\log p(X, Z|\theta)$ is straightforward
- ▶ Only information we have about Z is through its posterior distribution $p(Z|X, \theta)$
- ▶ **Idea:** consider instead the expectation of $\log p(X, Z|\theta)$ under posterior $p(Z|X, \theta)$

Expectation Maximization

- ▶ **Goal:** find maximum likelihood solutions for models having latent variables
- ▶ **Notation:**
 - X : observed variables $\{X, Z\}$: complete data set
 - Z : latent variables $\{X\}$: incomplete data set
 - θ : model parameters
- ▶ **Idea:** consider instead the expectation of $\log p(X, Z|\theta)$ under posterior $p(Z|X, \theta)$

E-step

Calculate *expectation* of $\log p(X, Z|\theta)$ under $p(Z|X, \theta)$



M-step

Maximize this expectation for model parameters θ

General EM Algorithm

Input: joint distribution $p(X, Z|\theta)$

Returns: maximum likelihood estimations for θ

1. Initialize θ^{old}
2. Loop until convergence*:
 - 2.1. Evaluate $p(Z|X, \theta^{\text{old}})$. 
 - 2.2. Update θ^{new} as 
$$\theta^{\text{new}} = \underset{\theta}{\operatorname{argmax}} \{ \sum_Z p(Z|X, \theta^{\text{old}}) \log p(X, Z|\theta) \}$$
 - 2.3. $\theta^{\text{old}} \leftarrow \theta^{\text{new}}$

* for convergence check either the log likelihood or model parameters for changes

Gaussian Mixtures Revisited

- ▶ Incomplete-data log likelihood

$$\log p(X|\theta) = \log[\sum_Z p(X, Z|\theta)]$$

$$\log p(X|\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_n \log[\sum_k \tau_k \mathcal{N}(x_n|\mu_k, \Sigma_k)]$$

← general EM

← EM for GMM

- ▶ Complete-data log likelihood

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

$$\log p(X, Z|\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$= \log \prod_n \prod_k \tau_k^{z_{n,k}} \cdot \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{n,k}}$$

$$= \sum_n \sum_k z_{n,k} \cdot (\log \tau_k + \log \mathcal{N}(x_n|\mu_k, \Sigma_k))$$

- ▶ Posterior

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

$$p(Z|X, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \prod_n \prod_k \tau_k^{z_{n,k}} \cdot \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{n,k}}$$

- ▶ Expectation of complete-data log likelihood under posterior

$$\mathbb{E}_{Z \sim p(Z|X, \theta)} [\log p(X, Z|\theta)]$$

$$\mathbb{E}_{Z \sim p(Z|X, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma})} [\log p(X, Z|\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma})]$$

$$= \sum_n \sum_k p(z_{n,k}|x_n, \mu_k, \Sigma_k) \cdot (\log \tau_k + \log \mathcal{N}(x_n|\mu_k, \Sigma_k))$$

$$= \sum_n \sum_k \frac{\tau_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_j \tau_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} \cdot (\log \tau_k + \log \mathcal{N}(x_n|\mu_k, \Sigma_k))$$



EM increases Likelihood

Why does every iteration of EM cause an increase in the (log) likelihood?

- ▶ **Goal:** find maximum likelihood solutions for models having latent variables
- ▶ **Notation:**
 - X : observed variables $\{X, Z\}$: complete data set
 - Z : latent variables $\{X\}$: incomplete data set
 - θ : model parameters
- ▶ $q(Z)$: any distribution over Z

$$\log p(X|\theta) = \underbrace{\sum_Z q(Z) \log \left[\frac{p(X, Z|\theta)}{q(Z)} \right]}_{=: \mathcal{L}(q, \theta)} + \underbrace{\sum_Z q(Z) \log \left[\frac{q(Z)}{p(Z|X, \theta)} \right]}_{= KL(q(z) \parallel p(Z|X, \theta))}$$

(for decomposition see slide 32)

EM increases Likelihood

$$\log p(X|\theta) = \mathcal{L}(q, \theta) + KL(q(z) \parallel p(Z|X, \theta))$$

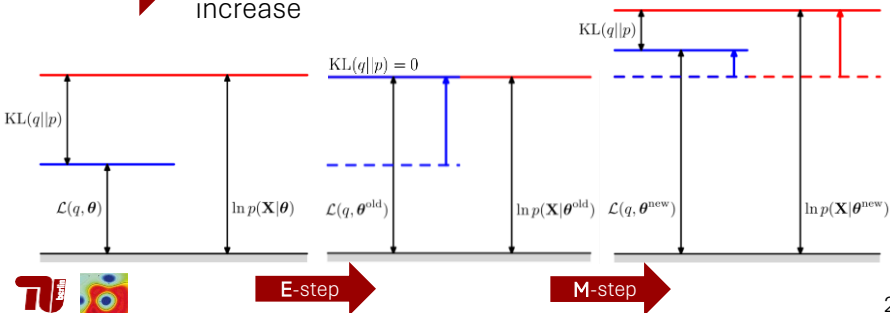
► Because the $KL \geq 0$, $\mathcal{L}(q, \theta)$ is a lower bound on $\log p(X|\theta)$

E-step

Fixes θ , and maximizes the lower bound $\mathcal{L}(q, \theta)$ wrt. $q(Z)$ which causes $KL(q(z) \parallel p(Z|X, \theta)) = 0$

M-step

Fixes $q(Z)$, and maximizes the lower bound $\mathcal{L}(q, \theta)$ wrt. θ which causes $\mathcal{L}(q, \theta)$, correspondingly $\log p(X|\theta)$, to increase



Summary

- ▶ Clustering is an unsupervised learning problem with the goal to partition the data set into similar groups (clusters).
- ▶ K -means is a simple algorithm which provides a clustering solution, and is guaranteed to converge to a (local) optimum
- ▶ Gaussian Mixture Models are a powerful tool to model complex data distributions. They provide a soft-clustering as a byproduct. Their solution can be found by EM.
- ▶ Expectation Maximization finds maximum likelihood solutions for models with latent variables and has a wide application range (not just GMMs).

Additional Slides

Decomposition of $\log p(X|\theta)$

$$\log p(X|\theta) = \underbrace{\sum_Z q(Z) \log \left[\frac{p(X, Z|\theta)}{q(Z)} \right]}_{=: \mathcal{L}(q, \theta)} + \underbrace{\sum_Z q(Z) \log \left[\frac{q(Z)}{p(Z|X, \theta)} \right]}_{= KL(q(z) \parallel p(Z|X, \theta))}$$

$$\begin{aligned} & \sum_Z q(Z) \log \left[\frac{p(X, Z|\theta)}{q(Z)} \right] + \sum_Z q(Z) \log \left[\frac{q(Z)}{p(Z|X, \theta)} \right] \\ &= \sum_Z q(Z) \log p(X, Z|\theta) - \cancel{q(Z) \log q(Z)} + \cancel{q(Z) \log q(Z)} - q(Z) \log p(Z|X, \theta) \\ &= \sum_Z q(Z) \cdot (\log p(X, Z|\theta) - \log p(Z|X, \theta)) \\ &\stackrel{*}{=} \sum_Z q(Z) \log p(X|\theta) = \log p(X|\theta) \cdot \sum_Z q(Z) = \log p(X|\theta) \end{aligned}$$

*(product rule with logarithm: $\log p(X|\theta) = \log p(X, Z|\theta) - \log p(Z|X, \theta)$)

EM increases Likelihood

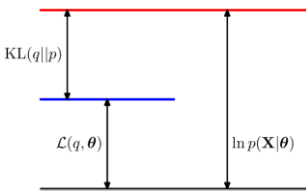
► $\log p(X|\theta)$ can be decomposed into two non-negative terms (for derivation see slide 31):

$$\log p(X|\theta) = \mathcal{L}(q, \theta) + KL(q(z) \parallel p(Z|X, \theta))$$

► $KL(q(z) \parallel p(Z|X, \theta)) \geq 0$ measures the divergence between $q(z)$ and $p(Z|X, \theta)$ with

$$KL(q(z) \parallel p(Z|X, \theta)) = 0 \Leftrightarrow q(z) = p(Z|X, \theta).$$

► Because $\log p(X|\theta) = \mathcal{L}(q, \theta) + KL(q \parallel p)$ and $KL(q \parallel p) \geq 0$, it follows that $\mathcal{L}(q, \theta) \leq \log p(X|\theta)$. $\mathcal{L}(q, \theta)$ is a lower bound on $\log p(X|\theta)$.

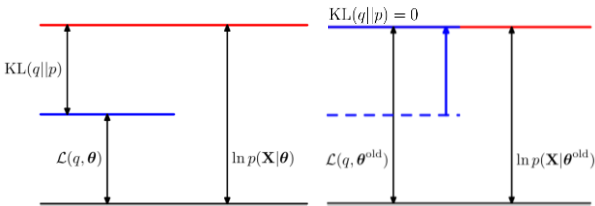


EM increases Likelihood

E-step

Fixes θ , and maximizes the lower bound $\mathcal{L}(q, \theta)$ wrt. $q(Z)$ which causes $KL(q(Z) \parallel p(Z|X, \theta)) = 0$

- ▶ Maximization of $\mathcal{L}(q, \theta)$ wrt. $q(Z)$ causes $q(Z)$, and correspondingly $\mathcal{L}(q, \theta) = \sum_Z q(Z) \log \left[\frac{p(X, Z|\theta)}{q(Z)} \right]$ to change. A change in $q(Z)$ does not affect $\log p(X|\theta)$ because it does not depend on $q(Z)$.
- ▶ Because $\mathcal{L}(q, \theta) \leq \log p(X|\theta)$, its maximal value is obtained when $\mathcal{L}(q, \theta) = \log p(X|\theta)$. Because $\log p(X|\theta) = \mathcal{L}(q, \theta) + KL(q \parallel p)$ this implies $KL(q \parallel p) = 0$ at the maximum.
- ▶ In other words, the E-step causes $q(Z) = p(Z|X, \theta)$.



E-step



EM increases Likelihood

M-step

Fixes $q(Z)$, and maximizes $\mathcal{L}(q, \theta)$ wrt. θ which causes $\mathcal{L}(q, \theta)$, correspondingly $\log p(X|\theta)$, to increase

- ▶ Maximization of $\mathcal{L}(q, \theta)$ wrt. θ gives new model parameters θ^{new} .
- ▶ Because the posterior $p(Z|X, \theta)$ also depends on θ , it will change too so that $q(Z) \neq p(Z|X, \theta^{\text{new}})$ and $KL(q \parallel p) > 0$.
- ▶ Because $\log p(X|\theta)$ depends on θ , it will change as well. Because $\log p(X|\theta^{\text{new}}) = \mathcal{L}(q, \theta^{\text{new}}) + KL(q(z) \parallel p(Z|X, \theta^{\text{new}}))$ with $\mathcal{L}(q, \theta^{\text{new}}) \geq \mathcal{L}(q, \theta^{\text{old}})$ and $KL(q(z) \parallel p(Z|X, \theta^{\text{new}})) > 0$, it follows that $\log p(X|\theta^{\text{new}}) \geq \log p(X|\theta^{\text{old}})$.
- ▶ In other words, EM causes the likelihood to increase in every iteration.

