第3.2节 贝叶斯学习

内　容

- 贝叶斯理论概述
- Brute-Force贝叶斯分类器
- 两种概率学习算法
  - 贝叶斯最优分类器
  - 朴素贝叶斯分类器
- EM算法与混合模型

概述

- 贝叶斯推理提供了一种概率手段，基于如下的假设：待考察的量遵循某概率分布，且可根据这些概率及已观察到的数据进行推理，以作出最优的决策。
- 贝叶斯推理为衡量多个假设的置信度提供了定量的方法
- 贝叶斯推理为直接操作概率的学习算法提供了基础，也为其他算法的分析提供了理论框架

概述

- 贝叶斯学习算法与机器学习相关的两个原因：
  - 贝叶斯学习算法能够计算显示假设概率
  - 贝叶斯方法为理解多数学习算法提供了一种有效的分析手段，而这些算法不一定直接操纵概率数据，比如
    - 决策树是概率
    - 神经网络是概率
    - SVM是概率…

Machine Learning: a Probabilistic Perspective" by Kevin Patrick Murphy

贝叶斯方法的难度

- 问题之一：需要概率的先验知识
  - 当概率预先未知时，可以基于背景知识、预先准备好的数据以及基准分布的假定来估计这些概率
- 问题之二：确定贝叶斯最优假设的计算代价比较大
  - 在某些特定情形下，大多通过条件独立性假设,降低计算代价

内 容

- 贝叶斯理论概述
- Brute-Force贝叶斯分类器
- 两种直接操作概率的学习算法
  - 贝叶斯最优分类器
  - 朴素贝叶斯分类器
- EM算法与混合模型

贝叶斯公式

- 贝叶斯公式提供了从先验概率P(h)、P(X)和P(X |h)计算后验概率P(h| X)的方法

$$P(h\mid X)=\frac{P(X\mid h)P(h)}{P(X)}$$

- P(h| X)随着P(h)和P(X |h)的增长而增长，随着P(X)的增长而减少
  - 即如果X独立于h被观察到的可能性越大，那么X对h的支持度越小

最大后验假设

- 学习器在候选假设集合H中寻找给定数据X时可能性最大的假设h，h被称为极大后验假设（MAP）
- 确定MAP的计算式如下

$$h_{MAP} = \arg\max_{h\in H} P(h \mid X) = \arg\max_{h\in H} \frac{P(X \mid h)P(h)}{P(X)} = \arg\max_{h\in H} P(X \mid h)P(h)$$

最后一步，去掉了P(X)，因为它是不依赖于h的常量

极大似然假设

- 在某些情况下，可假定H中每个假设有相同的先验概率，这样式子可以进一步简化，只需考虑P(X |h)来寻找极大可能假设。
- P(X |h)常被称为给定h时数据X的似然度，而使P(X |h)最大的假设被称为极大似然假设

$$h_{ML} = \arg\max_{h \in H} P(X \mid h)$$

- 只要假设取值的概率之和为1，假设空间H可扩展为任意互斥命题集合

Brute-Force贝叶斯学习总结

- 概念学习问题
  - 有限假设空间H定义在实例空间X上，任务是学习某个目标概念c。
- Brute-Force MAP学习算法
  - 对于H中每个假设h，计算后验概率
  - 输出有最高后验概率的假设

$$P(h|X)=\frac{P(X|h)P(h)}{P(X)}$$

$$h_{MAP}=\arg\max_{h\in H} P(h|X)$$

- 上面算法需要较大计算量
  - 因为它要计算每个假设的后验概率，对于大的假设空间显得不切实际，但是它提供了一个标准以判断其他概念学习算法的性能

内 容

- 贝叶斯理论概述
- Brute-Force贝叶斯分类器
- 两种概率学习算法
  - 贝叶斯最优分类器
  - 朴素贝叶斯分类器
- EM算法与混合模型

贝叶斯最优分类器

- 前面我们讨论的问题是：
  - 给定训练数据，最可能的假设是什么？
- 另一个相关的更有意义的问题是：
  - 给定训练数据，对新实例的最可能的分类是什么？
- 显然，第二个问题的解决可以将第一个问题的结果（MAP）应用到新实例上得到
- 还存在更好的算法？

贝叶斯最优分类器

- 例子
  - 考虑一个包含三个假设h_1，h_2，h_3的假设空间。假定已知训练数据时三个假设的后验概率分别是0.4，0.3，0.3，因此h_1为MAP假设。若一新实例x被h_1分类为正，被h_2和h_3分类为负，计算所有假设，x为正例的概率为0.4，为反例的概率为0.6。这时最可能的分类与MAP假设生成的分类不同。

贝叶斯最优分类器

- 一般而言，新实例的最可能分类可通过合并所有假设的预测得到，用后验概率来加权。
- 如果新实例的可能分类可取某集合Y中的任一值y_j，那么概率$P(y_j| X)$表示新实例分类为y_j的概率

$$P(y_j \mid X) = \sum_{h_i \in H} P(y_j \mid h_i) P(h_i \mid X)$$

- 新实例的最优分类为使$P(y_j| X)$最大的y_j值，贝叶斯最优分类器为：

$$\arg\max_{y_j \in Y} \sum_{h_i \in H} P(y_j \mid h_i) P(h_i \mid X)$$

内 容

- 贝叶斯理论概述
- Brute-Force贝叶斯分类器
- 两种概率学习算法
  - 贝叶斯最优分类器
  - 朴素贝叶斯分类器
- EM算法与混合模型

朴素贝叶斯分类器*

- 工程应用的学习任务：**每个实例x可由特征联合描述，而目标函数f(x)从某有限集Y中取值，忽略假设**
- 贝叶斯方法的新实例分类目标是在给定描述实例的特征<$x_1,\ldots,x_n$>下，得到最可能的目标值y_{MAP}

$$y_{MAP} = \underset{y_j}{\arg\max}\, P(y_j \mid x_1,\ldots,x_K)$$

- 使用贝叶斯公式变化上式

$$\begin{aligned} y_{MAP} &= \underset{y_j \in V}{\arg\max} \frac{P(x_1,\ldots,x_K \mid y_j)P(y_j)}{P(x_1,\ldots,x_K)} \\ &= \underset{y_j \in V}{\arg\max}\, P(x_1,\ldots,x_K \mid y_j)P(y_j) \end{aligned}$$

朴素贝叶斯分类器*

- 基于训练数据估计式(上页)中的两个数据项的值
  - 估计P(y_j)很容易：计算每个目标值y_j出现在训练数据中的频率
  - **估计 $P(x_1,\ldots,x_K \mid y_j)$ 遇到数据稀疏问题，除非有一个非常大的训练数据集，否则难以可靠估计**
- 朴素贝叶斯分类器引入一个简单的假设避免数据稀疏问题：在给定目标值时，属性值之间相互条件独立，即

$$P(x_1,\ldots,x_K \mid y_j)=\prod_i P(x_i \mid y_j)$$

朴素贝叶斯分类器*

- 朴素贝叶斯分类器的定义：
  - 从训练数据中估计不同 $P(x_i \mid y_j)$ 项的数量比要估计 $P(x_1,\ldots,x_K \mid y_j)$ 项所需的量小得多
  - 只要条件独立性得到满足，朴素贝叶斯分类y_{NB}等于MAP分类，否则是近似
  - 朴素贝叶斯分类器与其他已介绍的学习方法的一个区别：
    - 没有明确地搜索可能假设空间的过程（假设的形成不需要搜索，只是简单地计算训练样例中不同数据组合的出现频率）

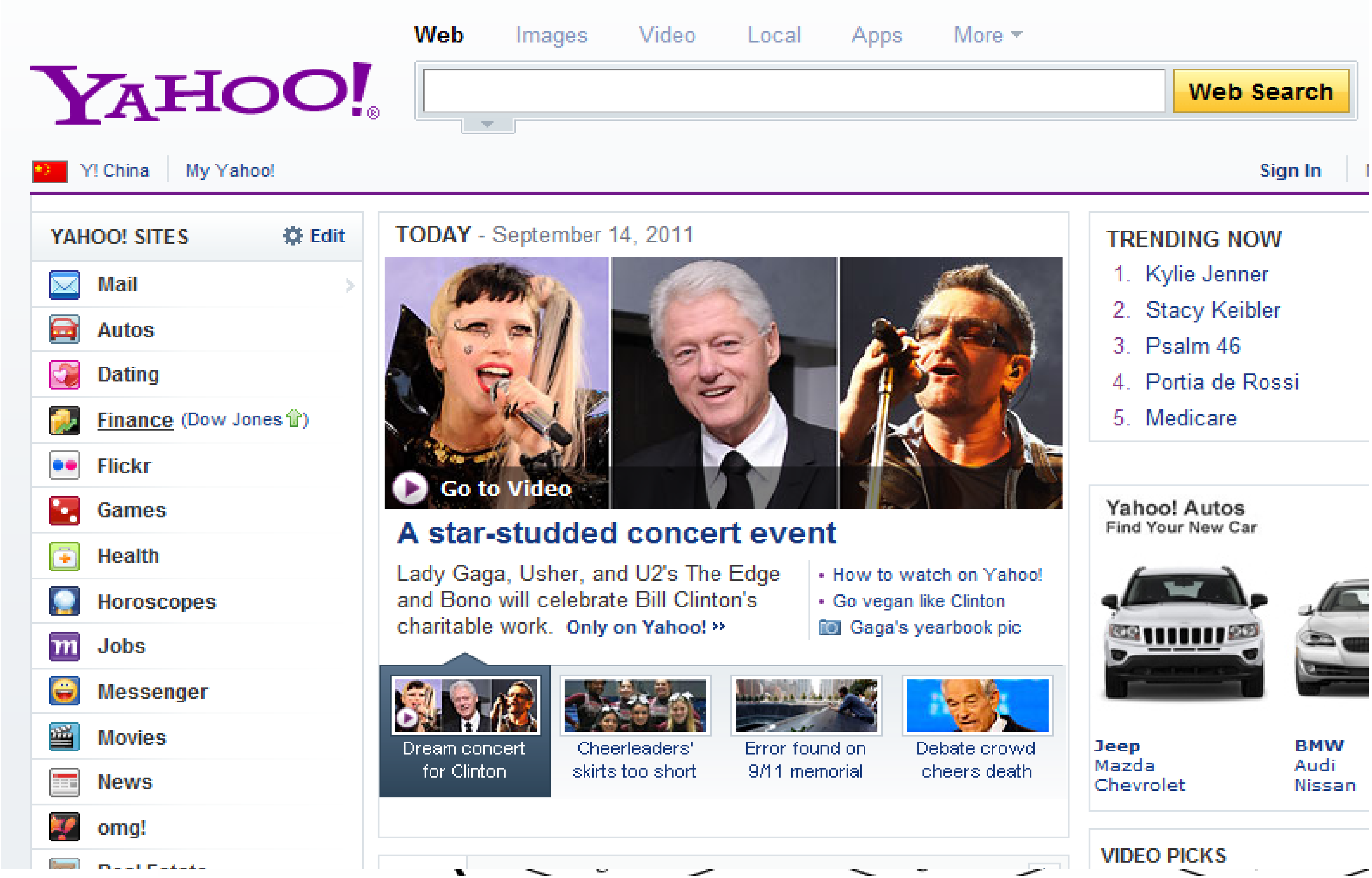
$$y_{NB} = \arg\max_{y_j \in V} P(y_j)\prod_i P(x_i \mid y_j)$$

朴素贝叶斯分类器计算是否去打球

表-1是否去打球的数据统计—训练数据

编号	天气	温度	湿度	风	是否去打球
1	晴天	炎热	高	弱	不去
2	晴天	炎热	高	强	不去
3	阴天	炎热	高	弱	去
4	下雨	适中	高	弱	去
5	下雨	寒冷	正常	弱	去
6	下雨	寒冷	正常	强	不去
7	阴天	寒冷	正常	强	去
8	晴天	适中	高	弱	不去
9	晴天	寒冷	正常	弱	去
10	下雨	适中	正常	弱	去
11	晴天	适中	正常	强	去
12	阴天	适中	高	强	去
13	阴天	炎热	正常	弱	去
14	下雨	适中	高	强	不去

举例：学习分类文本



Web
Images
Video
Local
Apps
More
YAHOO!
Web Search
Y! China
My Yahoo!
Sign In
YAHOO! SITES
Edit
Mail
Autos
Dating
Finance (Dow Jones)
Flickr
Games
Health
Horoscopes
Jobs
Messenger
Movies
News
omg!
TODAY - September 14, 2011
Go to Video
A star-studded concert event
Lady Gaga, Usher, and U2's The Edge and Bono will celebrate Bill Clinton's charitable work. Only on Yahoo!
How to watch on Yahoo!
Go vegan like Clinton
Gaga's yearbook pic
Dream concert for Clinton
Cheerleaders' skirts too short
Error found on 9/11 memorial
Debate crowd cheers death
TRENDING NOW
1. Kylie Jenner
2. Stacy Keibler
3. Psalm 46
4. Portia de Rossi
5. Medicare
Yahoo! Autos
Find Your New Car
Jeep
Mazda
Chevrolet
BMW
Audi
Nissan
VIDEO PICKS

内 容

- 贝叶斯理论概述
- Brute-Force贝叶斯分类器
- 两种概率学习算法
  - 贝叶斯最优分类器
  - 朴素贝叶斯分类器
- EM算法与混合模型

EM算法与GMMs

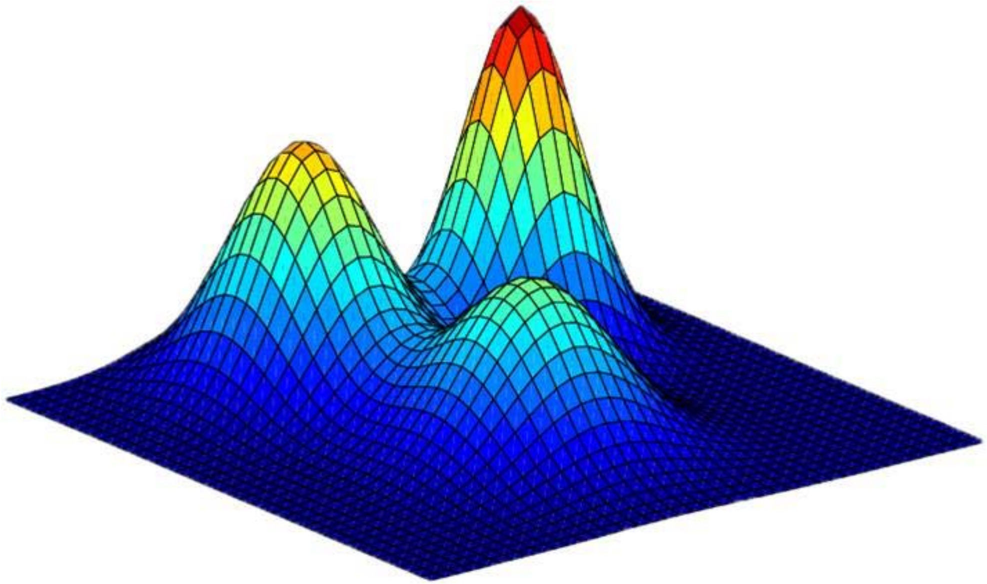
- 如何通过多个假设获得更准确的概率分布?
- EM算法是存在隐含变量时广泛使用的一种学习方法，可用于变量的值从来没有被直接观察到的情形，只要这些变量所遵循的概率分布的一般形式已知
  - 混合概率模型学习
  - 期望最大学习

Gaussian Mixture Models (GMMs) 高斯混合模型

$$P(x) = \sum_{k=1}^{K} w_k \cdot N\left(\mu_k, \sigma_k\right)$$

$$P(x) = \sum_{k=1}^{K} w_k \cdot \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)$$

$$P(x) = \sum_{k=1}^{K} w_k \cdot \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)\right)$$



GMMS的参数估计

- 当样本$x^{(1)},\ldots,x^{(N)}$符合单一高斯分布假设时，很容易计算该分布的参数的极大似然假设。以其均值参数为例：

$$\mu_{ML} = \arg\min_{\mu} \sum_{i=1}^{N} (x^{(i)} - \mu)^2 = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$$

- 问题涉及k个不同高斯分布，而且**不知道哪个样本是哪个分布（假设）产生的**。这是一个**隐参数模型，一种Chicken-And-Egg 问题**
- 每个样本的完整描述$y^{(i)}=<x^{(i)}, z_{ij}>$，其中$x^{(n)}$是第i个样本的观测值，$z_{ij}$表示样本$x^{(i)}$属于第j个正态分布，是隐含变量

GMMS的参数估计-EM算法

Expectation Maximum （EM）算法

- EM算法根据当前假设<$\mu_1 ... \mu_k$>，迭代地(iteratively)估计隐含变量z_{ij}的期望值，并用隐含变量的期望值重新计算极大似然假设
- 先将假设随机初始化为h=<$\mu_1, \ldots, \mu_k$>
  - 计算每个隐含变量z_{ij}的期望值$E[z_{ij}]$
  - 计算一个新的极大似然假设h'=<$\mu'_1, \ldots, \mu'_k$>，假定每个隐含变量z_{ij}所取值是第一步得到的期望值$E[z_{ij}]$。
  - 将假设替换为h'=<$\mu'_1, \ldots,\mu'_k$>，然后循环迭代

GMMS的参数估计-EM算法

- 第一步：计算E[z_{ij}]，表示样本$x^{(i)}$由第j个高斯分布生成的期望

$$E[z_{ij}] = \frac{p(x = x^{(i)} \mid \mu = \mu_j)}{\sum_{k=1}^{K} p(x = x^{(i)} \mid \mu = \mu_k)} = \frac{e^{-\frac{1}{2\sigma^2}(x^{(i)} - \mu_j)^2}}{\sum_{k=1}^{K} e^{-\frac{1}{2\sigma^2}(x^{(i)} - \mu_k)^2}}$$

- 第二步：使用第一步得到的E[z_{ij}]来导出新的极大似然假设

$$\mu_j = \frac{\sum_{i=1}^{N} E[z_{ij}] x^{(i)}}{\sum_{i=1}^{N} E[z_{ij}]}$$

第二步中表达式类似于求均值，只是变成了加权样本均值

GMMS的参数估计-EM算法

- EM算法的要点：
  - 当前的假设用于估计未知变量，而这些变量的期望值再被用于改进假设
- 原理
  - 算法的每一次循环中，**EM算法能使似然P(X|h)增加**，除非P(X|h)达到局部最大，
  - 算法符合极大似然估计原理，收敛到一个局部最大似然假设

GMMS的参数估计-EM算法

- 贝叶斯问题框架
  - 估计k个正态分布的均值$\theta=\langle\mu_1 \ldots \mu_k\rangle$
  - 观察到的数据是$X=\{x^{(i)}\}$
  - 隐含变量$Z=\{z_{i1},\ldots,z_{ik}\}$表示k个正态分布中哪一个生成$x^{(i)}$
- 用于表达式L(h'|h)的推导
  - 单个样本的生成(Generative)概率

$$p(y^{(i)}\,|\,h')=p(x^{(i)},z_{i1},\ldots,z_{ik}\,|\,h')=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{k}z_{ij}(x^{(i)}-\mu_j')^2}$$

GMMS的参数估计-EM算法

- 所有实例的联合概率的对数

$$\ln P(Y \mid h') = \ln \prod_{i=1}^{N} p(y^{(i)} \mid h')$$

$$= \sum_{i=1}^{N} \ln p(y^{(i)} \mid h')$$

$$= \sum_{i=1}^{N} \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{j=1}^{k} z_{ij} (x^{(i)} - \mu_j^{'})^2 \right)$$

- 计算期望值**E**

$$E[\ln P(Y \mid h')] = E\left[\sum_{i=1}^{N} \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{j=1}^{K} z_{ij} (x^{(i)} - \mu_j^{'})^2 \right) \right]$$

$$= \sum_{i=1}^{N} \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{j=1}^{K} E[z_{ij}] (x^{(i)} - \mu_j^{'})^2 \right)$$

GMMS的参数估计-EM算法

- 求使L函数最大的假设**M**

$$\arg\max_{h'} L(h'|h) = \arg\max_{h'} \sum_{i=1}^{N}\left(\ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{j=1}^{K} E\left[z_{ij}\right]\left(x^{(i)} - \mu_j^{'}\right)^2\right)$$

- 解上式得到

$$\frac{\partial L}{\partial \mu_j{'}} = \sum_{i=1}^{N} E\left[z_{ij}\right]\left(x^{(i)} - \mu_j^{'}\right) = 0$$

- 得到

$$\mu_j \leftarrow \frac{\sum_{i=1}^{N} E\left[z_{ij}\right] x^{(i)}}{\sum_{i=1}^{N} E[z_{ij}]} \qquad \sigma_j{}^2 = \frac{\sum_{i=1}^{N} E\left[Z_{ij}\right]\left(x^{(i)} - \mu_j\right)^2}{\sum_{i=1}^{N} E\left[Z_{ij}\right]} \qquad E[z_{ij}] \leftarrow \frac{e^{-\frac{1}{2\sigma^2}(x^{(i)}-\mu_j)2}}{\sum_{j=1}^{K} e^{-\frac{1}{2\sigma^2}(x^{(i)}-\mu_j)^2}}$$

GMMS的参数估计

- EM算法总结
  - 在数据分布范围内初始化 μ_k ，用整体数据方差初始化每个混合模型的方差
  - While ($\sum_{k=1}^{K}\left|\mu_k - \mu_k^{'}\right| > \varepsilon$)

{

- E step

$$E[z_{ij}] \leftarrow \frac{e^{-\frac{1}{2\sigma^2}(x^{(i)}-\mu_j)2}}{\sum_{j=1}^{K} e^{-\frac{1}{2\sigma^2}(x^{(i)}-\mu_j)^2}}$$

- M Step

$$\mu_j \leftarrow \frac{\sum_{i=1}^{N} E\left[z_{ij}\right] x^{(i)}}{\sum_{i=1}^{N} E[z_{ij}]} \qquad \sigma_j{}^2 \leftarrow \frac{\sum_{i=1}^{N} E\left[Z_{ij}\right]\left(x^{(i)}-\mu_j\right)^2}{\sum_{i=1}^{N} E\left[Z_{ij}\right]}$$

}

GMMS的参数估计

- 高维数据EM算法
  - 在数据分布范围内初始化 μ_k ，用整体数据协方差初始化每个混合模型的协方差
  - While ($\sum_{k=1}^{K}\left|\mu_k - \mu_k^{'}\right| > \varepsilon$)
  - {
    - E step

$$E[z_{ij}] \leftarrow \frac{|\Sigma_j|^{-1/2} e^{-\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1}(x^{(i)}-\mu_j)}}{\sum_{k=1}^{K} |\Sigma_k|^{-1/2} e^{-\frac{1}{2}(x^{(i)}-\mu_k)^T \Sigma_k^{-1}(x^{(i)}-\mu_k)}}$$

    - M Step

$$\mu_j \leftarrow \frac{\sum_{i=1}^{N} E\left[z_{ij}\right] x^{(i)}}{\sum_{i=1}^{N} E[z_{ij}]} \quad \Sigma_j \leftarrow \frac{\sum_{i=1}^{N} E\left[Z_{ij}\right] \cdot \left\{\left(x^{(i)}-\mu_j\right)^T \left(x^{(i)}-\mu_j\right)\right\}}{\sum_{i=1}^{N} E\left[Z_{ij}\right]}$$

  - }

小结

- 概率学习方法利用关于不同假设的先验概率，估计后验值
- 贝叶斯方法确定的极大后验概率假设最可能成为最优假设
- 朴素贝叶斯分类器增加了独立性：特征在给定实例的分类时条件独立
- EM算法提供了一个通用的算法，在存在隐含变量时进行学习。算法开始于任意的初始假设，然后迭代地计算隐藏变量的期望值，再重新计算极大似然假设，这个过程收敛到一个局部极大似然假设和隐含变量的估计值, 可以进行多假设模型的参数估计与求解