Everything observed	kMeans Clustering 000000	Mixtures of Gaussians	The General EM Algorithm

Lecture 6 Expectation Maximisation

University of Amsterdam



Introduction	Everything observed	kMeans Clustering	Mixtures of Gaussians	The General EM Algorithm
	0000	000000	0000000	0000000000

- 2 Everything observed
- 8 kMeans Clustering
 - Problem description
 - An algorithm

4 Mixtures of Gaussians

- Problem description
- Optimising the likelihood with latent variables

5 The General EM Algorithm

- Definitions
- MAP learning with EM
- A closer look at the EM algorithm
- Extensions to EM

introduction	0000	000000	00000000	00000000000
1	Introduction			
2	Everything observe	ed		
3	kMeans ClusteringProblem descripAn algorithm	tion		
4	Mixtures of Gauss • Problem descrip • Optimising the	ians tion likelihood with	latent variables	
5	 The General EM A Definitions MAP learning w A closer look at Extensions to E 	Algorithm vith EM the EM algorit M	hm	

・ロト (個) (注) (注) (注) (こ)

kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm

Introduction

Introduction

Last week, we have seen how the factorisation of probability distributions could be represented as graphs.

- We have seen algorithms to efficiently compute marginal probabilities on such graphs.
- We have seen how this could be used to compute conditional probabilities.
- However in we needed to assume that the parameters of the distributions were known.

Today, we see how to learn these parameters efficiently.

UNIVERSITY OF AMSTERDAM

We want to find the parameters that maximise the likelihood

- MAP treatment is very similar
- (exact) full Bayesian treatment is often not tractable
- We'll see approximations in lecture 13

How can we find the maximum of the likelihood?

- If everything is observed, it's easy
- If we have latent variables, it's hard







・ロト・「聞・ 《聞・ 《聞・ 《日・

Introduction	0000	Kivieans Clustering	Mixtures of Gaussians	The General EIVI Algorithm
1	Introduction			
2	Everything observe	ed		
3	kMeans ClusteringProblem descripAn algorithm	; tion		
4	Mixtures of GaussProblem descripOptimising the	ians tion likelihood with	latent variables	
5	 The General EM A Definitions MAP learning w A closer look at Extensions to E 	Algorithm vith EM the EM algorit M	thm	
			< □ > < □ >	★ E + ★ E + ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

Means Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

Ř

INIVERSITY OF AMSTERDAM

Fully observed model

Example



Int	rod	luct	ion
	100		

Means Clustering

Mixtures of Gaussians

The General EM Algorithm 00000000000

Fully observed model



 $p(\mathbf{x}_i, z_i) = \begin{cases} p(\mathcal{C}_1) \, p(\mathbf{x}_i | \mathcal{C}_1) & \text{if } z_i = 1\\ p(\mathcal{C}_2) \, p(\mathbf{x}_i | \mathcal{C}_2) & \text{if } z_i = 0 \end{cases}$

Parametrisation:

$$p(C_1) = \pi$$

$$p(C_2) = 1 - \pi$$

$$p(\mathbf{x}_i | C_1) = \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$

$$p(\mathbf{x}_i | C_2) = \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

$$\Longrightarrow \boldsymbol{\theta} = \{\pi, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\}$$

(Complete) likelihood:

 $p(\mathbf{x}_i, \mathbf{z}_i | \boldsymbol{\theta}) = \left[p(\mathcal{C}_1) \, p(\mathbf{x}_i | \mathcal{C}_1) \right]^{z_i} \left[p(\mathcal{C}_2) \, p(\mathbf{x}_i | \mathcal{C}_2) \right]^{1-z_i}$



I so de se of el s s de de s de se	
111111111111111111111111111111111111111	۱.
minouuction	

<Means Clustering

Mixtures of Gaussians

The General EM Algorithm 00000000000

UNIVERSITY OF AMSTERDAM

Fully observed model



$$p(\mathbf{x}_i, z_i) = \begin{cases} p(\mathcal{C}_1) \, p(\mathbf{x}_i | \mathcal{C}_1) & \text{if } z_i = 1\\ p(\mathcal{C}_2) \, p(\mathbf{x}_i | \mathcal{C}_2) & \text{if } z_i = 0 \end{cases}$$

.

Parametrisation:

$$p(C_1) = \pi$$

$$p(C_2) = 1 - \pi$$

$$p(\mathbf{x}_i | C_1) = \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$

$$p(\mathbf{x}_i | C_2) = \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

$$\Longrightarrow \boldsymbol{\theta} = \{\pi, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\}$$

(Complete) likelihood:

$$p(\mathbf{x}_i, \mathbf{z}_i | \boldsymbol{\theta}) = \left[p(\mathcal{C}_1) \, p(\mathbf{x}_i | \mathcal{C}_1) \right]^{z_i} \left[p(\mathcal{C}_2) \, p(\mathbf{x}_i | \mathcal{C}_2) \right]^{1-z_i}$$

Intelligent Autonomous Systems

kMeans Clustering 000000 Mixtures of Gaussians

The General EM Algorithm 00000000000

Parameter optimisation

Complete likelihood:

$$p(\{\mathbf{x}_i, \mathbf{z}_i\}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \left[\pi \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \right]^{z_i} \left[(1-\pi) \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \right]^{1-z_i}$$

Maximise: take logarithm, set derivative = 0

$$\pi_{1} = \frac{\sum_{i=1}^{N} z_{i}}{N} \qquad \pi_{2} = \frac{\sum_{i=1}^{N} (1 - z_{i})}{N}$$
$$\mu_{1} = \frac{\sum_{i=1}^{N} z_{i} \mathbf{x}_{i}}{\sum_{i=1}^{N} z_{i}} \qquad \mu_{2} = \frac{\sum_{i=1}^{N} (1 - z_{i}) \mathbf{x}_{i}}{\sum_{i=1}^{N} z_{i}}$$
$$\mathbf{\Sigma}_{1} = \frac{\sum_{i=1}^{N} z_{i} \mathbf{x}_{i} \mathbf{x}_{i}^{\top}}{\sum_{i=1}^{N} z_{i}} - \mu_{1} \mu_{1}^{\top} \qquad \mathbf{\Sigma}_{2} = \frac{\sum_{i=1}^{N} (1 - z_{i}) \mathbf{x}_{i} \mathbf{x}_{i}^{\top}}{\sum_{i=1}^{N} z_{i}} - \mu_{2} \mu_{2}^{\top}$$

IAS Intelligent Autonomous Systems UNIVERSITY OF AMSTERDAM

Everything observed	kMeans Clu
0000	

Mixtures of Gaussians

The General EM Algorithm

Hidden class label

So, what happens when the class label is not observed?





Everything observed 000●	kMeans Clustering	Mixtures of Gaussians	The General EM Algorithm

So, what happens when the class label is not observed?

• List all possible assignments, pick best



Ř



Everything observed	kMeans Clustering 000000	Mixtures of Gaussians	The General EM Algorithm

So, what happens when the class label is not observed?

• List all possible assignments, pick best



Ä



Everything observed	kMeans Clustering 000000	Mixtures of Gaussians	The General EM Algorithm

So, what happens when the class label is not observed?





Everything observed	kMeans Clustering 000000	Mixtures of Gaussians	The General EM Algorithm

So, what happens when the class label is not observed?





Everything observed	kMeans Clustering 000000	Mixtures of Gaussians	The General EM Algorithm

Ä

INIVERSITY OF AMSTERDAM

Hidden class label

So, what happens when the class label is not observed?



Everything observed	kMeans Clustering 000000	Mixtures of Gaussians	The General EM Algorithm

So, what happens when the class label is not observed?

• List all possible assignments, pick best





< ∃→

э

Everything observed	kMeans Clustering 000000	Mixtures of Gaussians	The General EM Algorithm

JNIVERSITY OF AMSTERDAM

イロト イポト イヨト イヨト

э

Hidden class label

So, what happens when the class label is not observed?



Everything observed	kMeans Clustering 000000	Mixtures of Gaussians	The General EM Algorithm

So, what happens when the class label is not observed?

• List all possible assignments, pick best



(日)、

ㅋ ㅋ



Introduction	Everything observed	kMeans Clustering	Mixtures of Gaussians	The General EM Algorithm

- 2 Everything observed
- kMeans Clustering
 Problem description
 - An algorithm

4 Mixtures of Gaussians

- Problem description
- Optimising the likelihood with latent variables

5 The General EM Algorithm

- Definitions
- MAP learning with EM
- A closer look at the EM algorithm
- Extensions to EM



kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 00000000000

Ř

UNIVERSITY OF

Amsterdam

Problem description

Example: kMeans clustering





Intelligent Autonomous Systems

 Introduction
 Everything observed 000
 kMeans Clustering 00000000
 Mixtures of Gaussians 00000000
 The General EM Algo 000000000

 Problem description

Properties of clustering

In order to cluster the data, we need:

- Some representation of what a cluster looks like
 - Let's assume for now that each cluster is fully defined by its centre.

• An assignment of each datapoint to one of the clusters

• Let's assume that this is defined by the Euclidean distance to the clusters' centres.

(日)、

The best configuration is the one where all datapoints are as close as possible to their cluster's centre





uction Everything observed kMeans Clustering 0000 0€0000 Mixtures of Gaussians

The General EM Algorithm 00000000000

・ロト ・ 理 ト ・ ヨ ト ・ ヨ ト

э

Problem description

Properties of clustering

In order to cluster the data, we need:

- Some representation of what a cluster looks like
 - Let's assume for now that each cluster is fully defined by its centre.
- An assignment of each datapoint to one of the clusters
 - Let's assume that this is defined by the Euclidean distance to the clusters' centres.

The best configuration is the one where all datapoints are as close as possible to their cluster's centre



on Everything observed ki 0000 O

kMeans Clustering ○●○○○○ Mixtures of Gaussians

The General EM Algorithm

Problem description

Properties of clustering

In order to cluster the data, we need:

- Some representation of what a cluster looks like
 - Let's assume for now that each cluster is fully defined by its centre.
- An assignment of each datapoint to one of the clusters
 - Let's assume that this is defined by the Euclidean distance to the clusters' centres.

The best configuration is the one where all datapoints are as close as possible to their cluster's centre



Mixtures of Gaussians

The General EM Algorithm

An algorithm

Optimising the clustering

An exhaustive search for the optimal clustering is intractable and requires C^N operations

- Where C is the number of clusters
- and N is the number of datapoints.

How do we find the optimal clustering without exhaustive search? Solve the clustering iteratively:

- Initialise the cluster means at random
- Repeat until convergence
 - Assign each data point to the closest cluster mean
 - Opdate each cluster's centre according to the associated data



Mixtures of Gaussians

The General EM Algorithm

Optimising the clustering

An exhaustive search for the optimal clustering is intractable and requires ${\cal C}^N$ operations

- Where C is the number of clusters
- and N is the number of datapoints.

How do we find the optimal clustering without exhaustive search? Solve the clustering iteratively:

- Initialise the cluster means at random
- Repeat until convergence
 - Assign each data point to the closest cluster mean
 - Opdate each cluster's centre according to the associated data

JNIVERSITY OF AMSTERDAM

Everything observed

kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

An algorithm

An iterative algorithm



××



Everything observed

kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

Ř

An algorithm





Everything observed

kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

An algorithm

An iterative algorithm



Ř

IAS Intelligent Autonomous Systems

Everything observed

kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

An algorithm

An iterative algorithm



IAS Intelligent Autonomous Systems

Everything observed

kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

Ř

An algorithm





Everything observed

kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

An algorithm

An iterative algorithm



Ř

Everything observed 0000 kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

Ř

An algorithm





Everything observed 0000 kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

An algorithm





Everything observed 0000 kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 00000000000

An algorithm





Everything observed 0000 kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

An algorithm




Introduction

Everything observed 0000 kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 00000000000

An algorithm

An iterative algorithm





kMeans Clustering 000000

・ロト ・ 雪 ト ・ ヨ ト

э

An algorithm

Problems with kMeans

There are some disadvantages to kMeans:

Euclidean distance:

- Only useful for some types of data
- Not robust to outliers.
- Sensitive to scaling of data
- Solution: Other distance measures
- Hard assignments At each iteration, each datapoint is assigned to exactly one cluster, even for doubtful cases.



Ř

Intr	odi	ictic	hn.
	out		

kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

An algorithm





kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

ž

An algorithm





kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

Ř

An algorithm







kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

An algorithm

k-means sensitivity to scaling



Ř



kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

Ř

An algorithm







kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

Ř

An algorithm





kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

An algorithm

Intelligent Autonomous Systems

k-means sensitivity to scaling



Ř



kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

An algorithm

k-means sensitivity to scaling





kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

Ř

An algorithm







kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

An algorithm

k-means sensitivity to scaling



Ř



kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

An algorithm

k-means sensitivity to scaling



Ř

Everything observed	kMeans Clustering 000000	Mixtures of Gaussians	The General EM Algorithm

Introduction

- 2 Everything observed
- kMeans Clustering
 Problem description
 - An algorithm

Mixtures of Gaussians

- Problem description
- Optimising the likelihood with latent variables

5 The General EM Algorithm

- Definitions
- MAP learning with EM
- A closer look at the EM algorithm
- Extensions to EM

 Introduction
 Everything observed 0000
 kMeans Clustering 00000
 Mixtures of Gaussians 00000000
 The General EM Algorit 000000000

 Problem description
 Free Clustering
 Free Cl

Mixtures of Gaussians

A mixture of Gaussians is a linear combination of Gaussians:

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

where $0 \leqslant \pi_k \leqslant 1$ and

$$\sum_{k=1}^{K} \pi_k = 1 \tag{2}$$





ž

 Introduction
 Everything observed 0000
 kMeans Clustering 00000
 Mixtures of Gaussians 0000000
 The General EM Algor 0000000000

 Problem description
 Problem description
 Problem description
 Problem description

Mixtures of Gaussians

A mixture of Gaussians is a linear combination of Gaussians:

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

where $0 \leqslant \pi_k \leqslant 1$ and

$$\sum_{k=1}^{K} \pi_k = 1 \tag{2}$$



▲□▶ ▲圖▶ ▲臣▶ ▲臣▶ ―臣 … のへで

ž



 Introduction
 Everything observed 0000
 kMeans Clustering 00000
 Mixtures of Gaussians ●0000000
 The General EM Algorith 000000000

 Problem description

Mixtures of Gaussians

A mixture of Gaussians is a linear combination of Gaussians:

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

where $0 \leq \pi_k \leq 1$ and

$$\sum_{k=1}^{K} \pi_k = 1 \tag{2}$$





ž

	Everything observed	kMeans Clustering 000000	Mixtures of Gaussians 0●000000	The General EM Algorithm
Problem descript	ion			

Alternative view of the mixture model

We introduce a binary random variable z in 1-of-K encoding, the probability of z can be writen as

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

We choose the conditional distribution of \mathbf{x} given a z_k as

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

Then we have:

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

K

Intelligent Autonomous Systems

3

・ロト ・ 理 ト ・ ヨ ト ・ ヨ ト

Mixtures of Gaussians 0000000

UNIVERSITY OF AMSTERDAM

Optimising the likelihood with latent variables

Optimising the likelihood

Similar to kMeans, finding the optimal parameters for $p(\mathbf{x})$ in a mixture of Gaussians is hard.

However, with our new representation, we can now work with $p(\mathbf{x}, \mathbf{z})$ rather than $p(\mathbf{x})$. In particular, consider the log-likelihood of N datapoints \mathbf{X} :

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]$$
(3)

Setting the first derivative with respect to μ_k equal to zero, gives:

$$0 = -\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{\ell} \pi_\ell \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)} \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)$$
(4)

Mixtures of Gaussians 0000000

Optimising the likelihood with latent variables

Optimising the likelihood

Similar to kMeans, finding the optimal parameters for $p(\mathbf{x})$ in a mixture of Gaussians is hard.

However, with our new representation, we can now work with $p(\mathbf{x}, \mathbf{z})$ rather than $p(\mathbf{x})$. In particular, consider the log-likelihood of N datapoints \mathbf{X} :

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]$$
(3)

Setting the first derivative with respect to μ_k equal to zero, gives:

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{\ell} \pi_{\ell} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_{\ell}, \boldsymbol{\Sigma}_{\ell})}}_{p(z_k | \mathbf{x}_n)} \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)$$
(4)
Autonomous Systems



The values for μ_k , Σ_k and π_k that maximise the likelihood are:

$$\pi_k = \frac{N_k}{N}$$
$$\mu_k = \frac{1}{N_k} \sum_{n=1}^N p(z_k | \mathbf{x}_n) \mathbf{x}_n$$
$$\mathbf{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^N p(z_k | \mathbf{x}_n) (\mathbf{x}_n - \mu_k) (\mathbf{x}_n - \mu_k)^\top$$

where we defined $N_k = \sum_{n=1}^{N} p(z_k | \mathbf{x}_n)$

However $p(z_k|\mathbf{x}_n)$ is a function of $\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K$

Intelligent Autonomous Systems



The values for μ_k , Σ_k and π_k that maximise the likelihood are:

$$\pi_k = \frac{N_k}{N}$$
$$\mu_k = \frac{1}{N_k} \sum_{n=1}^N p(z_k | \mathbf{x}_n) \mathbf{x}_n$$
$$\mathbf{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^N p(z_k | \mathbf{x}_n) (\mathbf{x}_n - \mu_k) (\mathbf{x}_n - \mu_k)^\top$$

where we defined $N_k = \sum_{n=1}^N p(z_k | \mathbf{x}_n)$

However $p(z_k|\mathbf{x}_n)$ is a function of $\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K$

Intelligent Autonomous Systems

Introduction

Everything observed

Means Clustering

Mixtures of Gaussians

The General EM Algorithm 00000000000

UNIVERSITY OF AMSTERDAM

Optimising the likelihood with latent variables

The EM algorithm for Gaussian Mixtures

The Expectation-Maximisation algorithm is an iterative update where:

- E-step Compute the posterior probabilities of the latent variables given the data and the current parameters (also called *responsibilities*), $p(z_k|\mathbf{x}_n, \boldsymbol{\theta})$
- M-step Optimise the expectation of the complete log-likelihood with respect to the parameters

In practice, stop when the increase in likelihood falls below a certain threshold.



	Everything observed	kMeans Clustering 000000	Mixtures of Gaussians	The General EM Algorithm
Optimising the like	lihood with latent variables			



Example: Mixtures of Gaussians



Introduction

Everything observe

kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 00000000000

JNIVERSITY OF AMSTERDAM

Optimising the likelihood with latent variables

About EM for mixtures of Gaussians

Some things to notice

 The problem is ill-posed: consider a component k with covariance σl. If the mean of the component falls exactly on a data point, its contribution to the likelihood is

$$\ln p(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{2\pi\sigma}}$$
(5)

which, in the limit of $\sigma \rightarrow 0$ goes to infinity.

- A suitable prior on heta avoids this problem
- The kMeans algorithm is equivalent with EM for a Gaussian mixture model, where the covariance is σI for *all* mixture components in the limit $\sigma \rightarrow 0$.

Everything observed	kľ

kMeans Clustering

Mixtures of Gaussians

The General EM Algorithm 0000000000

Optimising the likelihood with latent variables

Example



	Everything observed	kMeans Clustering 000000	Mixtures of Gaussians	The General EM Algorithm
1	Introduction			
2	Everything observe	ed		
3	kMeans ClusteringProblem descripAn algorithm	tion		
4	Mixtures of Gaussi Problem descrip Optimising the l	a <mark>ns</mark> tion ikelihood with	latent variables	
5	The General EM ADefinitionsMAP learning wA closer look at	lgorithm ith EM the EM algorit	thm	

◆□ ▶ < 圖 ▶ < 圖 ▶ < 圖 ▶ < 圖 • 의 Q @</p>

• Extensions to EM

kMeans Clusterin

Mixtures of Gaussians

The General EM Algorithm

Definitions

The General EM Algorithm

In general, the EM algorithm is defined as follows. Optimising the *complete* log-likelihood

$$p(\mathbf{X}, \mathbf{Z}|\theta)$$
 (6)

would be easy, but we only observe **X**. So let's optimise our best estimate of the complete log-likelihood: the expectation of the complete log-likelihood under our current parameter estimates θ^{old} :

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})]$$
(7)

$$=\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta) \ln p(\mathbf{X}, \mathbf{Z}|\theta)$$
(8)





In the E-step we evaluate the distribution of the latent variables

$$p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \tag{9}$$

so that we can compute the expectation of the complete log-likelihood (although we do not need to compute that explicitly)

 \bullet In the M-step we maximise the complete log-likelihood with respect to θ

$$\boldsymbol{\theta}^{new} \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) \tag{10}$$

IAS Intelligent Autonomous Systems



EM for MAP learning

If we incorporate a prior over θ , $p(\theta)$, the EM algorithm changes slightly:

E-Step : $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$ does not depend on $p(\theta^{old})$ since θ^{old} are given, so the E-step remains identical.

M-Step : We now optimise

$$\mathbb{E}\left[\ln(\rho(\mathbf{X}, \mathbf{Z}|\theta)\rho(\theta))\right]$$
(11)

where $p(\theta)$ is constant with respect to $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$, so that we get:

$$\theta^{new} = \underset{\theta}{\arg \max} \mathcal{Q}(\theta, \theta^{old}) + \ln p(\theta)$$
 (12)

IAS Intelligent Autonomous Systems Ř



Consider the log likelihood

$$\ell(\boldsymbol{\theta}) = \ln p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{z}} \ln p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$$
(13)

$$= \sum_{\mathbf{z}} \ln[q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})}]$$
(14)

$$\geq \sum_{\mathbf{z}} q(\mathbf{Z}) \ln[\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}]$$
(15)

(日)、

æ

By Jensen's inequality





	Everything observed	kMeans Clustering 000000	Mixtures of Gaussians	The General EM Algorithm
A closer look at	the EM algorithm			
Jensen'	s inequality			ž



UNIVERSITY OF AMSTERDAM

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 のへぐ

	Everything observed 0000	kMeans Clustering 000000	Mixtures of Gaussians	The General EM Algorithm
A closer look at	the EM algorithm			
Mutual	entropy			Š
Cons	ider the quantity	/		Univ
Σ	$\sum_{\mathbf{z}} q(\mathbf{z}) \ln[\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})}]$	$\left(\frac{\theta}{\theta}\right)$]		ERSITY O
	$=\sum_{z}q(z)$	\mathbf{z}) [In $p(\mathbf{z} \mathbf{x}, \boldsymbol{ heta})$ -	$+ \ln p(\mathbf{x} m{ heta}) - \ln q$	(z)] Amste
	$=\sum q($	\mathbf{z}) ln $\frac{p(\mathbf{z} \mathbf{x}, \boldsymbol{\theta})}{q(\mathbf{z})}$ -	$+ \ln p(\mathbf{x} m{ heta}) \sum q(\mathbf{x} m{ heta})$	z) am

$$\sum_{\mathbf{z}} q(\mathbf{z}) \ln[\frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})}]$$

= $\sum_{\mathbf{z}} q(\mathbf{z}) [\ln p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) + \ln p(\mathbf{x}|\boldsymbol{\theta}) - \ln q(\mathbf{z})]$
= $\sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})}{q(\mathbf{z})} + \ln p(\mathbf{x}|\boldsymbol{\theta}) \underbrace{\sum_{\mathbf{z}} q(\mathbf{z})}_{=1}$

so that

$$\ln p(\mathbf{x}|\theta) = \underbrace{\sum_{\mathbf{z}} q(\mathbf{z}) \ln[\frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z})}]}_{\mathcal{L}(q, \theta)} + \underbrace{\left(-\sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{z}|\mathbf{x}, \theta)}{q(\mathbf{z})}\right)}_{\mathcal{K}L(q||p)}$$

	Everything observed	kMeans Clustering 000000	Mixtures of Gaussians	The General EM Algorithm ○○○○○○●○○○○
A closer look at	the EM algorithm			
E-Step				××××

During the E-step, we maximise $\mathcal{L}(q, \theta)$ with respect to $q(\mathbf{z})$, leaving θ^{old} untouched. Since $\ell(\theta^{old})$ does not depend on $q(\mathbf{z})$, this can only be achieved by setting

JNIVERSITY OF AMSTERDAM


	Everything observed	kMeans Clustering 000000	Mixtures of Gaussians	The General EM Algorithm ○○○○○○●○○○
A closer look at th	ne EM algorithm			
M-Step				×

During the M-step, we maximise $\mathcal{L}(q, \theta)$ with respect to θ , leaving $q(\mathbf{z})$ untouched. Since $q(\mathbf{x}) = p(\mathbf{z}|\mathbf{x}, \theta^{old})$,

$$\begin{split} \mathcal{L}(q, \theta) &= \sum_{\mathbf{z}} p(\mathbf{z} | \mathbf{x}, \theta) \ln p(\mathbf{z}, \mathbf{x} | \theta) - \sum_{\mathbf{z}} p(\mathbf{z} | \mathbf{x}, \theta) \ln p(\mathbf{z} | \mathbf{x}, \theta) \\ &= \mathcal{Q}(\theta, \theta) + \mathsf{H}[q(\mathbf{z})] \end{split}$$

- Maximising L(q, θ) with respect to θ changes ln p(z|x, θ), so that KL(q||p) increases.
- $p(\mathbf{x}|\boldsymbol{\theta})$ therefore increases at least as much as $\mathcal{L}(q, \boldsymbol{\theta})$

Intelligent Autonomous Systems



Introduction

Everything observed

Means Clustering

Mixtures of Gaussians

The General EM Algorithm

A closer look at the EM algorithm

EM vs. Gradient Descent





Common extensions of EM

Sometimes, the following extensions of EM are used:

- When the datapoints are independent, the responsibilities z_n depend on x_n and θ only, so that the E and M step can be computed online rather than in batch. This can converge faster than the batch version.
- Sometimes the E-step or M-step (or both) remain intractable. Increasing the likelihood (rather than maximising it) still guarantees increasing the likelihood. This is called Generalised EM (GEM)





Today, we have seen:

- Learning when we can do inference
- Examples of the EM algorithm
- A formal analysis of EM
- Compared k-means with mixtures of Gaussians

Lab:

• Implement the EM algorithm for mixtures of Gaussians

IAS Intelligent Autonomous Systems (Bishop, p. 439–441) (Bishop, p. 423–439) (Bishop, p. 450–453)

・ ロ ト ・ 雪 ト ・ 目 ト

-

