周志华 著

MACHINE LEARNING

机器学习

清华大学出版社

崔磊

Tel: 15829735700(M)
QQ：362626744
E-Mail: leicui@nwu.edu.cn

本章课件致谢：霍轩

第七章：贝叶斯分类器

走进贝叶斯

- 贝叶斯简介：

  - 贝叶斯（约1701-1761），Thomas Bayes，英国数学家

  - 贝叶斯方法源于他为解决一个“逆向概率”问题写的一篇文章

什么是“逆概” 问题呢？

走进贝叶斯

比如商场举办了一个抽奖，抽奖桶里有10个球，其中2个白球，8个黑球，抽到白球就算你中奖。你伸手进去随便摸出1颗球，摸出中奖球的概率是多大?

走进贝叶斯

比如在上面的例子中，我们并不知道抽奖桶里黑白球的比例，而是摸出一个（一些）球，通过观察这个球的颜色，来预测这个桶里里白色球和黑色球的比例?

走进贝叶斯

Why贝叶斯

这是因为现实生活中的问题，大部分都是像上面的“**逆概率**”问题。

例如：天气预报说，明天降雨的概率是30%，这是什么意思呢?

我们无法像计算频率概率那样，重复地把明天过上100次，然后计算出大约有30次会下雨。

而是只能利用有限的信息（过去天气的测量数据），用贝叶斯定理来预测出明天下雨的概率是多少。

走进贝叶斯

Why贝叶斯

贝叶斯定理其实就是为了解决这类问题而诞生的，它可以根据过去的数据来预测出概率。

贝叶斯定理的思考方式为我们提供了明显有效的方法来帮助我们提供预测能力，以便更好地预测未来的商业、金融、以及日常生活。

设么是贝叶斯

- 条件概率： $P(A/B) = \dfrac{P(AB)}{P(B)}$
- 全概率公式： $P(A) = \sum_i P(A/B_i)P(B_i)$
- 贝叶斯(Bayes)公式： $P(B_i/A) = \dfrac{P(A/B_i)P(B_i)}{\sum_j P(A/B_j)P(B_j)}$

公式推导（加强记忆）

男生：60%
女生：40%

- 男生总是穿长裤，女生则一半穿长裤一半穿裙子
- **正向概率**：随机选取一个学生，他（她）穿长裤的概率和穿裙子的概率是多大
- **逆向概率**：迎面走来一个穿长裤的学生，你只看得见他（她）穿的是否长裤，而无法确定他（她）的性别，你能够推断出他（她）是女生的概率是多大吗？

公式推导（加强记忆）

- 假设学校里面人的总数是 U 个
- 穿长裤的（男生）：U * P(Boy) * P(Pants|Boy)
  - P(Boy) 是男生的概率 = 60%
  - P(Pants|Boy) 是条件概率，即在 Boy 这个条件下穿长裤的概率是多大，这里是 100%，因为所有男生都穿长裤
- 穿长裤的（女生）：U * P(Girl) * P(Pants|Girl)

公式推导（加强记忆）

- 求解：穿长裤的人里面有多少女生

  - 穿长裤总数：U * P(Boy) * P(Pants|Boy) + U * P(Girl) * P(Pants|Girl)

  - P(Girl|Pants) = U * P(Girl) * P(Pants|Girl)/穿长裤总数

    U * P(Girl) * P(Pants|Girl) / [U * P(Boy) * P(Pants|Boy) + U * P(Girl) * P(Pants|Girl)]

公式推导（加强记忆）

- 贝叶斯(Bayes)公式：

$$P(B/A) = \frac{P(B)\,P(A/B)}{P(A)}$$

③ 后验概率　① 先验概率　② 可能性函数

- 换个表达形式：

$$P(\text{类别}/\text{特征}) = \frac{P(\text{类别})\,P(\text{特征}/\text{类别})}{P(\text{特征})}$$

举例

举个栗子，给定数据如下：现在给我们的问题是，如果一对男女朋友，男生向女生求婚，男生的四个特点分别是“不帅”，“性格不好”，“身高矮”，“不上进”，请你判断一下女生是嫁还是不嫁？

帅？	性格好？	身高？	上进？	嫁与否
帅	不好	矮	不上进	不嫁
不帅	好	矮	上进	不嫁
帅	好	矮	上进	嫁
不帅	好	高	上进	嫁
帅	不好	矮	上进	不嫁
帅	不好	矮	上进	不嫁
帅	好	高	不上进	嫁
不帅	好	中	上进	嫁
帅	好	中	上进	嫁
不帅	不好	高	上进	嫁
帅	好	矮	不上进	不嫁
帅	好	矮	不上进	不嫁

举例

这是一个典型的分类问题，转为数学问题就是比较p(嫁|(不帅、性格不好、身高矮、不上进))与p(不嫁|(不帅、性格不好、身高矮、不上进))的概率，谁的概率大，我就能给出嫁或者不嫁的答案！

*** 这里我们联系到朴素贝叶斯公式 *：**

$$p(\text{嫁}|\text{不帅、性格不好、身高矮、不上进})=\frac{p(\text{不帅、性格不好、身高矮、不上进}|\text{嫁})*p(\text{嫁})}{p(\text{不帅、性格不好、身高矮、不上进})}$$

问题转化：我们需要求p(嫁|(不帅、性格不好、身高矮、不上进)，这是我们不知道的，但是通过贝叶斯公式可以转化为好求的三个量.

举例

那么这三个量是如何求得？

$$p(\text{嫁}|\text{不帅、性格不好、身高矮、不上进})=\frac{p(\text{不帅、性格不好、身高矮、不上进}|\text{嫁})*p(\text{嫁})}{p(\text{不帅、性格不好、身高矮、不上进})}$$

p(不帅、性格不好、身高矮、不上进|嫁) = p(不帅|嫁)*p(性格不好|嫁)*p(身高矮|嫁)*p(不上进|嫁)

那么我就要分别统计后面几个概率，也就得到了左边的概率。

问题：这样做，对吗？？？

举例

为什么需要假设特征之间相互独立呢?

① 我们这么想，假如没有这个假设，那么我们对右边联合概率的估计其实是不好的，比如我们这个例子有4个特征，其中帅包括{帅，不帅}，性格包括{不好，好，爆好}，身高包括{高，矮，中}，上进包括{不上进，上进}，那么四个特征的联合概率分布总共是4维空间，总个数为2*3*3*2=36个。

② 假如我们没有假设特征之间相互独立，那么我们统计的时候，就需要在整个特征空间中去找，比如统计p(不帅、性格不好、身高矮、不上进|嫁),我们就需要在嫁的条件下，去找四种特征全满足分别是不帅，性格不好，身高矮，不上进的人的个数，这样的话，由于数据的稀疏性，很容易统计到0的情况。 这样是不合适的。

举例

回到问题： 我们将上面公式整理一下如下：

$$p(\text{嫁}\mid\text{不帅、性格不好、身高矮、不上进})=\frac{p(\text{不帅、性格不好、身高矮、不上进}\mid\text{嫁})*p(\text{嫁})}{p(\text{不帅、性格不好、身高矮、不上进})}$$

$$=\frac{p(\text{不帅}\mid\text{嫁})*p(\text{性格不好}\mid\text{嫁})*p(\text{身高矮}\mid\text{嫁})*p(\text{不上进}\mid\text{嫁})*p(\text{嫁})}{p(\text{不帅})*p(\text{性格不好})*p(\text{身高矮})*p(\text{不上进})}$$

举例—计算过程

Step1: 先验概率 → p(嫁)=？？

首先我们整理训练数据中，嫁的样本数如下：

帅？	性格好？	身高？	上进？	嫁与否
帅	好	矮	上进	嫁
不帅	好	高	上进	嫁
帅	好	高	不上进	嫁
不帅	好	中	上进	嫁
帅	好	中	上进	嫁
不帅	不好	高	上进	嫁

则 p(嫁) = 6/12（总样本数） = 1/2

举例—计算过程

Step2: 求类条件概率 → p(不帅|嫁)、p(性格不好|嫁)、p(性格不好|嫁)、p（矮|嫁）

① p(不帅|嫁)=?

统计满足样本数如下：

帅？	性格好？	身高？	上进？	嫁与否
不帅	好	高	上进	嫁
不帅	好	中	上进	嫁
不帅	不好	高	上进	嫁

则p(不帅|嫁) = 3/6 = 1/2 在嫁的条件下，看不帅有多少

② p(性格不好|嫁)= ？

统计满足样本数如下：

帅？	性格好？	身高？	上进？	嫁与否
不帅	不好	高	上进	嫁

则p(性格不好|嫁)= 1/6

举例—计算过程

③ p（矮|嫁） = ?

统计满足样本数如下：

帅？	性格好？	身高？	上进？	嫁与否
帅	不好	矮	不上进	不嫁
不帅	好	矮	上进	不嫁
帅	好	矮	上进	嫁
不帅	好	高	上进	嫁
帅	不好	矮	上进	不嫁
帅	不好	矮	上进	不嫁
帅	好	高	不上进	嫁
不帅	好	中	上进	嫁
帅	好	中	上进	嫁
不帅	不好	高	上进	嫁
帅	好	矮	不上进	不嫁
帅	好	矮	不上进	不嫁

则p(矮|嫁) = 1/6

举例—计算过程

④ p(不上进|嫁) = ?

统计满足样本数如下:

帅？	性格好？	身高？	上进？	嫁与否
帅	好	高	不上进	嫁

则p(不上进|嫁) = 1/6

举例—计算过程

Step3: 下面开始求分母，p(不帅)、p（性格不好）、p（矮）、p（不上进）

统计样本如下：

帅？	性格好？	身高？	上进？	嫁与否
帅	不好	矮	不上进	不嫁
不帅	好	矮	上进	不嫁
帅	好	矮	上进	嫁
不帅	好	高	上进	嫁
帅	不好	矮	上进	不嫁
帅	不好	矮	上进	不嫁
帅	好	高	不上进	嫁
不帅	好	中	上进	嫁
帅	好	中	上进	嫁
不帅	不好	高	上进	嫁
帅	好	矮	不上进	不嫁
帅	好	矮	不上进	不嫁

则：p（不帅） = 4/12 = 1/3　　p（性格不好） = 4/12 = 1/3

p（身高矮） = 7/12　　p（不上进） = 4/12 = 1/3

举例—计算过程

结果：

$$p(\text{嫁}|\text{不帅、性格不好、身高矮、不上进})=\frac{p(\text{不帅、性格不好、身高矮、不上进}|\text{嫁})*p(\text{嫁})}{p(\text{不帅、性格不好、身高矮、不上进})}$$

$$=\frac{p(\text{不帅}|\text{嫁})*p(\text{性格不好}|\text{嫁})*p(\text{身高矮}|\text{嫁})*p(\text{不上进}|\text{嫁})*p(\text{嫁})}{p(\text{不帅})*p(\text{性格不好})*p(\text{身高矮})*p(\text{不上进})}$$

$$=(1/2*1/6*1/6*1/6*1/2)/(1/3*1/3*7/12*1/3)$$

* **同样的方法来求** p(不嫁|不帅，性格不好，身高矮，不上进)

贝叶斯决策论

- 贝叶斯决策论（Bayesian decision theory）是在概率框架下实施决策的基本方法。
  - 在分类问题情况下，在所有相关概率都已知的理想情形下，贝叶斯决策考虑如何基于这些概率和误判损失来选择最优的类别标记。

贝叶斯决策论

- 假设有 N 种可能的类别标记，即 $y = \{c_1, c_2, \ldots, c_N\}$ ，λ_{ij} 是将一个真实标记为 c_j 的样本误分类为 c_i 所产生的损失。基于后验概率 $P\{c_i \mid \mathbf{x}\}$ 可获得将样本 $\mathbf{x}$ 分类为 c_i 所产生的期望损失（expected loss），即在样本上的“条件风险”（conditional risk）

$$R(c_i \mid \mathbf{x}) = \sum_{j=1}^{N} \lambda_{ij} P(c_j \mid \mathbf{x}) \qquad (7.1)$$

- 我们的任务是寻找一个判定准则 $h : X \mapsto Y$ 以最小化总体风险

$$R(h) = \mathbf{E}_x \left[R(h(\mathbf{x}) \mid \mathbf{x}) \right] \qquad (7.2)$$

贝叶斯决策论

- 显然，对每个样本 $\mathbf{x}$，若 h 能最小化条件风险 $R(h(\mathbf{x}) \mid \mathbf{x})$，则总体风险 $R(h)$ 也将被最小化。

贝叶斯决策论

- 显然，对每个样本 $\mathbf{x}$，若 h 能最小化条件风险 $R(h(\mathbf{x}) \mid \mathbf{x})$ ，则总体风险 $R(h)$ 也将被最小化。
- 这就产生了贝叶斯判定准则（Bayes decision rule）： 为最小化总体风险，只需在每个样本上选择那个能使条件风险 $R(c \mid \mathbf{x})$ 最小的类别标记，即

$$h^*(x) = \underset{c \in y}{\operatorname{argmin}} R(c \mid x) \qquad (7.3)$$

  - 此时，被称为贝叶斯最优分类器(Bayes optimal classifier)，与之对应的总体风险 $R(h^*)$ 称为贝叶斯风险 (Bayes risk)
  - $1 - R(h^*)$ 反映了分类起所能达到的最好性能，即通过机器学习所能产生的模型精度的理论上限。

贝叶斯决策论

- 具体来说，若目标是最小化分类错误率，则误判损失 λ_{ij}可写为

$$\lambda_{i,j} \begin{cases} 0, & \text{if } i = j; \\ 1, & \text{otherwise}, \end{cases} \tag{7.4}$$

贝叶斯决策论

- 具体来说，若目标是最小化分类错误率，则误判损失 λ_{ij}可写为

$$\lambda_{i,j} \begin{cases} 0, & \text{if } i = j; \\ 1, & \text{otherwise}, \end{cases} \tag{7.4}$$

- 此时条件风险

$$R(c \mid \mathbf{x}) = 1 - P(c \mid \mathbf{x}) \tag{7.5}$$

贝叶斯决策论

- 具体来说，若目标是最小化分类错误率，则误判损失 λ_{ij}可写为

$$\lambda_{i,j} \begin{cases} 0, & \text{if } i = j; \\ 1, & \text{otherwise}, \end{cases} \tag{7.4}$$

- 此时条件风险

$$R(c \mid \mathbf{x}) = 1 - P(c \mid \mathbf{x}) \tag{7.5}$$

- 于是，最小化分类错误率的贝叶斯最优分类器为

$$h^*(x) = \underset{c \in y}{\operatorname{argmax}} P(c \mid x) \tag{7.6}$$

  - 即对每个样本 $\mathbf{x}$，选择能使后验概率$P(c \mid \mathbf{x})$ 最大的类别标记。

贝叶斯决策论

- 不难看出，使用贝叶斯判定准则来最小化决策风险，首先要获得后验概率 $P(c \mid \mathbf{x})$ 。
- 然而，在现实中通常难以直接获得。机器学习所要实现的是基于有限的训练样本尽可能准确地估计出后验概率 $P(c \mid \mathbf{x})$ 。
- 主要有两种策略：
  - 判别式模型（discriminative models）
    - 给定 $\mathbf{x}$，通过直接建模 $P(c \mid \mathbf{x})$，来预测 c
    - 决策树，BP神经网络，支持向量机
  - 生成式模型（generative models）
    - 先对联合概率分布 $P(\mathbf{x}, c)$ 建模，再由此获得 $P(c \mid \mathbf{x})$
    - 生成式模型考虑 $P(c \mid \mathbf{x}) = \dfrac{P(\mathbf{x}, c)}{P(\mathbf{x})}$ (7.7)

贝叶斯决策论

- 生成式模型

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x}, c)}{P(\mathbf{x})} \tag{7.7}$$

贝叶斯决策论

- 生成式模型

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x}, c)}{P(\mathbf{x})} \tag{7.7}$$

- 基于贝叶斯定理，$P(c \mid \mathbf{x})$ 可写成

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} \tag{7.8}$$

贝叶斯决策论

- 生成式模型

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x}, c)}{P(\mathbf{x})} \tag{7.7}$$

- 基于贝叶斯定理，$P(c \mid \mathbf{x})$ 可写成

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} \tag{7.8}$$

先验概率
样本空间中各类样本所占的比例，可通过各类样本出现的频率估计（大数定理）

贝叶斯决策论

- 生成式模型

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x}, c)}{P(\mathbf{x})} \tag{7.7}$$

- 基于贝叶斯定理，$P(c \mid \mathbf{x})$ 可写成

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} \tag{7.8}$$

先验概率
样本空间中各类样本所占的比例，可通过各类样本出现的频率估计（大数定理）

“证据” （evidence）因子，与类标记无关

贝叶斯决策论

- 生成式模型

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x}, c)}{P(\mathbf{x})} \tag{7.7}$$

- 基于贝叶斯定理，$P(c \mid \mathbf{x})$ 可写成

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} \tag{7.8}$$

类标记 c相对于样本$\mathbf{x}$的“类条件概率” (class-conditional probability), 或称“似然”。

先验概率
样本空间中各类样本所占的比例，可通过各类样本出现的频率估计（大数定理）

“证据” (evidence) 因子，与类标记无关

章节目录

极大似然估计

- 估计类条件概率的常用策略：先假定其具有某种确定的概率分布形式，再基于训练样本对概率分布参数估计。

- 记关于类别 c 的类条件概率为 $P(\mathbf{x} \mid c)$ ，
  - 假设 $P(\mathbf{x} \mid c)$ 具有确定的形式被参数 $\boldsymbol{\theta}_c$ 唯一确定，我们的任务就是利用训练集 D 估计参数 $\boldsymbol{\theta}_c$

极大似然估计

- 估计类条件概率的常用策略：先假定其具有某种确定的概率分布形式，再基于训练样本对概率分布参数估计。
- 记关于类别 C 的类条件概率为 $P(\mathbf{x} \mid c)$ ，
  - 假设 $P(\mathbf{x} \mid c)$ 具有确定的形式被参数 $\boldsymbol{\theta}_c$ 唯一确定，我们的任务就是利用训练集 D 估计参数 $\boldsymbol{\theta}_c$
- 概率模型的训练过程就是参数估计过程，统计学界的两个学派提供了不同的方案：
  - 频率主义学派 (Frequentist)认为参数虽然未知，但却存在客观值，因此可通过优化似然函数等准则来确定参数值。
  - 贝叶斯学派 (Bayesian)认为参数是未观察到的随机变量、其本身也可有分布，因此可假定参数服从一个先验分布，然后基于观测到的数据计算参数的后验分布。

极大似然估计

- 令 D_c 表示训练集中第 c 类样本的组合的集合，假设这些样本是独立的，则参数 $\boldsymbol{\theta}_c$ 对于数据集 D_c 的似然是

$$P(D_c \mid \boldsymbol{\theta_c}) = \prod_{\mathbf{x} \in D_c} P(\mathbf{x} \mid \boldsymbol{\theta_c}) \qquad (7.9)$$

  - 对 $\boldsymbol{\theta}_c$ 进行极大似然估计，寻找能最大化似然 $P(D_c \mid \theta_c)$ 的参数值 $\hat{\theta}_c$ 。直观上看，极大似然估计是试图在 θ_c 所有可能的取值中，找到一个使数据出现的“可能性”最大值。

极大似然估计

- 令 D_c 表示训练集中第 c 类样本的组合的集合，假设这些样本是独立的，则参数 $\boldsymbol{\theta}_c$ 对于数据集 D_c 的似然是

$$P(D_c \mid \boldsymbol{\theta_c}) = \prod_{\mathbf{x} \in D_c} P(\mathbf{x} \mid \boldsymbol{\theta_c}) \qquad (7.9)$$

  - 对 $\boldsymbol{\theta}_c$ 进行极大似然估计，寻找能最大化似然 $P(D_c \mid \theta_c)$ 的参数值 $\hat{\theta}_c$ 。直观上看，极大似然估计是试图在 θ_c 所有可能的取值中，找到一个使数据出现的“可能性”最大值。

- 式 (7.9)的连乘操作易造成下溢，通常使用对数似然(log-likelihood)

$$\begin{aligned} LL(\boldsymbol{\theta_c}) &= \log P(D_c \mid \boldsymbol{\theta_c}) \\ &= \sum_{\mathbf{x} \in D_c} \log P(\mathbf{x} \mid \boldsymbol{\theta}_c) \end{aligned} \qquad (7.10)$$

- 此时参数 $\boldsymbol{\theta}_c$ 的极大似然估计 $\hat{\boldsymbol{\theta}}_c$ 为 $\hat{\boldsymbol{\theta}}_{\boldsymbol{c}} = \underset{\boldsymbol{\theta}_c}{\operatorname{argmax}} LL(\boldsymbol{\theta}_c)$ (7.11)

举个例子

问题：现在需要调查西北大学信科院学生的身高分布，如何进行？

- 统计每一个学生的身高（效率太低）
- 使用概率统计的思想，即抽样，根据样本估计总体。

假设信科院所有学生的身高服从正态分布$N(\mu,\sigma^2)$，此分布的两个参数，均值μ和标准差σ未知。假设随机抽到了200个人的身高样本数据，利用这200个人的身高来估计均值μ和标准差σ。

数学语言描述：

为了统计西北大学信科院学生的身高分布，按照概率密度$p(x|\theta)$，即$N(\mu,\sigma^2)$，抽取了200个人的身高，得到样本集$X=\{x_1,x_2,\dots,x_N\}$，其中，x_i 表示抽到的第 i 个人的身高，N表示样本个数。通过样本集X来估计总体分布的未知参数$\theta=[\mu,\sigma]^T$。

问题来了：如何估计参数$\theta=[\mu,\sigma]^T$？

举个例子

问题一：抽到这200个人的身高的概率是多少？

$$L(\theta) = L(x_1, x_2, \ldots, x_N; \theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

为什么取对数？

函数$L(\theta)$为参数θ对于样本集X的似然函数（Likelihood Function），记为$L(\theta)$。对$L(\theta)$取对数，得到对数似然函数，如下式：

$$H(\theta) = \ln L(\theta) = \ln \prod_{i=1}^{n} p(x_i|\theta) = \sum_{i=1}^{n} \ln p(x_i|\theta)$$

问题二：信科院那么多学生，为什么就恰好抽到了这200个人的身高呢？

$$\hat{\theta} = argmax\, L(\theta)$$

其中，$\hat{\theta}$为θ的极大似然估计值。

问题三：如何求极大似然函数？

求偏导数，让这些偏导数等于0，假设有n个参数，就有n个方程组成的方程组，求解方程组就能得到$\hat{\theta}$。

举个例子

所以

极大似然估计，是参数估计的方法之一。说的是已知某个随机样本满足某种概率分布，但是其中具体的参数不清楚，参数估计就是通过若干次试验，观察其结果，利用结果推出参数的大概值。极大似然估计是建立在这样的思想上：已知某个参数能使这个样本出现的概率最大，我们当然不会再去选择其他小概率的样本，所以干脆就把这个参数作为估计的真实值。

求极大似然函数估计值的一般步骤：

（1） 写出似然函数；

（2） 对似然函数取对数，并整理；

（3） 求导数；

（4） 解似然方程 。（注：有可能解不出方程组，得不到解析解）

极大似然估计

- 例如，在连续属性情形下，假设概率密度函数 $p(\mathbf{x} \mid c) \sim N(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)$，则参数 $\boldsymbol{\mu}_c$ 和 $\boldsymbol{\sigma}_c^2$ 的极大似然估计为

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{|D_c|} \sum_{\mathbf{x} \in D_c} \mathbf{x} \tag{7.12}$$

$$\hat{\boldsymbol{\sigma}}_c^2 = \frac{1}{|D_c|} \sum_{\mathbf{x} \in D_c} (\mathbf{x} - \hat{\boldsymbol{\mu}}_c)(\mathbf{x} - \hat{\boldsymbol{\mu}}_c)^{\mathrm{T}} \tag{7.13}$$

- 也就是说，通过极大似然法得到的正态分布均值就是样本均值，方差就是 $(\mathbf{x} - \hat{\boldsymbol{\mu}}_c)(\mathbf{x} - \hat{\boldsymbol{\mu}}_c)^{\mathrm{T}}$ 的均值，这显然是一个符合直觉的结果。
- 需注意的是，这种参数化的方法虽能使类条件概率估计变得相对简单，但估计结果的准确性严重依赖于所假设的概率分布形式是否符合潜在的真实数据分布。

章节目录

朴素贝叶斯分类器

- 估计后验概率 $P(c \mid \mathbf{x})$ 主要困难：类条件概率 $P(\mathbf{x} \mid c)$ 是所有属性上的联合概率难以从有限的训练样本估计获得。

- 朴素贝叶斯分类器(Naïve Bayes Classifier)采用了“属性条件独立性假设”(attribute conditional independence assumption)：每个属性独立地对分类结果发生影响。

- 基于属性条件独立性假设，(7.8)可重写为

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^{d} P(x_i \mid c) \qquad (7.14)$$

  - 其中 d 为属性数目，x_i 为 $\mathbf{x}$ 在第 i 个属性上的取值。

朴素贝叶斯分类器

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^{d} P(x_i \mid c) \qquad (7.14)$$

朴素贝叶斯分类器

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^{d} P(x_i \mid c) \qquad (7.14)$$

由于对所有类别来说 $P(x)$ 相同，因此基于式 (7.6)的贝叶斯判定准则有

$$h_{nb}(\mathbf{x}) = \underset{c \in y}{\operatorname{argmax}} P(c) \prod_{i=1}^{d} P(x_i \mid c) \qquad (7.15)$$

- 这就是朴素贝叶斯分类器的表达式子

朴素贝叶斯分类器

- 朴素贝叶斯分类器的训练器的训练过程就是基于训练集 D 估计类先验概率 $P(c)$ 并为每个属性估计条件概率 $P(x_i \mid c)$ 。
  - 令 D_c 表示训练集 D 中第 c 类样本组合的集合，若有充足的独立同分布样本，则可容易地估计出类先验概率

$$P(c) = \frac{|D_c|}{D} \qquad (7.16)$$

  - 对离散属性而言，令 D_{c,x_i} 表示 D_c 中在第 i 个属性上取值为 x_i 的样本组成的集合，则条件概率 $P(x_i \mid c)$ 可估计为

$$P(x_i \mid c) = \frac{|D_{c,x_i}|}{D} \qquad (7.17)$$

  - 对连续属性而言可考虑概率密度函数，假定 $p(x_i \mid c) \sim N(\mu_{c,i}, \sigma_{c,i}^2)$ ，其中 $\mu_{c,i}$ 和 $\sigma_{c,i}^2$ 分别是第 c 类样本在第 i 个属性上取值的均值和方差，则有

$$P(x_i \mid c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2{\sigma_{c,i}}^2}\right) \qquad (7.18)$$

朴素贝叶斯分类器

- 例子：用西瓜数据集3.0训练一个朴素贝叶斯分类器，对测试例“测1”进行分类

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

拉普拉斯修正

- 若某个属性值在训练集中没有与某个类同时出现过，则直接计算会出现问题，比如“敲声=清脆”测试例，训练集中没有该样例，因此连乘式计算的概率值为0，无论其他属性上明显像好瓜，分类结果都是“好瓜=否”，这显然并不合理。

拉普拉斯修正

- 为了避免其他属性携带的信息被训练集中未出现的属性值“抹去”，在估计概率值时通常要进行“拉普拉斯修正”（Laplacian correction）
  - 令 N 表示训练集 D 中可能的类别数，N_i 表示第 i 个属性可能的取值数，则式 (7.16)和 (7.17)分别修正为

$$\hat{P}(c) = \frac{|D_c| + 1}{|D| + N} \quad (7.19) \qquad \hat{P}(x_i \mid c) = \frac{|D_{c,x_i}| + 1}{|D| + N_i} \quad (7.20)$$

- 现实任务中，朴素贝叶斯分类器的使用：速度要求高，“查表”；任务数据更替频繁，“懒惰学习”（lazy learning）；数据不断增加，增量学习等等。

章节目录

半朴素贝叶斯分类器

- 为了降低贝叶斯公式中估计后验概率的困难，朴素贝叶斯分类器采用的属性条件独立性假设；对属性条件独立假设进行一定程度的放松，由此产生了一类称为“半朴素贝叶斯分类器” (semi-naïve Bayes classifiers)

半朴素贝叶斯分类器

- 半朴素贝叶斯分类器最常用的一种策略：“独依赖估计”（One-Dependent Estimator,简称ODE），假设每个属性在类别之外最多仅依赖一个其他属性，即

$$P(c \mid x) \propto P(c) \prod_{i=1}^{d} P(x_i \mid c, pa_i)$$

  - 其中 pa_i 为属性 x_i 所依赖的属性，称为 x_i 的父属性

- 对每个属性 x_i ，若其父属性 pa_i 已知，则可估计概值 $P(x_i \mid c, pa_i)$ ，于是问题的关键转化为如何确定每个属性的父属性

SPODE

- 最直接的做法是假设所有属性都依赖于同一属性，称为“超父”(super-parent)，然后通过交叉验证等模型选择方法来确定超父属性，由此形成了SPODE (Super-Parent ODE)方法。

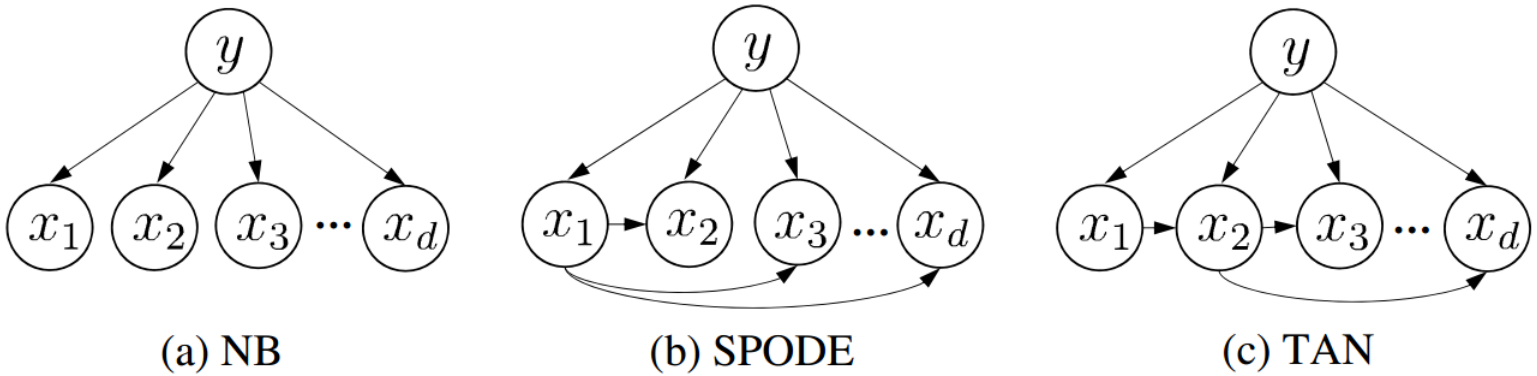


图7.1 朴素贝叶斯分类器与两种半朴素分类器所考虑的属性依赖关系

- 在图7.1 (b)中，x_1 是超父属性。

TAN

- TAN (Tree augmented Naïve Bayes) [Friedman et al., 1997] 则在最大带权生成树 (Maximum weighted spanning tree) 算法 [Chow and Liu, 1968] 的基础上，通过以下步骤将属性间依赖关系简约为图7.1 (c)。

  - 计算任意两个属性之间的条件互信息 (conditional mutual information)

$$I(x_i, x_j \mid y) = \sum_{x_i, x_j; c \in y} P(x_i, x_j \mid c) \log \frac{P(x_i, x_j \mid c)}{P(x_i \mid c) P(x_j \mid c)}$$

  - 以属性为结点构建完全图，任意两个结点之间边的权重设为 $I(x_i, x_j \mid y)$
  - 构建此完全图的最大带权生成树，挑选根变量，将边设为有向；
  - 加入类别节点y，增加从y到每个属性的有向边。

AODE

- AODE (Averaged One-Dependent Estimator) [Webb et al. 2005] 是一种基于集成学习机制、更为强大的分类器。
  - 尝试将每个属性作为超父构建 SPODE
  - 将具有足够训练数据支撑的SPODE集群起来作为最终结果

$$P(c \mid \mathbf{x}) \propto \sum_{i=1;|D_{x_i} \geq m'|}^{d} P(c, x_i) \prod_{j=1}^{d} P(x_j \mid c, x_i)$$

其中，D_{x_i} 是在第 i 个属性上取值 x_i 的样本的集合，m' 为阈值常数

$$\hat{P}(x_i, c) = \frac{|D_{c,x_i}| + 1}{|D| + N_i} \qquad \hat{P}(x_j \mid c, x_i) = \frac{|D_{c,x_i,x_j}| + 1}{|D_{c,x_i}| + N_j}$$

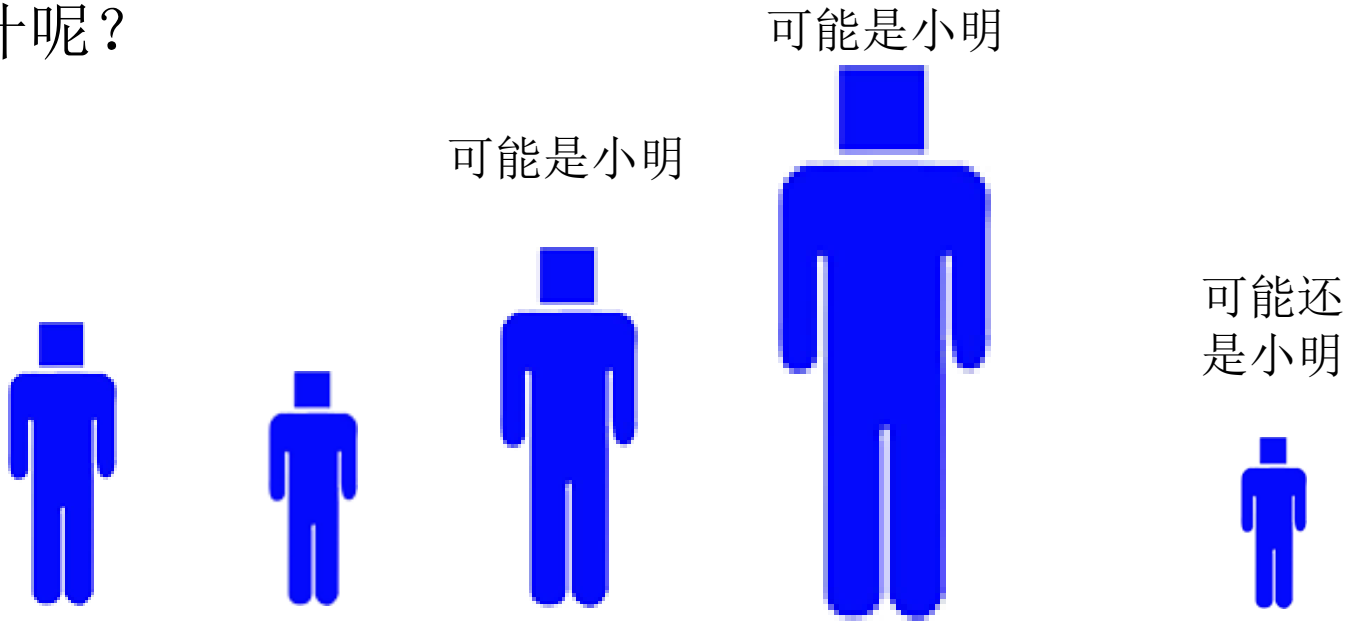
其中，N_i 是在第 i 个属性上取值数，D_{c,x_i} 是类别为 c 且在第 i 个属性上取值为 x_i 的样本集合，

章节目录

章节目录

EM算法

- 男生和女生分别服从两种不同的正态分布，即男生$\in N(\mu_1, \sigma_1^2)$，女生$\in N(\mu_2, \sigma_2^2)$，现在该如何评估学生的身高分布呢？
- 随机抽取100个男生和100个女生，分别进行极大似然估计，分别求出男生和女生的分布。
- 假如在抽取过程中，没有记录每一个同学是男生还是女生，对于每一位同学，只有他们身高数据，没有性别信息。又该如何进行参数估计呢？



EM算法

X: 观测数据 $X = (x_1, x_2, x_3 \ldots\ldots x_N)$

θ:参数 $\theta = \{p_1, p_2 \ldots\ldots p_k, \mu_1, \mu_2 \ldots \mu_k, \sigma_1, \sigma_2 \ldots \sigma_k\}$

$$\hat{\theta}_{MLE} = \arg\max_{\theta}(\log p(X)) = \arg\max_{\theta} \log \prod_{i=1}^{N} p(x_i)$$

$$= \arg\max_{\theta} \sum_{i=1}^{N} \log p(x_i)$$

$$= \arg\max_{\theta} \sum_{i=1}^{N} \log \sum_{k=1}^{K} p_k \times N(X \mid \mu_k, \sigma_k)$$

有可能是高维高斯分布，还乘上概率，还连加，表达式太复杂。得不到解析解，所以引出了EM方法。

EM算法

诗词解读

混高辞

作者：李某某

高者是小明，矮者亦小明；

小明一起走，安能辨我是雄雌？

对于抽取到的每一个身高样本，无法知道它是来自男生的，还是来自女生的，因此，无法分别对男女生的身高样本进行极大似然估计，从而估计出两个分布的参数。

世纪难题：先有鸡还是先有蛋

对于每一个样本，只有当我们对这两个分布的参数作出了准确的估计的时候，才能知道到底哪些人是男生是女生。

混高辞 II

作者：李李某某

问女何所 Σ，问女何所 μ，女亦无所 Σ，女亦无所 μ。

阿明无参数，小明无分布，愿用EM，从此替爷征。

EM算法

EM(Expectation Maximization)方法解决：

- 先设定男生和女生的身高分布参数(初始值)，例如男生的身高分布为$N(\mu_1 = 172, \sigma_1^2 = 5^2)$， 女生的身高分布为$N(\mu_2 = 162, \sigma_2^2 = 5^2)$，当然了，刚开始肯定没那么准；
- 然后计算出每个人更可能属于第一个还是第二个正态分布中的（例如，这个人的身高是180，那很明显，他极大可能属于男生），这个是属于E步；
- 按上面的方法将这 200 个人分为男生和女生两部分，接着根据极大似然估计分别对男生和女生的身高分布参数进行估计，这个是属于M步；
- 经过M步，这两个分布的参数得到更新，每一个样本来自女生分布还是男生分布的概率又变了，那么需要调整E步；

……如此往复，直到参数基本不再发生变化或满足结束条件为止。

EM算法

EM吟

作者：李某某

每个分布赋初值，不断迭代去更新。
往复计算浑不怕，要留逼近在人间。

- EM算法是一种迭代算法，用于含有隐变量的概率模型参数的极大似然估计，或极大后验概率估计。

- EM算法的每次迭代由两步组成：E步，求期望（expectation）；M步，求极大（maximization）。

EM算法

- “不完整”的样本：西瓜已经脱落的根蒂，无法看出是“蜷缩”还是“坚挺”,则训练样本的“根蒂”属性变量值未知，如何计算?

- 未观测的变量称为“隐变量”(latent variable)。令 $\mathbf{X}$ 表示已观测变量集，$\mathbf{Z}$ 表示隐变量集，若预对模型参数 Θ 做极大似然估计，则应最大化对数似然函数

$$LL(\boldsymbol{\Theta} \mid \mathbf{X}, \mathbf{Z}) = \ln P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta}) \quad (7.34)$$

EM算法

- 未观测的变量称为“隐变量”(latent variable)。令 $\mathbf{X}$表示已观测变量集，$\mathbf{Z}$ 表示隐变量集，若预对模型参数 Θ做极大似然估计，则应最大化对数似然函数

$$LL(\boldsymbol{\Theta} \mid \mathbf{X}, \mathbf{Z}) = \ln P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta}) \tag{7.34}$$

- 由于$\mathbf{Z}$是隐变量，上式无法直接求解。此时我们可以通过对 $\mathbf{Z}$ 计算期望，来最大化已观测数据的对数“边际似然”(marginal likelihood)

$$LL(\boldsymbol{\Theta} \mid \mathbf{X}) = \ln P(\mathbf{X} \mid \boldsymbol{\Theta}) = \ln \sum_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta}) \tag{7.35}$$

EM算法

EM (Expectation-Maximization)算法 [Dempster et al., 1977]是常用的估计参数隐变量的利器。

- 当参数Θ已知 - > 根据训练数据推断出最优隐变量 $\mathbf{Z}$ 的值（**E步**）
- 当$\mathbf{Z}$ 已知 - > 对Θ 做极大似然估计(M步)

于是，以初始值Θ^0 为起点，对式子(7.35),可迭代执行以下步骤直至收敛:

- 基于 Θ^t 推断隐变量 $\mathbf{Z}$ 的期望,记为 $\mathbf{Z}^t$;
- 基于已观测到变量 $\mathbf{X}$ 和$\mathbf{Z}^t$对参数 Θ 做极大似然估计，记为 Θ^{t+1} ;
- 这就是EM算法的原型。

EM算法

进一步，若我们不是取Z的期望，而是基于 Θ^t 计算隐变量 $\mathbf{Z}$ 的概率分布 $P(\mathbf{Z} \mid \mathbf{X}, \Theta^t)$ EM算法的两个步骤是：

- E步(Expectation):以当前参数 Θ^t 推断隐变量分布 $P(\mathbf{Z} \mid \mathbf{X}, \Theta^t)$ ，并计算对数似然 $LL(\Theta \mid \mathbf{X}, \mathbf{Z})$ 关于 $\mathbf{Z}$ 的期望:

$$Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^t) = \mathbb{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\Theta}^t} LL(\boldsymbol{\Theta} \mid \mathbf{X}, \mathbf{Z}) \qquad (7.36)$$

- M步（Maximization）：寻找参数最大化期望似然，即

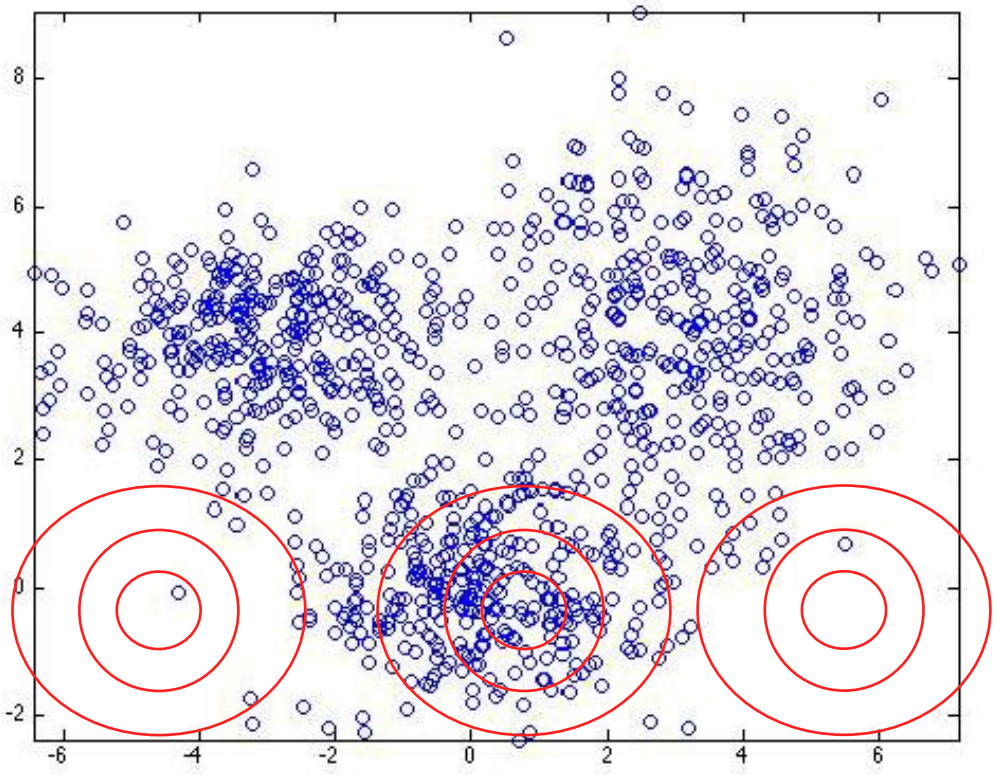
$$\boldsymbol{\Theta}^{t+1} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^t) \qquad (7.37)$$

- EM算法使用两个步骤交替计算：第一步计算期望(E步)，利用当前估计的参数值计算对数似然的参数值；第二步最大化(M步)，寻找能使E步产生的似然期望最大化的参数值……直至收敛到全局最优解。

EM算法

迭代过程总览

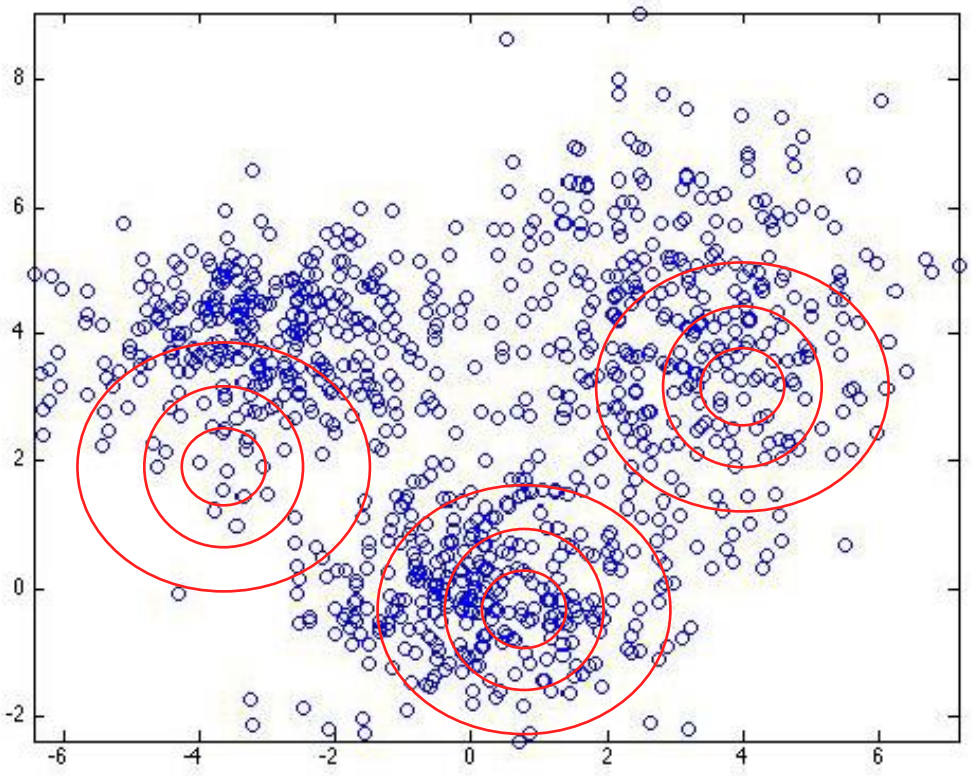
θ_1



EM算法

迭代过程总览

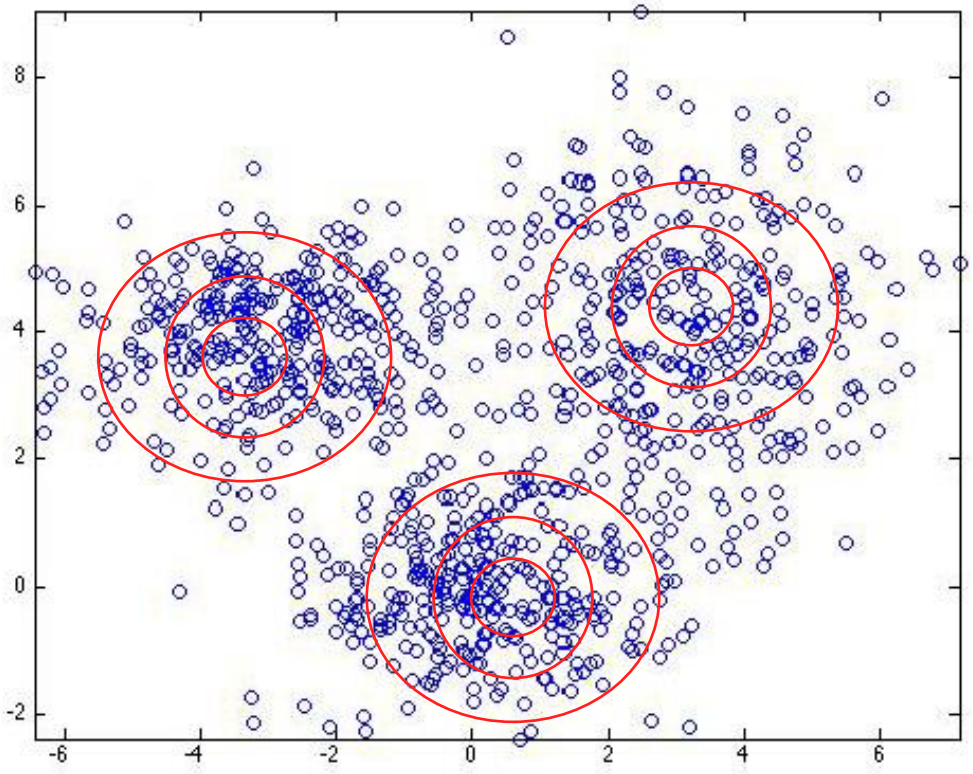
θ_2



EM算法

迭代过程总览

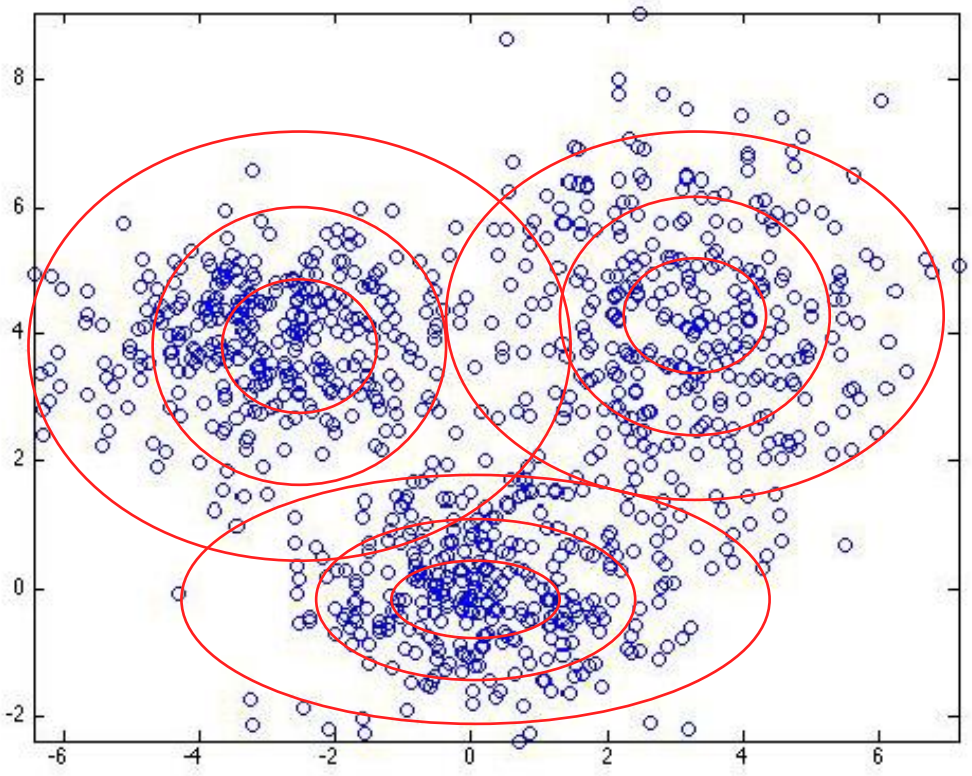
θ_3



EM算法

迭代过程总览

θ_4



小结

- 贝叶斯决策论
- 极大似然估计
- 朴素贝叶斯分类器
- 半朴素贝叶斯分类器
- 贝叶斯网
- EM算法