EM算法原理及其应用

罗维

大纲

- 基础知识
- EM算法应用举例
- EM算法及其证明
- EM算法的变种

EM算法的名字由来

E 步

Expectation

期望

M 步

Maximization

最大化

EM（Expectation Maximization, 期望最大化）算法

笼统的EM算法描述

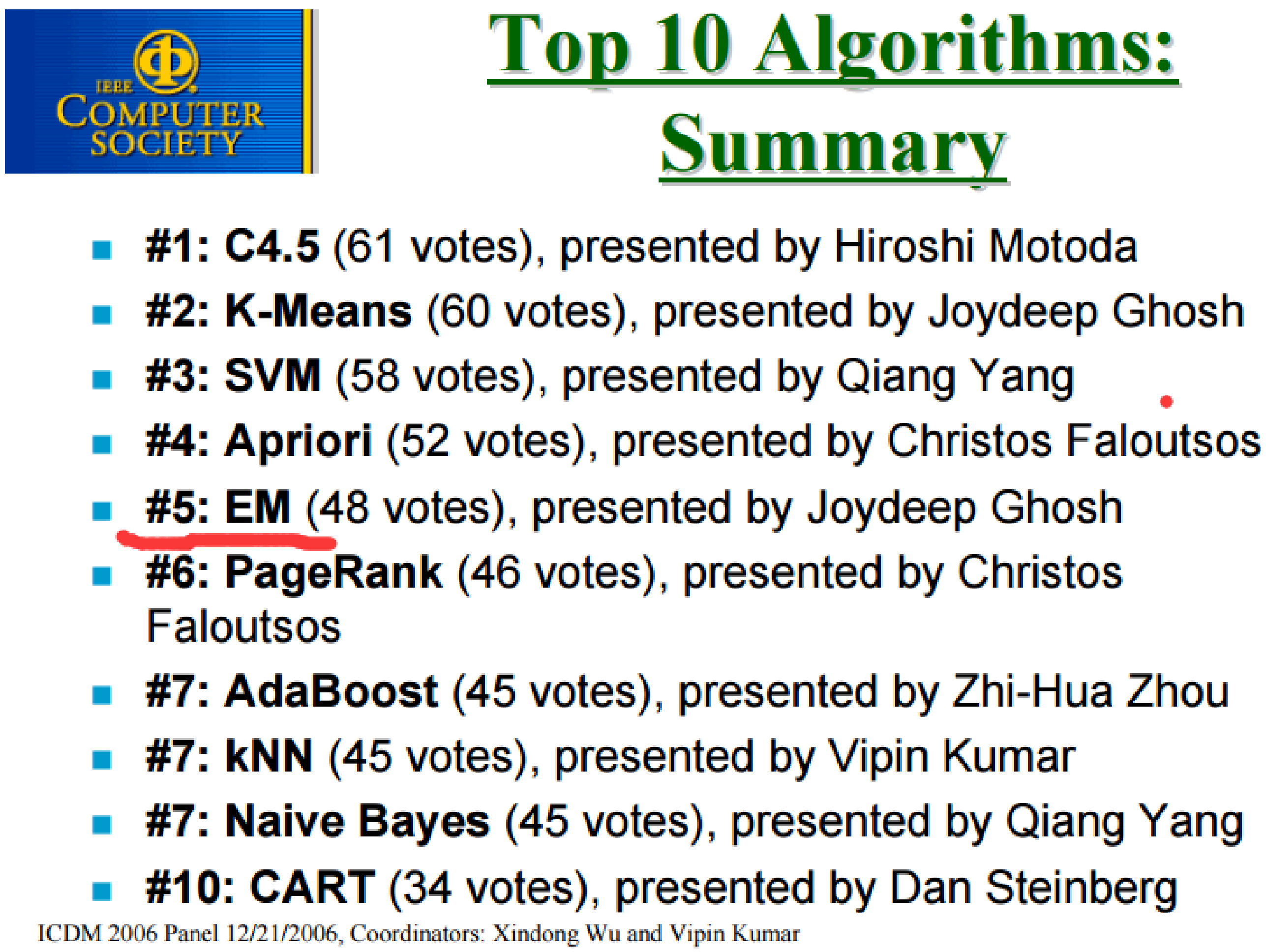
```
Loop {
    E步：求期望（expectation）        → 什么函数关于什么分布的期望？
    M步：求极大（maximization）       → 关于什么函数的最大化？
}
```

- “鸡生蛋，蛋生鸡”

When & What

- 1977年由Dempster等人 总结 提出

- 一种优化算法框架，用于含有 隐变量（hidden variable）的概率模型参数的极大似然估计（Maximum Likelihood Estimation, MLE），或极大后验概率估计（Maximum A Posterior estimation, MAP）



Top 10 Algorithms: Summary

- **#1: C4.5** (61 votes), presented by Hiroshi Motoda
- **#2: K-Means** (60 votes), presented by Joydeep Ghosh
- **#3: SVM** (58 votes), presented by Qiang Yang
- **#4: Apriori** (52 votes), presented by Christos Faloutsos
- **#5: EM** (48 votes), presented by Joydeep Ghosh
- **#6: PageRank** (46 votes), presented by Christos Faloutsos
- **#7: AdaBoost** (45 votes), presented by Zhi-Hua Zhou
- **#7: kNN** (45 votes), presented by Vipin Kumar
- **#7: Naive Bayes** (45 votes), presented by Qiang Yang
- **#10: CART** (34 votes), presented by Dan Steinberg

ICDM 2006 Panel 12/21/2006, Coordinators: Xindong Wu and Vipin Kumar

[1] 机器学习十大算法(ICDM2006) （英文）
http://119.90.25.20/www.cs.uvm.edu/~icdm/algorithms/ICDM06-Panel.pdf

[2] 机器学习十大算法(ICDM2006) （中文翻译）
www.itfront.cn/attachment.aspx?attachmentid=1565

涉及到的基本概念

- 无监督学习
- 生成式模型
- 隐变量
- 先验概率、后验概率、似然概率

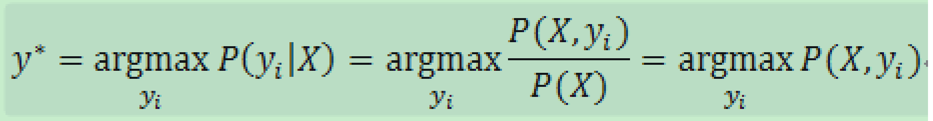
无监督学习

- 无监督学习：样本没有标注
  - 聚类
  - 概率密度估计

- 变量说明：X表示样本（标量 or 向量），y表示标注（标量）

方法	处理怎样的数据	模型举例
无监督学习	(X)	k-Means, HMM, GMM, pLSA, LDA…
有监督学习	(X, y)	Naïve Bayes, NN, LR, ME, SVM, GBDT…
半监督学习	(X) + (X, y)	self-training, co-training, S3VM…
强化学习	(action, state, award)	Markov Decision Process（MDP）

生成式模型

- 生成式模型
  - 带有一个故事（称呼为生成故事）
  - 

$$y^* = \operatorname*{argmax}_{y_i} P(y_i|X) = \operatorname*{argmax}_{y_i} \frac{P(X,y_i)}{P(X)} = \operatorname*{argmax}_{y_i} P(X,y_i)$$

方法	对什么建模	模型举例
生成式模型	P(X, y_i)	Naïve Bayes, HMM, GMM, pLSA, LDA …
判别式模型	P(y_i | X)	NN, LR, ME, SVM, GBDT …

参考文献：http://luowei828.blog.163.com/blog/static/310312042010228247264 71/

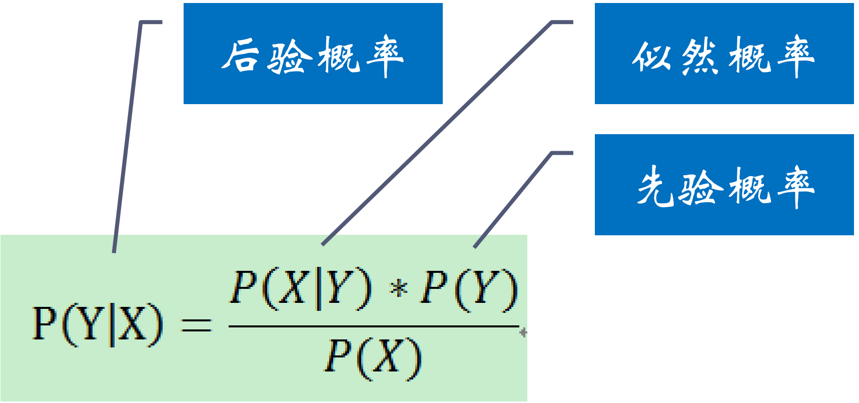
隐变量

- 隐变量（latent variable）
  - 可能是建模时就带有隐变量，也可能是为了求解方便而引入
  - 含有隐变量，通常的MLE估计、MAP估计没法实施

模型	模型的隐变量是什么
k-Means	样本所属的聚类中心点
HMM	隐含状态 (E.g. 词性 for 词性标注；状态 for 其他序列标注模型)
GMM	样本所属的高斯分布 (Gaussian Distribution)
topic model (E.g. pLSA, LDA)	topic
IBM Model for Word Alignment	词语对齐 E.g. (布什与沙龙举行了会谈) (Bush held a meeting with Sharon)

先验概率、后验概率、似然概率

- **贝叶斯公式**
  - X: 标量或者向量
  - Y: 标量或者向量



后验概率

似然概率

先验概率

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)}$$

涉及到的数学基础

- 凸集 (convex sets)
- 凸函数 (convex functions)
- Jensen不等式
- KL距离
- 高斯分布

凸集

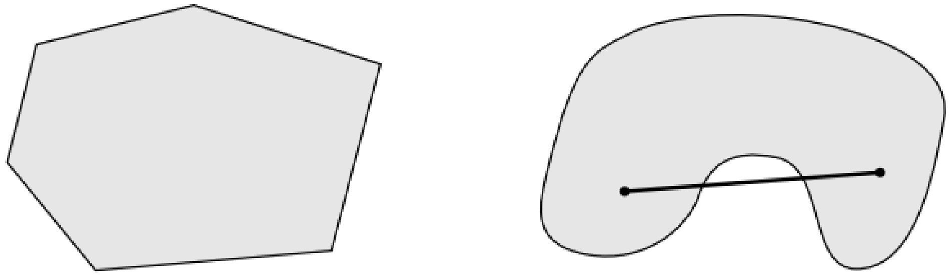
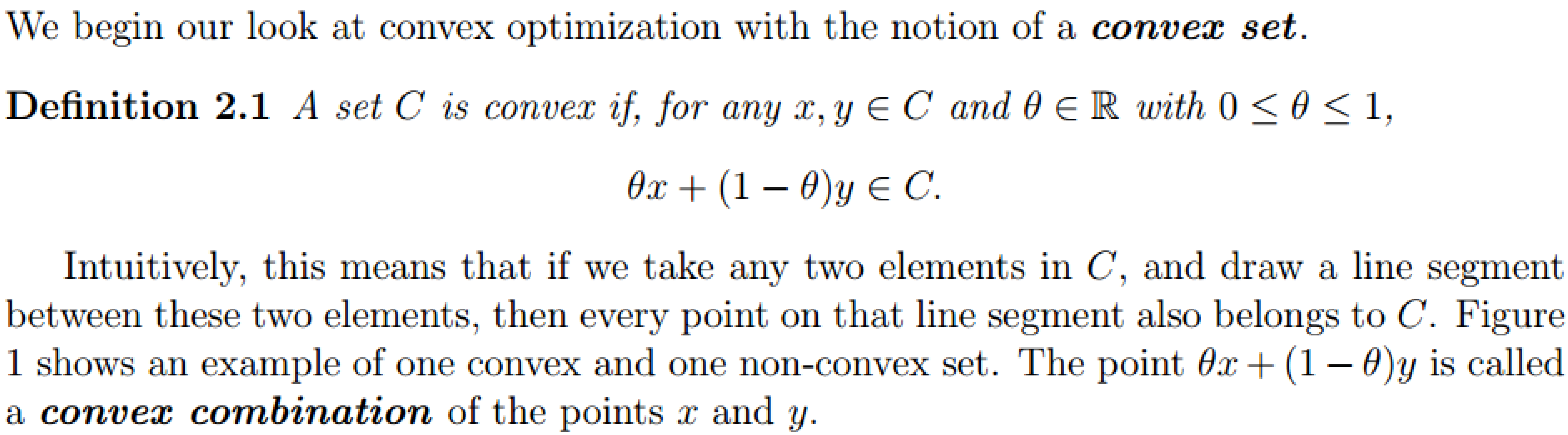


Figure 1: Examples of a convex set (a) and a non-convex set (b).



We begin our look at convex optimization with the notion of a ***convex set***.

Definition 2.1 *A set C is convex if, for any $x, y \in C$ and $\theta \in \mathbb{R}$ with $0 \leq \theta \leq 1$,*

$$\theta x + (1 - \theta)y \in C.$$

Intuitively, this means that if we take any two elements in C, and draw a line segment between these two elements, then every point on that line segment also belongs to C. Figure 1 shows an example of one convex and one non-convex set. The point $\theta x + (1 - \theta)y$ is called a ***convex combination*** of the points x and y.

参考文献：http://cs229.stanford.edu/section/cs229-cvxopt.pdf

凸函数

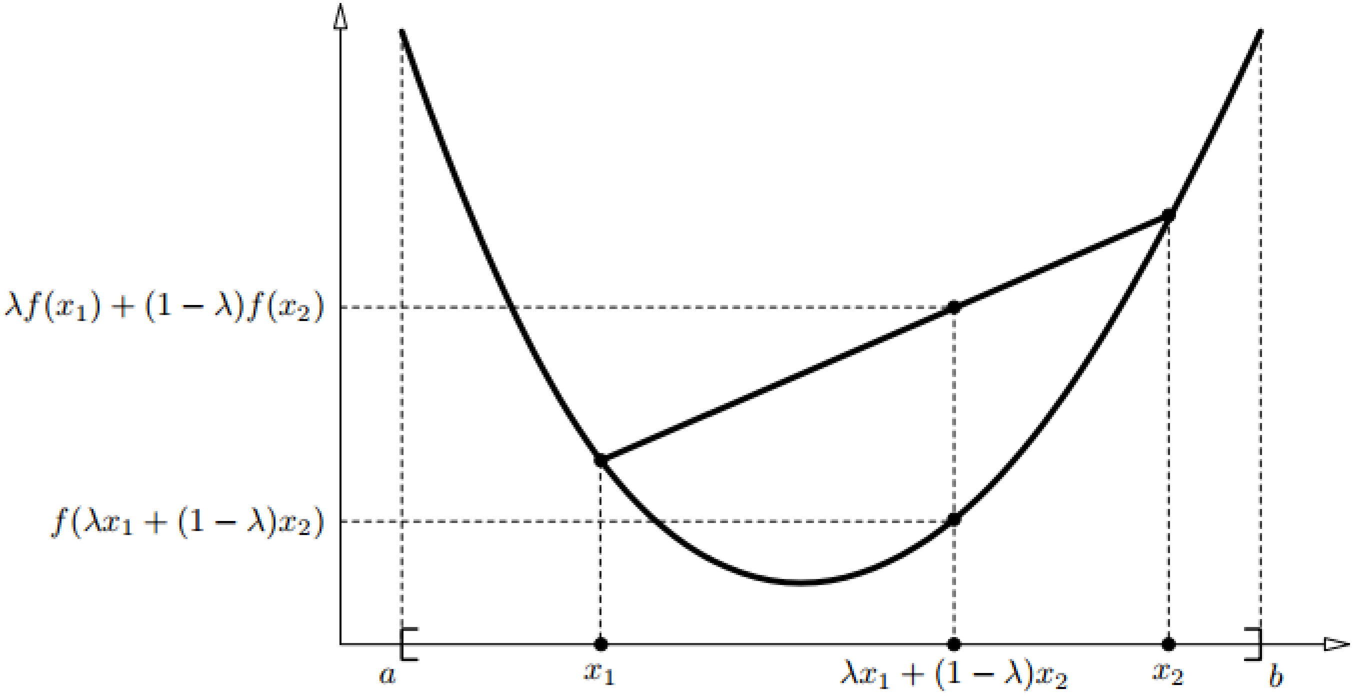
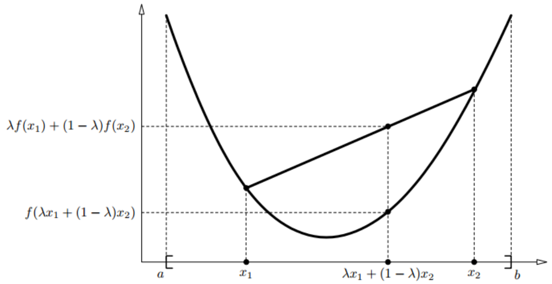

$\lambda f(x_1) + (1-\lambda) f(x_2)$
$f(\lambda x_1 + (1-\lambda)x_2)$
a
x_1
$\lambda x_1 + (1-\lambda)x_2$
x_2
b

Figure 1: f is *convex* on $[a,b]$ if $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$ $\forall x_1, x_2 \in [a,b], \quad \lambda \in [0,1]$.

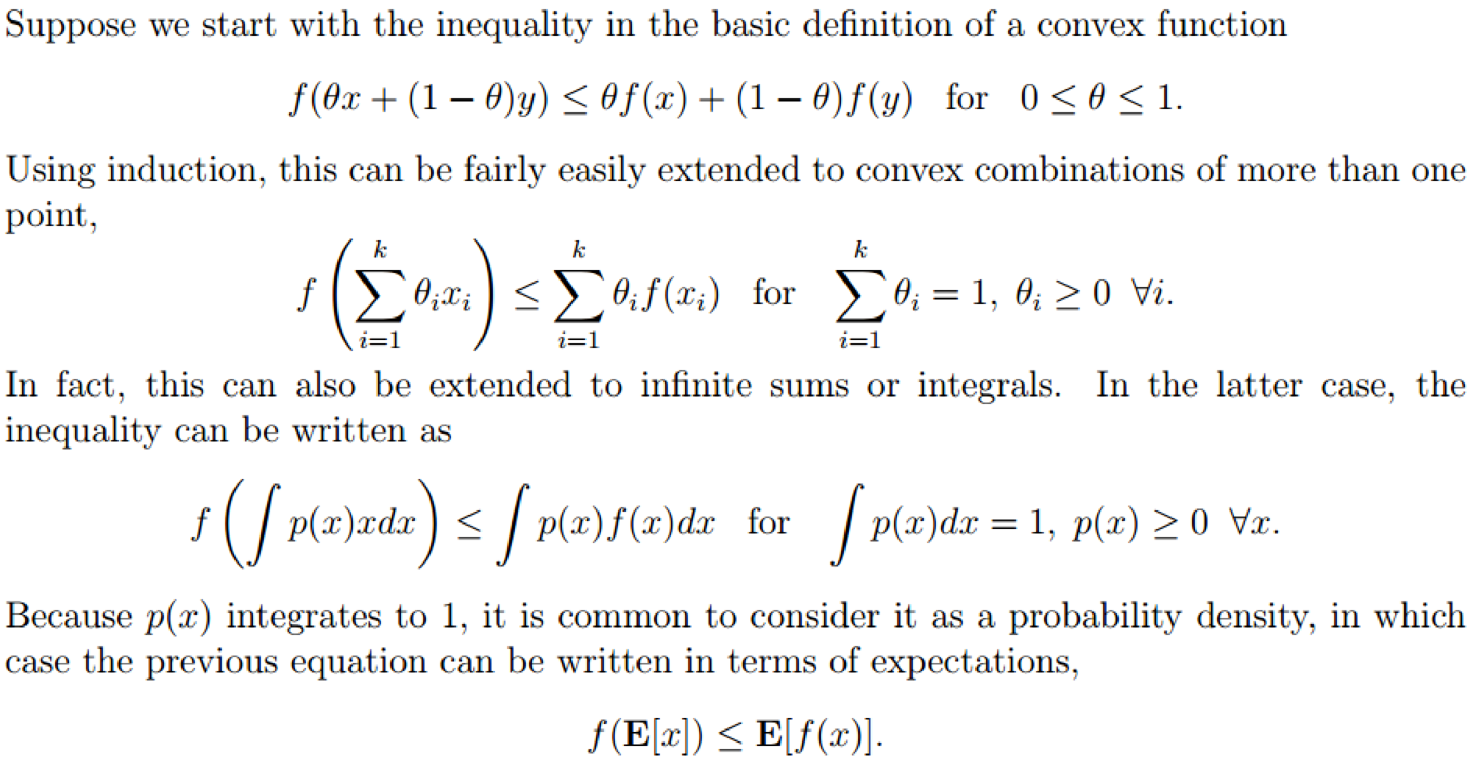
参考文献： https://www.cs.utah.edu/~piyush/teaching/EM_algorithm.pdf

Jensen不等式



$\lambda f(x_1) + (1-\lambda) f(x_2)$
$f(\lambda x_1 + (1-\lambda)x_2)$
a
x_1
$\lambda x_1 + (1-\lambda)x_2$
x_2
b

Figure 1: f is *convex* on $[a,b]$ if $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$ $\forall x_1, x_2 \in [a,b], \;\; \lambda \in [0,1]$.



Suppose we start with the inequality in the basic definition of a convex function

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y) \quad \text{for} \quad 0 \leq \theta \leq 1.$$

Using induction, this can be fairly easily extended to convex combinations of more than one point,

$$f\left(\sum_{i=1}^{k} \theta_i x_i\right) \leq \sum_{i=1}^{k} \theta_i f(x_i) \quad \text{for} \quad \sum_{i=1}^{k} \theta_i = 1, \; \theta_i \geq 0 \;\; \forall i.$$

In fact, this can also be extended to infinite sums or integrals. In the latter case, the inequality can be written as

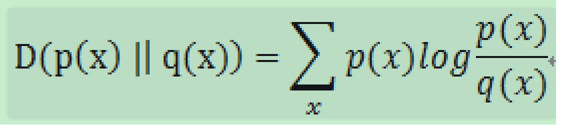
$$f\left(\int p(x) x dx\right) \leq \int p(x) f(x) dx \quad \text{for} \quad \int p(x) dx = 1, \; p(x) \geq 0 \;\; \forall x.$$

Because $p(x)$ integrates to 1, it is common to consider it as a probability density, in which case the previous equation can be written in terms of expectations,

$$f(\mathbf{E}[x]) \leq \mathbf{E}[f(x)].$$

参考文献：http://cs229.stanford.edu/section/cs229-cvxopt.pdf

KL距离

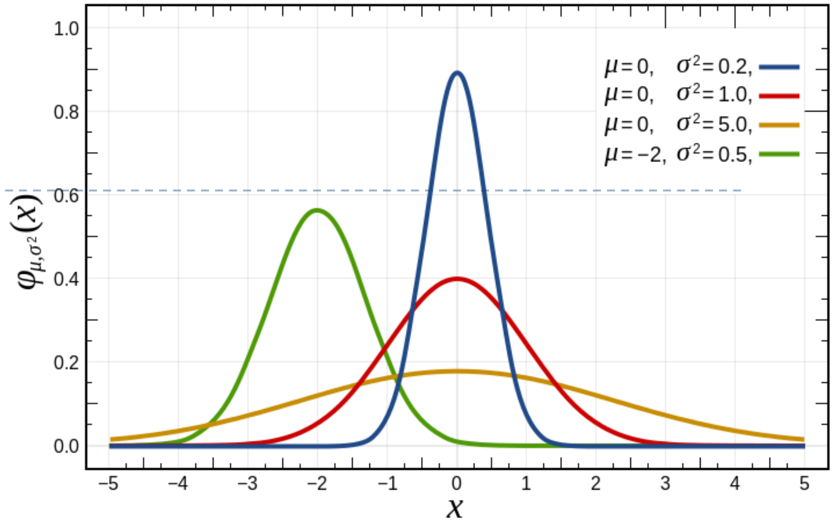
- KL距离
  - 又称 KL散度（Kullback–Leibler divergence）
  - 又称 相对熵（relative entropy）
  - 

    $$D(p(x) \,||\, q(x)) = \sum_{x} p(x) log \frac{p(x)}{q(x)}$$

- 2个主要性质
  - 非对称：D(p(x) || q(x)) 与 D(q(x) || p(x))不一定相等
  - 恒大于等于0：当且仅当p(x)=q(x)时，D(p(x) || q(x)) = 0

参考文献：https://www.cs.princeton.edu/courses/archive/fall11/cos597D/L03.pdf

高斯分布



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{2.42}$$

where μ is the mean and σ^2 is the variance. For a D-dimensional vector $\mathbf{x}$, the multivariate Gaussian distribution takes the form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{2.43}$$

where $\boldsymbol{\mu}$ is a D-dimensional mean vector, $\boldsymbol{\Sigma}$ is a $D \times D$ covariance matrix, and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

参考文献：PRML书的第2.3节

大纲

- 基础知识
- EM算法应用举例
- EM算法及其证明
- EM算法的变种

EM算法应用：以k-Means为例

- 一种聚类算法

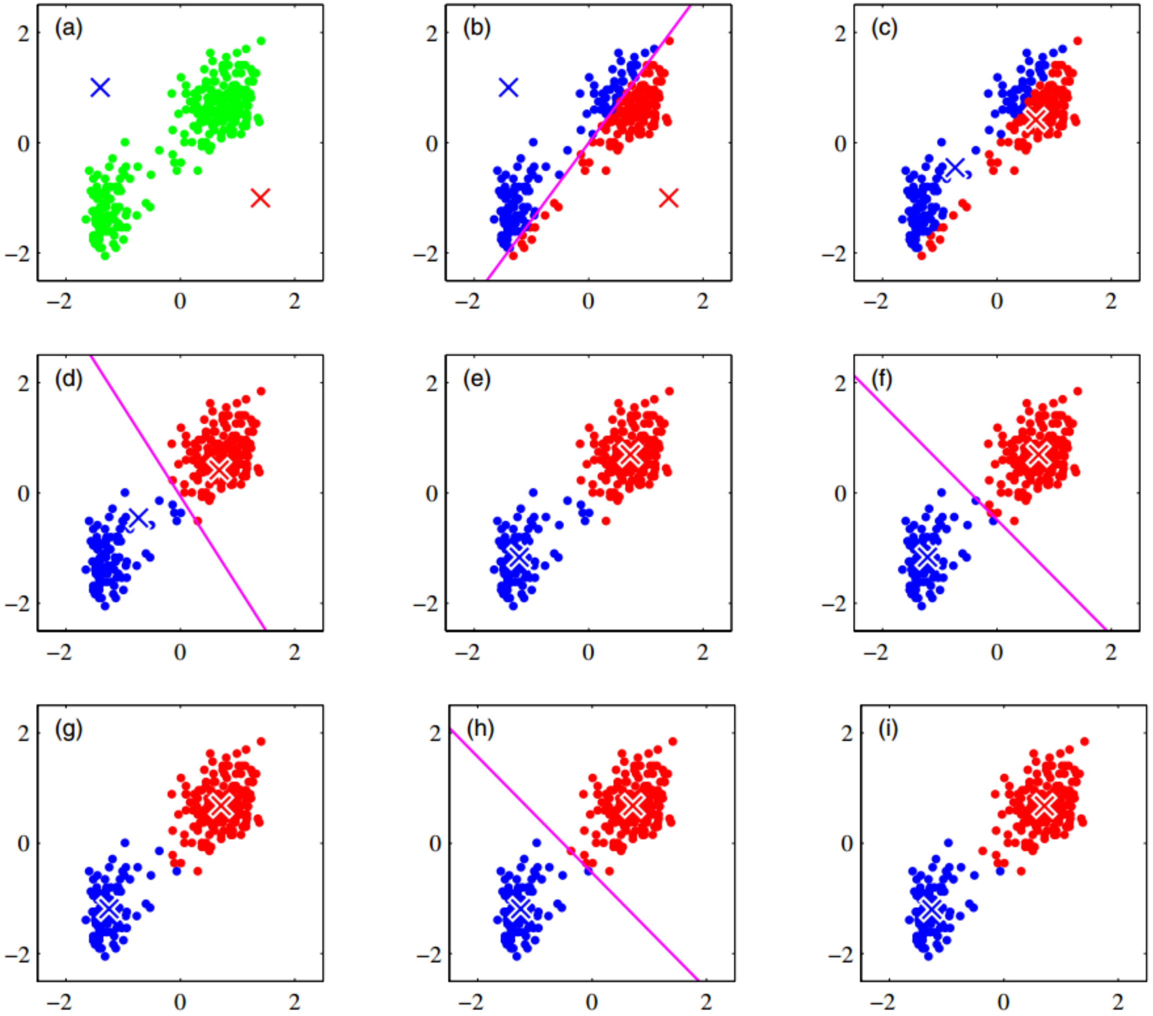
当有数据集{x₁, x₂, x₃, … , x_N}时，希望找到K个聚类中心点，使得

$$J=\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\|\mathbf{x}_n-\boldsymbol{\mu}_k\|^2$$

最小。其中：

(1) r_{nk}是1-of-K编码的K维变量，只有一个维度的值为1，其他维度的值为0。表示节点n是否属于聚类k。

(2) μ_k表示聚类k的中心点的向量。



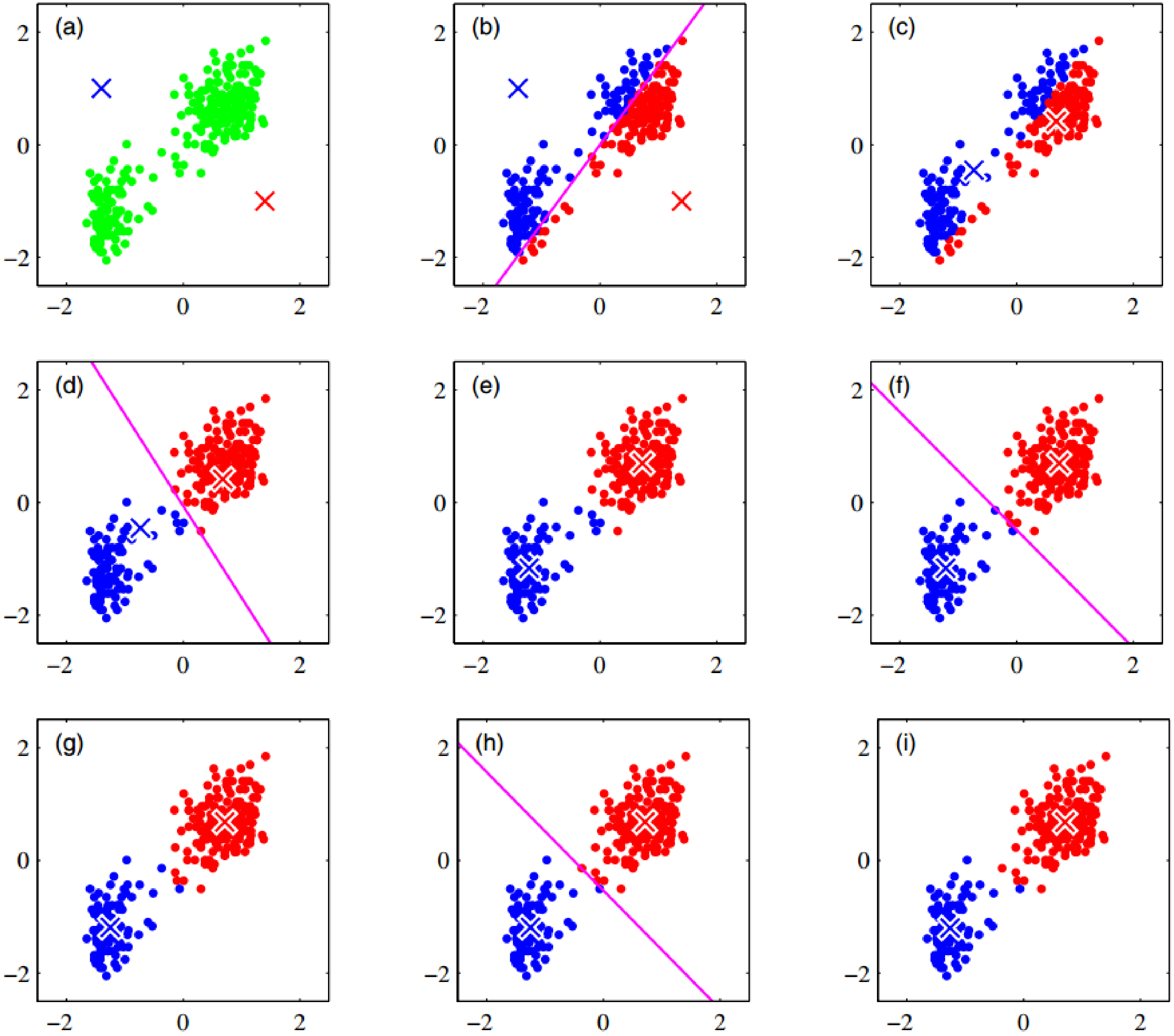
k-Means算法

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

- 目标函数有2类参数 r_{nk}和μ_k 要学习
- 算法流程

```
初始化 $\mu_k$
Loop {
    E步  $r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$
    M步  $\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$
}
```

- 是一种坐标调参法
- 是一种EM算法。hard-EM



k-Means算法

- 优点
  - 方法简单，易理解
- 缺点
  - 局部最优解
  - 计算量大
- 算法改进
  - k-Means++：效果优化
  - 基于三角不等式的性能优化
  - 基于k-d树的性能优化

EM算法应用：以GMM为例

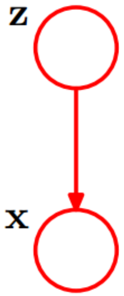
- GMM: Gaussian Mixture Model
- 高斯混合模型 or 混合高斯模型

- 一种混合模型
- 一种概率密度估计模型
- 一种图模型

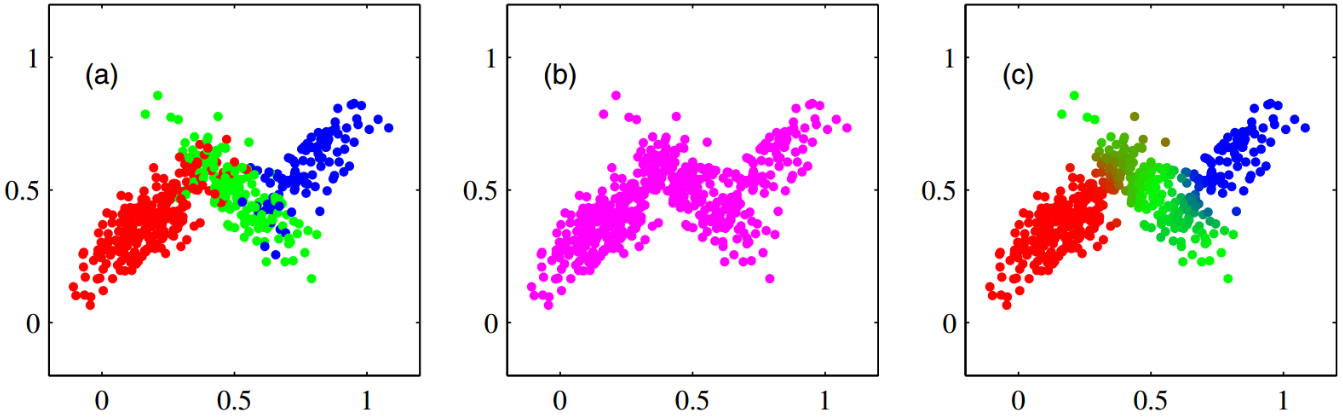
不但能做概率密度估计，

也可以做聚类

Soft-EM算法



Graphical representation of a mixture model, in which the joint distribution is expressed in the form $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$.



GMM

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

约束条件：

(1) $0 \leqslant \pi_k \leqslant 1$

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

引入隐变量z，1-of-K编码的K维向量。只有一个维度的值为1，其他维度的值为0。

$$p(z_k = 1) = \pi_k$$

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

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

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

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

GMM

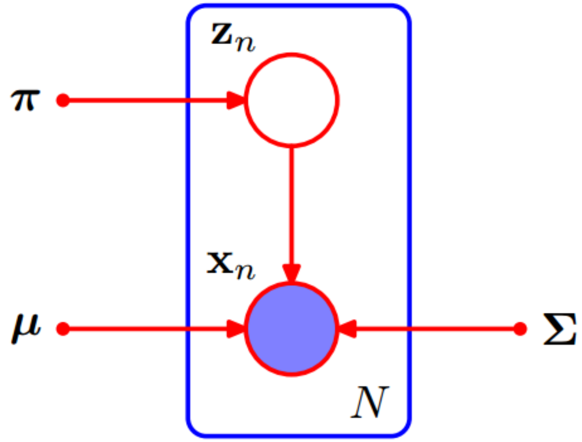
- 计算条件概率P(z|x) (很重要的一个数)

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

We shall 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 $\mathbf{x}$. As we shall see later, $\gamma(z_k)$ can also be viewed as the *responsibility* that component k takes for 'explaining' the observation $\mathbf{x}$.

- 优化目标函数 log-likelihood

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



EM for GMM

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

对 μ$_k$ 求偏导

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)$$

进一步化简得到

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

其中N$_k$是

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

对 Σ$_k$ 求偏导

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

对 π$_k$ 求偏导

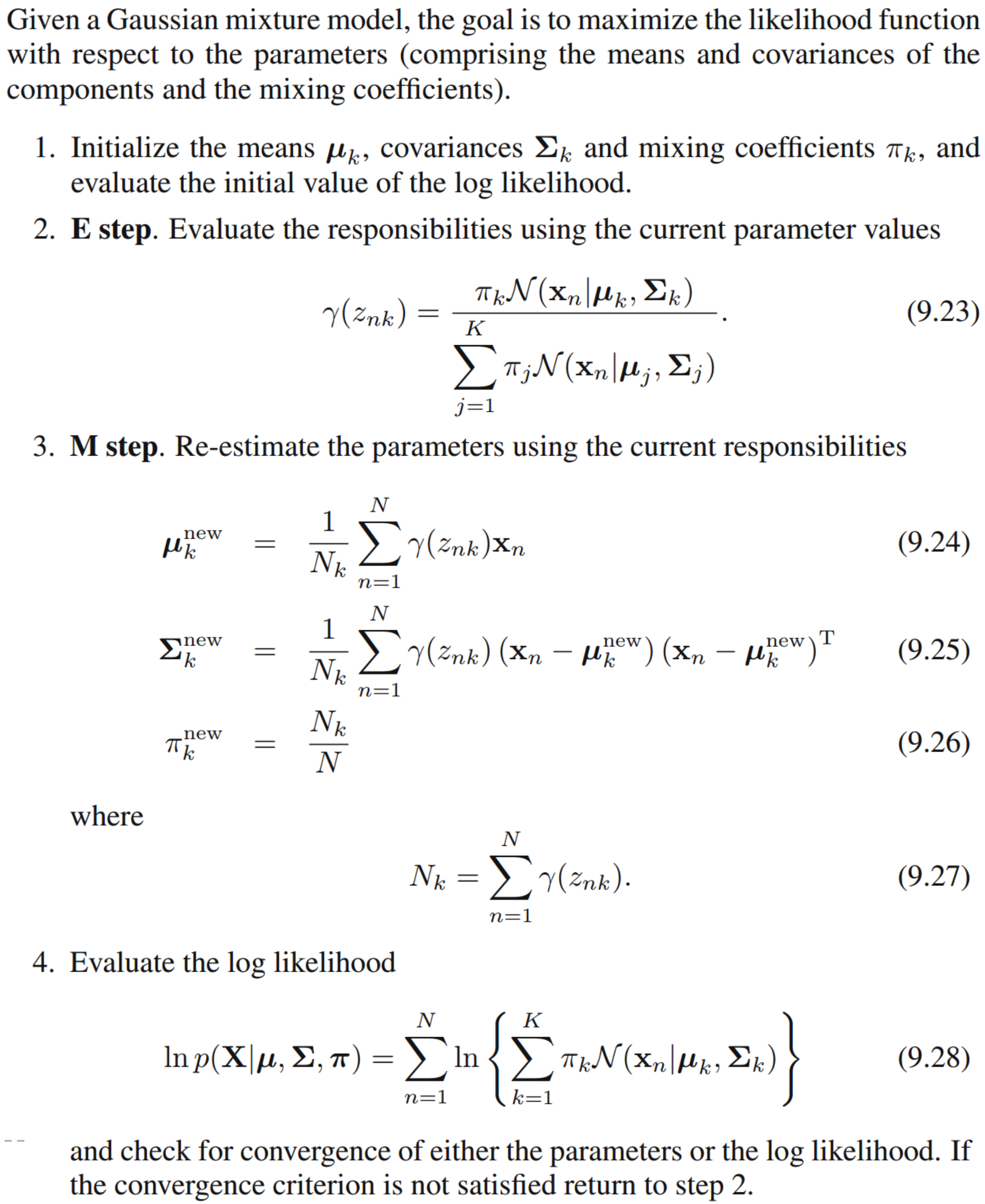
$$\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)$$

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} + \lambda$$

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

EM for GMM

- 算法伪代码



Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients).

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

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum\limits_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{9.23}$$

3. **M step**. Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \tag{9.24}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}} \tag{9.25}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N} \tag{9.26}$$

where

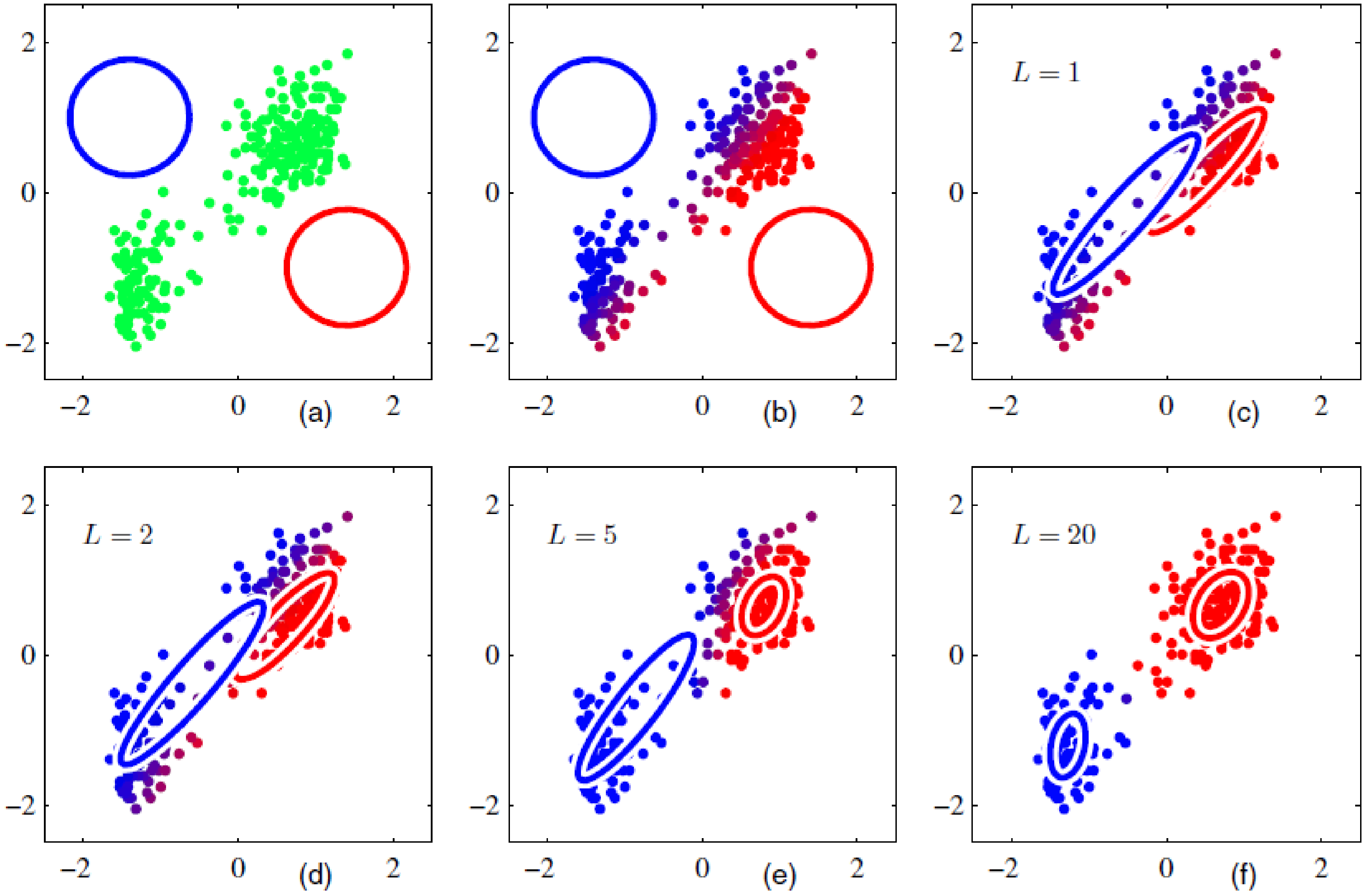
$$N_k = \sum_{n=1}^{N}\gamma(z_{nk}). \tag{9.27}$$

4. Evaluate the log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\} \tag{9.28}$$

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

EM for GMM



大纲

- 基础知识
- EM算法应用举例
- EM算法及其证明
- EM算法的变种

正经的EM算法陈述

"鸡生蛋，蛋生鸡"

- 以EM for MLE为例

- 变量说明：
  - X: 样本
  - Z: 隐变量
  - θ: 模型参数

- 优化目标（当Z为离散变量时）

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

{X, Z}: 完整数据；{X}: 不完整数据

P(X, Z | θ) 好算

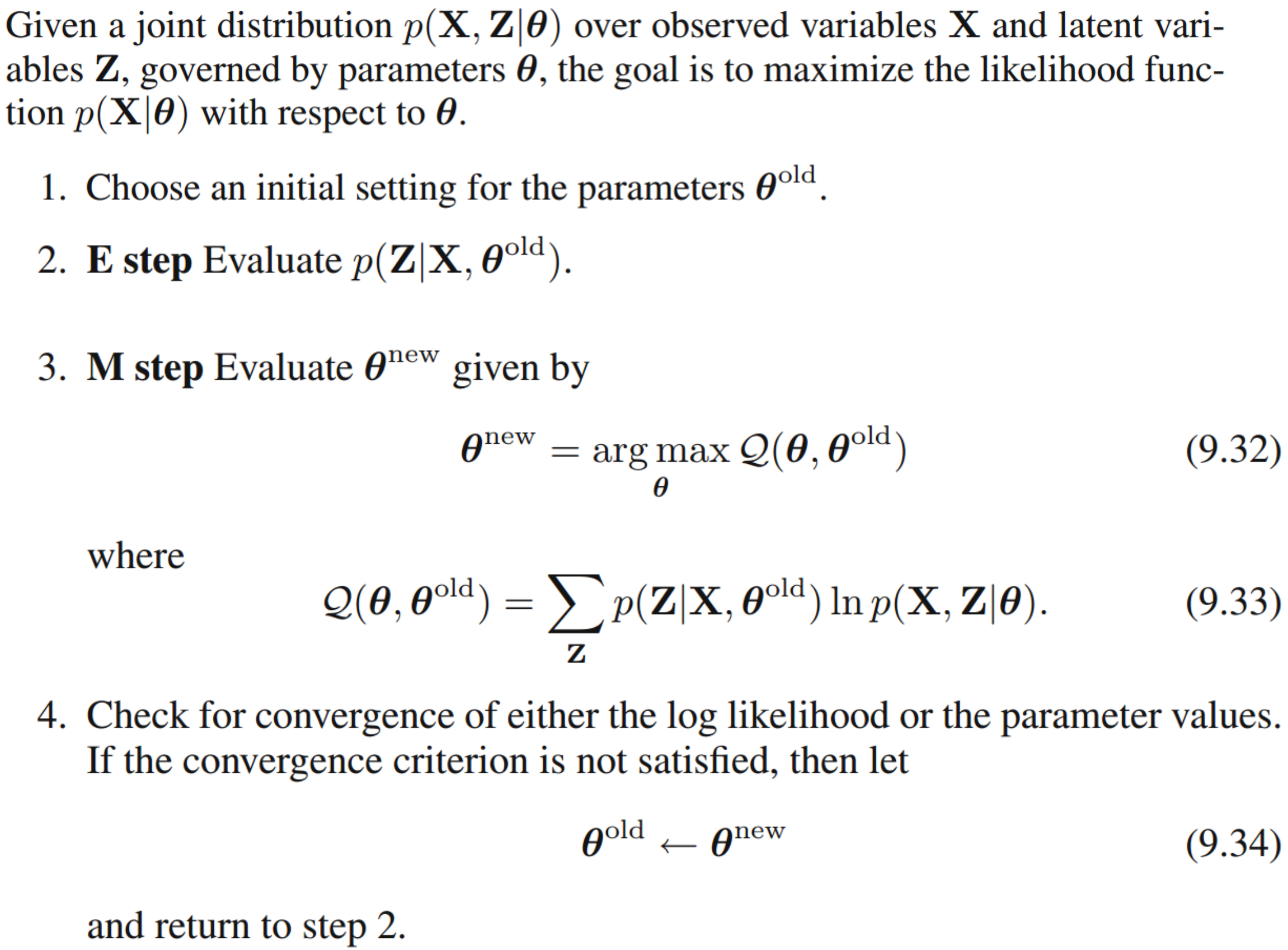
P(X | θ) 不好算

但P(Z | X, θ)好算

$$\sum_{Z} P(Z|X,\theta) lnP(X,Z|\theta)$$

The General EM Algorithm

```
Loop {
    E步：求期望（expectation）   → 什么函数关于什么分布的期望？
    M步：求极大（maximization）  → 关于什么函数的最大化？
}
```



Given a joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood function $p(\mathbf{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.

2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.

3. **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \underset{\boldsymbol{\theta}}{\arg\max}\, \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \tag{9.32}$$

where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \tag{9.33}$$

4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}} \tag{9.34}$$

and return to step 2.

EM算法的收敛性证明 (1)

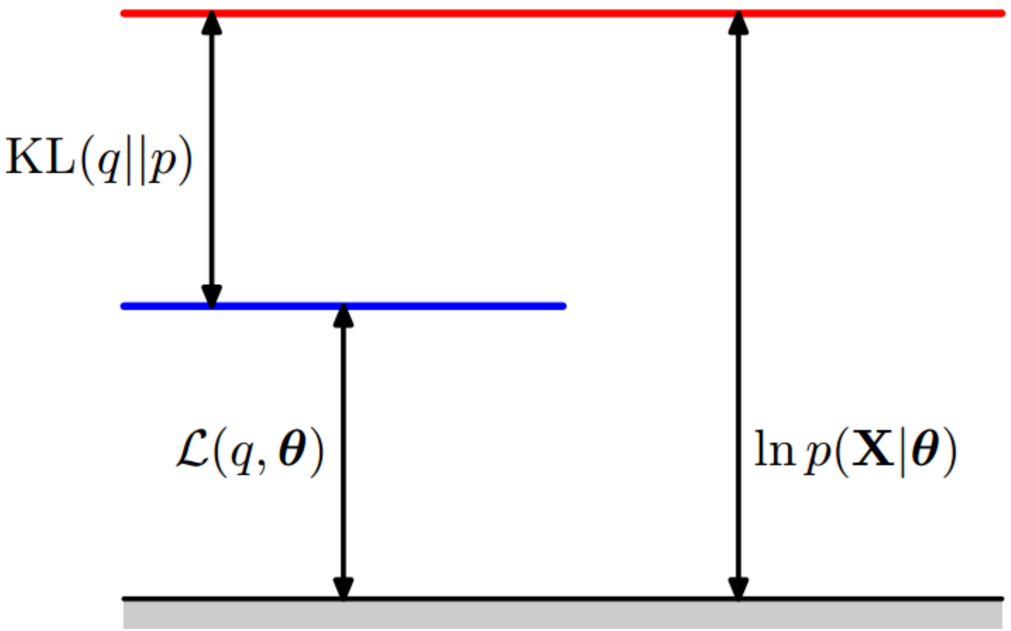
$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln\left\{\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})\right\}$$

基于等式 P(X | θ) = P(X, Z | θ) / P(Z | X, θ)，有

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q,\boldsymbol{\theta}) + \mathrm{KL}(q\|p)$$

其中，

$$\begin{aligned}\mathcal{L}(q,\boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\left\{\frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}\right\} \\ \mathrm{KL}(q\|p) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\left\{\frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})}\right\}\end{aligned}$$


$\mathrm{KL}(q\|p)$
$\mathcal{L}(q,\boldsymbol{\theta})$
$\ln p(\mathbf{X}|\boldsymbol{\theta})$

因为KL(q || p)恒大于等于0，所以L(q, θ)是ln P(X | θ)的下界

EM算法的收敛性证明 (2)

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q,\boldsymbol{\theta}) + \mathrm{KL}(q\|p)$$

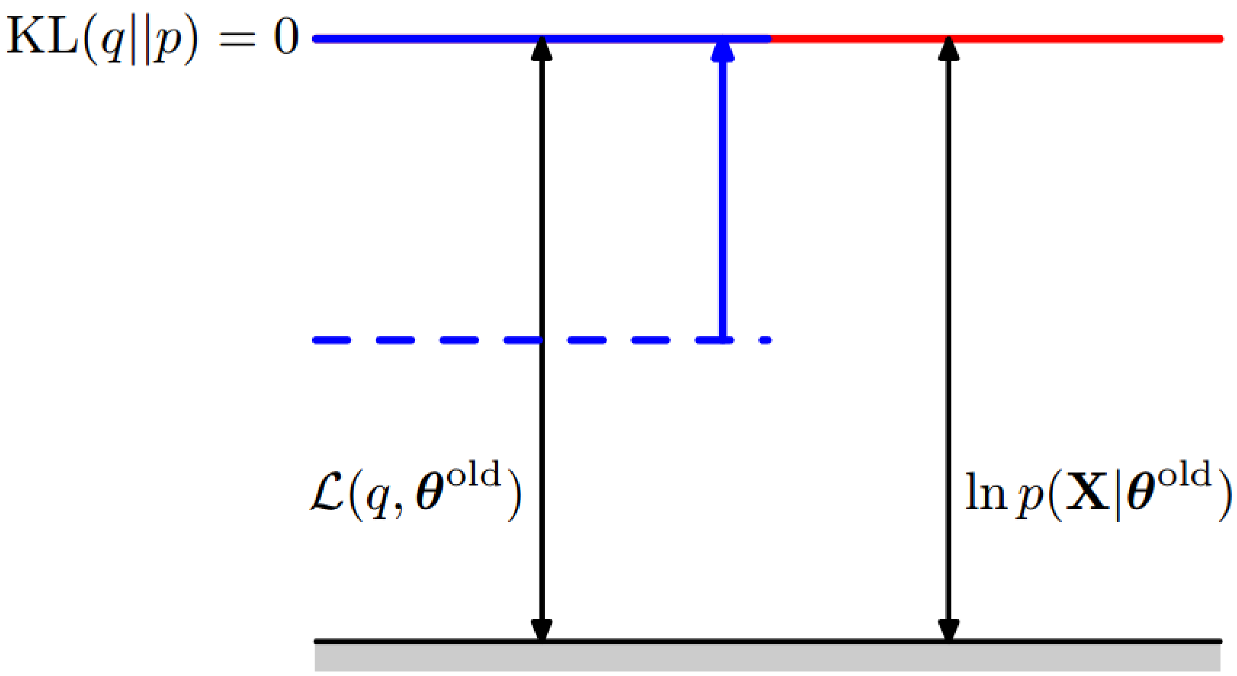
$$\mathcal{L}(q,\boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

$$\mathrm{KL}(q\|p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

假设当前轮θ的值为θ^{old}

固定θ^{old}，寻找q(Z)使得$L(q, \theta^{old})$最大

Illustration of the E step of the EM algorithm. The q distribution is set equal to the posterior distribution for the current parameter values $\boldsymbol{\theta}^{\mathrm{old}}$, causing the lower bound to move up to the same value as the log likelihood function, with the KL divergence vanishing.


$\mathrm{KL}(q\|p) = 0$
$\mathcal{L}(q,\boldsymbol{\theta}^{\mathrm{old}})$
$\ln p(\mathbf{X}|\boldsymbol{\theta}^{\mathrm{old}})$

EM算法的收敛性证明 (3)

$$\mathcal{L}(q,\boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

为什么是L(q, θ)？

将 q(Z) = P(Z | X, θ^{old})带入L(q, θ)表达式得到

$$\begin{aligned}\mathcal{L}(q,\boldsymbol{\theta}) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \\ &= \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}) + \text{const} \qquad (9.74)\end{aligned}$$

伪代码中
的Q函数

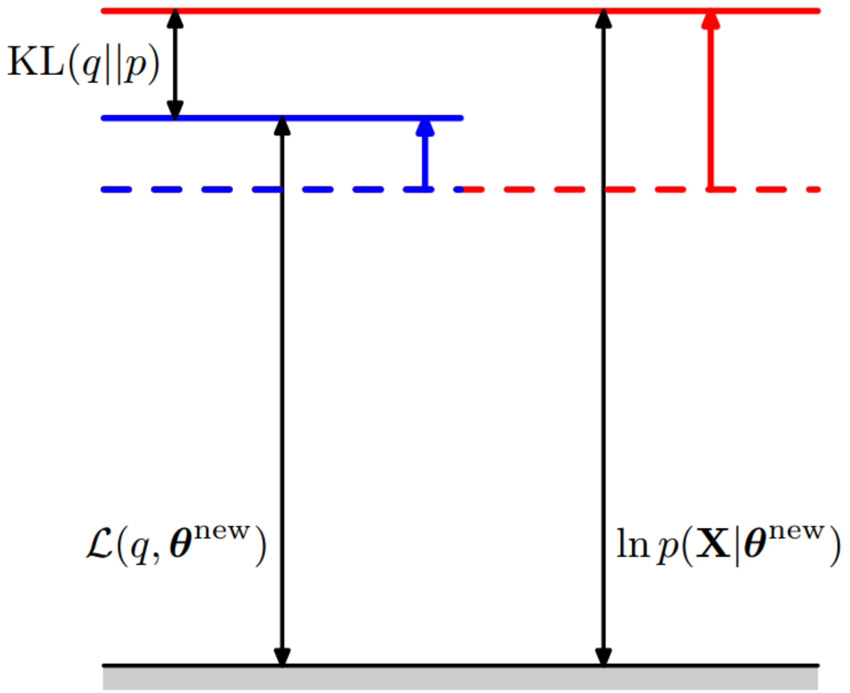
EM算法的收敛性证明 (4)

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q,\boldsymbol{\theta}) + \mathrm{KL}(q\|p)$$

$$\begin{aligned} \mathcal{L}(q,\boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \\ \mathrm{KL}(q\|p) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \end{aligned}$$

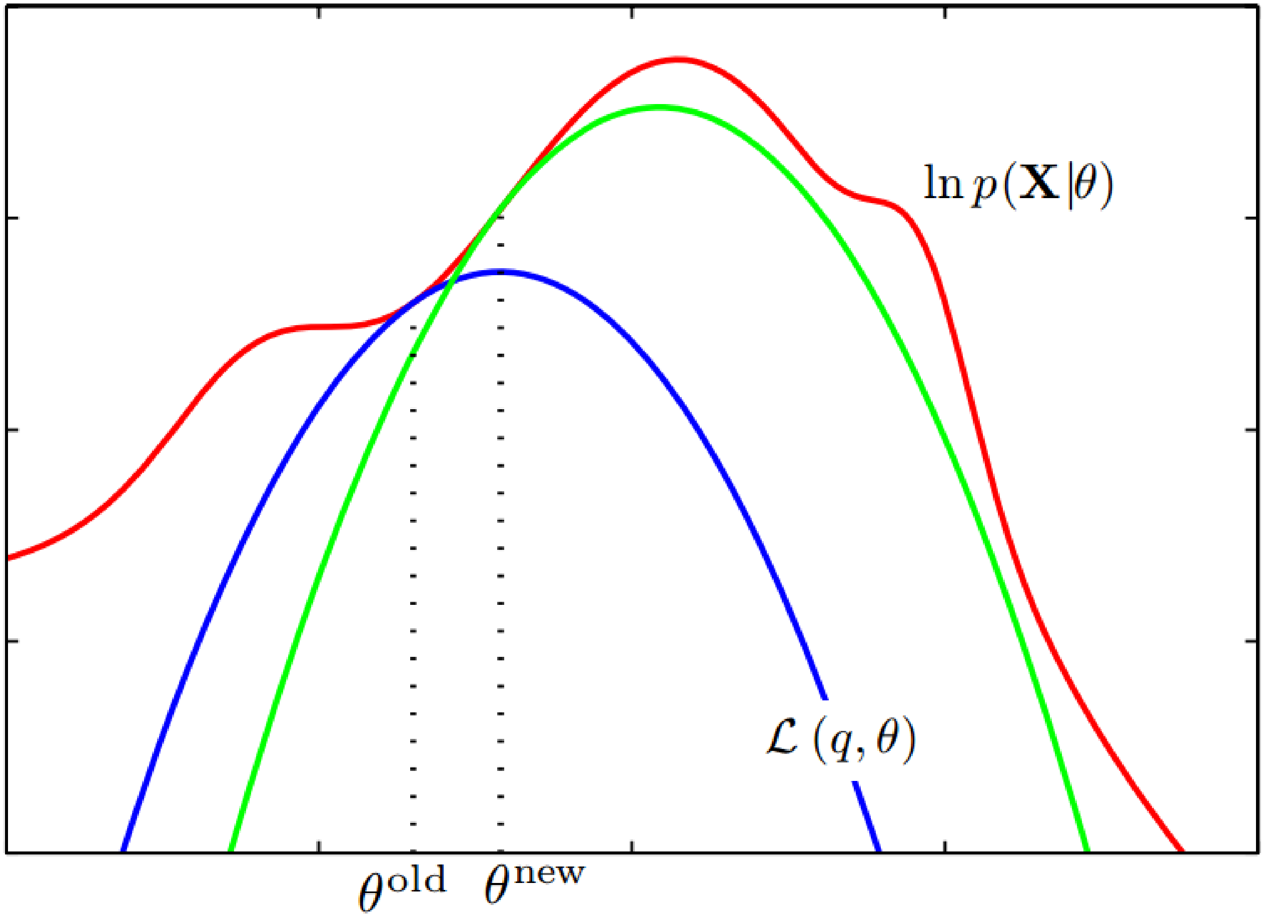
固定q(Z)，寻找θ^{new}使得L(q, θ)最大

Illustration of the M step of the EM algorithm. The distribution $q(\mathbf{Z})$ is held fixed and the lower bound $\mathcal{L}(q,\boldsymbol{\theta})$ is maximized with respect to the parameter vector $\boldsymbol{\theta}$ to give a revised value $\boldsymbol{\theta}^{\mathrm{new}}$. Because the KL divergence is nonnegative, this causes the log likelihood $\ln p(\mathbf{X}|\boldsymbol{\theta})$ to increase by at least as much as the lower bound does.


$\mathrm{KL}(q\|p)$
$\mathcal{L}(q,\boldsymbol{\theta}^{\mathrm{new}})$
$\ln p(\mathbf{X}|\boldsymbol{\theta}^{\mathrm{new}})$

EM算法的效果示意图

The EM algorithm involves alternately computing a lower bound on the log likelihood for the current parameter values and then maximizing this bound to obtain the new parameter values. See the text for a full discussion.



思考

EM for PLSA ?

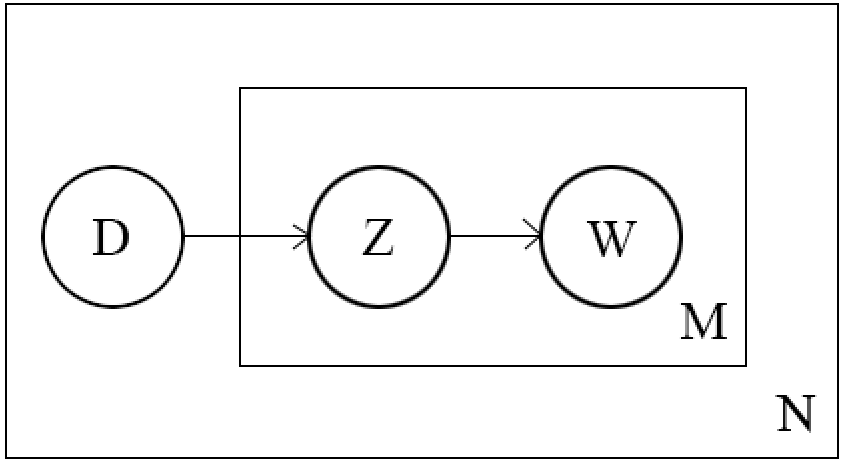
PLSA： Probabilistic Latent Semantic Analysis

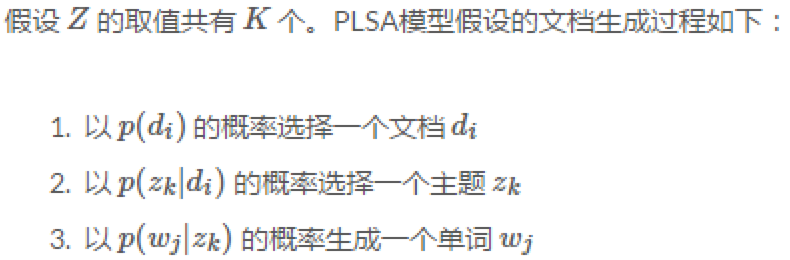
概率潜在语义分析

隐变量Z 表示 主题topic

{D,W, Z}表示完整数据

{D,W}表示不完整数据

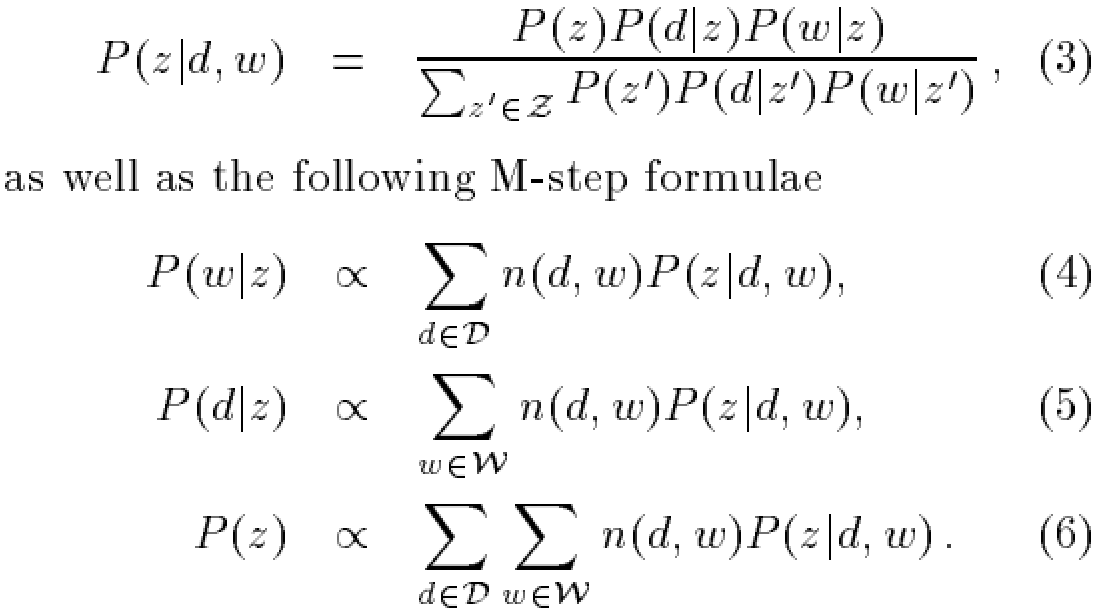




假设 Z 的取值共有 K 个。PLSA模型假设的文档生成过程如下：

1. 以 $p(d_i)$ 的概率选择一个文档 d_i
2. 以 $p(z_k|d_i)$ 的概率选择一个主题 z_k
3. 以 $p(w_j|z_k)$ 的概率生成一个单词 w_j

EM for PLSA



$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'\in\mathcal{Z}} P(z')P(d|z')P(w|z')}, \qquad (3)$$

as well as the following M-step formulae

$$P(w|z) \propto \sum_{d\in\mathcal{D}} n(d,w)P(z|d,w), \qquad (4)$$

$$P(d|z) \propto \sum_{w\in\mathcal{W}} n(d,w)P(z|d,w), \qquad (5)$$

$$P(z) \propto \sum_{d\in\mathcal{D}} \sum_{w\in\mathcal{W}} n(d,w)P(z|d,w)\,. \qquad (6)$$

参考文献

[1] Thomas Hofmann. Probabilistic Latent Semantic Analysis. UAI 1999.

[2] http://zhikaizhang.cn/2016/06/17/%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86%E4%B9%8BPLSA/

大纲

- 基础知识
- EM算法应用举例
- EM算法及其证明
- EM算法的变种

EM for MAP

- 极大后验概率估计（Maximum A Posterior estimation, MAP）
  - 优化目标 P(θ | X)。NOTE: 不同于MLE中的 P(X | θ)
  - 因为 P(θ | X) = P(θ, X) / P(X)，所以有
  - ln P(θ | X) = ln P(θ, X) - ln P(X)，继续展开有

$$\begin{aligned}\ln p(\boldsymbol{\theta}|\mathbf{X}) &= \overbrace{\mathcal{L}(q,\boldsymbol{\theta}) + \mathrm{KL}(q\|p)}^{\text{ln P(X | θ)}} + \ln p(\boldsymbol{\theta}) - \ln p(\mathbf{X}) \\ &\geqslant \mathcal{L}(q,\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - \overbrace{\ln p(\mathbf{X})}^{\text{常量}}.\end{aligned}$$

  - P(θ)是关于θ的先验分布
  - E步不变，M步加一项 ln P(θ) 后再求关于θ的最大值

Variational EM

- 当E步的$P(Z \mid X, \theta^{old})$不好计算时
- 可以基于KL距离找个与 $P(Z \mid X, \theta^{old})$ 分布近似的$q(Z)$来代替，且 $L(q, \theta^{old})$ 大于 $L(q^{old}, \theta^{old})$

Generalization EM

- 当M步不能基于梯度直接给出新参数的解析解时
- 意味着：在M步，还需要内嵌一个迭代算法计算θ^{new}
- 可以采用 非线性优化算法（比如共轭梯度法）或者坐标调参法

```
Loop {
    E步: 求期望（expectation）
    M步：求极大（maximization）Loop {
        计算$\theta^{new\_of\_inner\_loop}$
    } 得到 $\theta^{new}$
}
```

Online EM

- Online learning vs batch learning
- Online EM vs batch EM

其他变种

- 融入feature后的EM算法
  - Taylor Berg-Kirkpatrick, et al. Painless Unsupervised Learning with Features. ACL 2010.

主要参考文献

- Christopher M. Bishop. 《Pattern Recognition and Machine Learning》
- 李航. 《统计学习方法》

谢谢