



人工智能

ARTIFICIAL INTELLIGENCE

主讲：鲍军鹏 博士

西安交通大学电信学院计算机系

电子邮箱：dr.baojp@googlemail.com



版本：2.0

2010年1月

6.3 贝叶斯学习

- × 6.3.1 贝叶斯法则
- × 6.3.2 朴素贝叶斯方法
- × 6.3.3 贝叶斯网络
- × 6.3.4 EM算法
- × 6.3.5 用贝叶斯方法过滤垃圾邮件

6.3.1 贝叶斯法则

✖ 贝叶斯学习

- + 就是基于贝叶斯理论 (Bayesian Theory) 的机器学习方法。

✖ 贝叶斯法则

- + 也称为贝叶斯理论 (Bayesian Theorem, 或 Bayesian Rule, 或 Bayesian Law), 其核心就是贝叶斯公式。

贝叶斯公式

后验概率

先验概率

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

先验概率

先验概率 (Prior Probability)

- + 先验概率就是还没有训练数据之前，某个假设 h ($h \in H$) 的初始概率，记为 $P(h)$ 。
- + 先验概率反映了一个背景知识，表示 h 是一个正确假设的可能性有多少。
- + 如果没有这一先验知识，那么可以简单地将每一候选假设赋予相同的先验概率。

似然度

- + $P(d)$ 表示训练数据 d 的先验概率，也就是在任何假设都未知或不确定时 d 的概率。
- + $P(d|h)$ 表示已知假设 h 成立时 d 的概率，称之为类条件概率，或者给定假设 h 时数据 d 的**似然度**（Likelihood）。

后验概率

后验概率 (Posterior Probability)

- + 后验概率就是在数据d上经过学习之后，获得的假设h成立的概率，记为 $P(h|d)$ 。
- + $P(h|d)$ 表示给定数据d时假设h成立的概率，称为h的后验概率。

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

- + 后验概率是学习的结果，反映了在看到训练数据 d 之后，假设 h 成立的置信度。
- + 后验概率用作解决问题时的依据。
- + 对于给定数据根据该概率做出相应决策，例如判断数据的类别，或得出某种结论，或执行某种行动等等。

- + $P(h|d)$ 随着 $P(h)$ 和 $P(d|h)$ 的增长而增长，随着 $P(d)$ 的增长而减少。
- + 即如果 d 独立于 h 时被观察到的可能性越大，那么 d 对 h 的支持度越小。
- + 后验概率是对先验概率的修正。

- ✗ 后验概率 $P(h|d)$ 是在数据 d 上得到的学习结果，反映了数据 d 的影响。这个学习结果是和训练数据相关的。
- ✗ 与此相反，先验概率 $P(h)$ 是与训练数据 d 无关的，是独立于 d 的。

注意!

- ✖ 贝叶斯法则解决的机器学习任务一般是：
 - + 在给定训练数据 D 时，确定假设空间 H 中的最优假设。这是典型的分类问题。
- ✖ 贝叶斯法则基于假设的先验概率、给定假设下观察到不同数据的概率以及观察到的数据本身，提供了一种计算假设概率的方法。

贝叶斯最优假设

- ✗ 分类问题的最优假设（即最优结果），可以有不同定义。
 - + 例如，与期望误差最小的假设；或者能取得最小熵（Entropy）的假设等等。
- ✗ 贝叶斯分类器是指为在给定数据 d 、假设空间 H 中不同假设的先验概率以及有关知识下的最可能假设。
 - + 这个最可能假设可有不同选择。

极大后验假设

(1) 极大后验假设

(Maximum A Posteriori, 简称MAP假设)

极大后验假设 h_{MAP} ($h_{MAP} \in H$) 就是在候选假设集合 H 中寻找对于给定数据 d 使后验概率 $P(h|d)$ 最大的那个假设。

极大后验假设

$$h_{MAP} \equiv \arg \max_{h \in H} P(h | d)$$

$$= \arg \max_{h \in H} \frac{P(d | h)P(h)}{P(d)}$$

$$= \arg \max_{h \in H} P(d | h)P(h)$$

不依赖于
h的常
量

极大似然假设

(2) 极大似然假设

(Maximum Likelihood, 简称ML假设)

极大似然假设就是在候选假设集合H中选择使给定数据d似然度（即类条件概率） $P(d|h)$ 最大的假设，即ML假设 h_{ML} ($h_{ML} \in H$) 是满足下式的假设。

$$h_{ML} \equiv \arg \max_{h \in H} P(d | h)$$

- + 极大似然假设和极大后验假设有很强的关联性。
- + 由于数据似然度是先验知识，不需要训练就能知道。所以在机器学习实践中经常应用极大似然假设来指导学习。

贝叶斯最优分类器

(3) 贝叶斯最优分类器

(Bayes Optimal Classifier)

- + 贝叶斯最优分类器是对最大后验假设的发展。它并不是简单地直接选取后验概率最大的假设（模型）作为分类依据。
- + 而是对所有假设（模型）的后验概率做线性组合（加权求和），然后再选择加权和最大结果作为最优分类结果。

贝叶斯最优分类器

- + 设 V 表示类别集合，对于 V 中的任意一个类别 v_j ，概率 $P(v_j | d)$ 表示把数据 d 归为类别 v_j 的概率。
- + 贝叶斯最优分类就是使 $P(v_j | d)$ 最大的那个类别。贝叶斯最优分类器就是满足下式的分类系统。

$$\arg \max_{v_j \in V} \sum_{h_i \in H} P(v_j | h_i) P(h_i | d)$$

- ✘ 在相同的假设空间和相同的先验概率条件下，其它方法的平均性能不会比贝叶斯最优分类器更好。
- ✘ 虽然贝叶斯最优分类器能从给定训练数据中获得最好性能，但是其算法开销比较大。

注意!

贝叶斯分类器示例

例. 设对于数据 d 有假设 h_1 , h_2 , h_3 。它们的先验概率分别是

$$P(h_1)=0.3, P(h_2)=0.3, P(h_3)=0.4。$$

并且已知

$$P(d|h_1)=0.5, P(d|h_2)=0.3, P(d|h_3)=0.2。$$

又已知在分类集合 $V=\{+, -\}$ 上数据 d 被 h_1 分类为正, 被 h_2 和 h_3 分类为负。请分别依据MAP假设和贝叶斯最优分类器对数据 d 进行分类。

贝叶斯分类器示例

解：先分别计算出假设 h_1 , h_2 , h_3 的后验概率如下。

最优假设

$$P(h_1 | d) = \frac{P(d | h_1)P(h_1)}{P(d)} = \frac{0.5 \times 0.3}{0.5 \times 0.3 + 0.3 \times 0.3 + 0.2 \times 0.4} \approx 0.47$$

$$P(h_2 | d) = \frac{P(d | h_2)P(h_2)}{P(d)} = \frac{0.3 \times 0.3}{0.5 \times 0.3 + 0.3 \times 0.3 + 0.2 \times 0.4} \approx 0.28$$

$$P(h_3 | d) = \frac{P(d | h_3)P(h_3)}{P(d)} = \frac{0.2 \times 0.4}{0.5 \times 0.3 + 0.3 \times 0.3 + 0.2 \times 0.4} = 0.25$$

那么依据MAP假设, h_1 是最优假设, 所以数据 d 应分类为正。

贝叶斯分类器示例

对于贝叶斯最优分类器，再计算分类概率如下。

$$\sum_{h_i \in H} P(+ | h_i) P(h_i | d)$$

$$= P(+ | h_1) P(h_1 | d) + P(+ | h_2) P(h_2 | d) + P(+ | h_3) P(h_3 | d)$$

$$= 1 \times 0.47 + 0 \times 0.28 + 0 \times 0.25$$

$$= 0.47$$

$$\sum_{h_i \in H} P(- | h_i) P(h_i | d)$$

$$= P(- | h_1) P(h_1 | d) + P(- | h_2) P(h_2 | d) + P(- | h_3) P(h_3 | d)$$

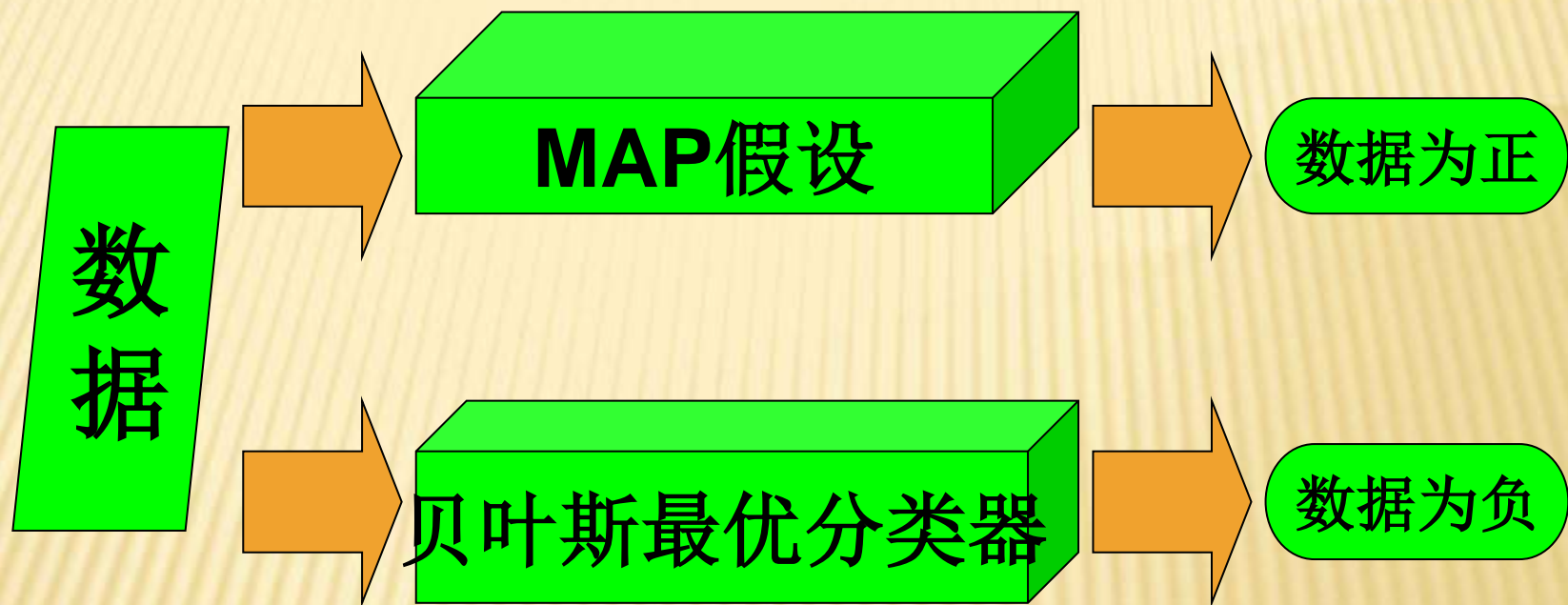
$$= 0 \times 0.47 + 1 \times 0.28 + 1 \times 0.25$$

$$= 0.53$$

$$0.53 > 0.47$$

那么依据贝叶斯最优分类器，数据d应该分类为**负**。

贝叶斯分类器



不同的方法结果不同!

贝叶斯学习的特点

- ✗ 贝叶斯学习为衡量多个假设的置信度提供了定量的方法，可以计算每个假设的显式概率，提供了一个客观的选择标准。
- ✗ 特性
 - + 观察到的每个训练样例可以增量地降低或升高某假设的估计概率。
 - + 先验知识可以与观察数据一起决定假设的最终概率。
 - + 允许假设做出不确定性的预测。例如前方目标是骆驼的可能性是90%，是马的可能性是5%。
 - + 新的实例分类可由多个假设一起做出预测，用它们的概率来加权。
 - + 即使在贝叶斯方法计算复杂度较高时，它仍可作为一个最优决策标准去衡量其它方法。

6.3.2 朴素贝叶斯方法

✕ 在机器学习中一个实例 x 往往有很多属性

$$\langle a_1, a_2, \dots, a_n \rangle$$

+ 其中每一维代表一个属性，该分量的数值就是所对应属性的值。

- ✘ 此时依据MAP假设的贝叶斯学习就是对一个数据 $\langle a_1, a_2, \dots, a_n \rangle$ ，求使其满足下式的目标值。其中 H 是目标值集合（样本类别的集合）。

$$\begin{aligned} h_{MAP} &= \arg \max_{h_i \in H} P(h_i | a_1, a_2, \dots, a_n) \\ &= \arg \max_{h_i \in H} \frac{P(a_1, a_2, \dots, a_n | h_i) P(h_i)}{P(a_1, a_2, \dots, a_n)} \\ &= \arg \max_{h_i \in H} P(a_1, a_2, \dots, a_n | h_i) P(h_i) \end{aligned}$$

- ✘ 估计每个 $P(h_i)$ 很容易，只要计算每个目标值 h_i 出现在训练数据中的频率就可以。
- ✘ 如果要如此估计所有的 $P(a_1, a_2, \dots, a_n | h_i)$ 项，则必须计算 a_1, a_2, \dots, a_n 的所有可能取值组合，再乘以可能的目标值数量。

- ✗ 假设一个实例有10个属性，每个属性有3个可能取值，而目标集合中有5个候选目标。那么 $P(a_1, a_2, \dots, a_n | h_i)$ 项就有 5×3^{10} 个。



不适合于高维数据！

朴素贝叶斯方法

- ✗ 朴素贝叶斯分类器采用最简单的假设：
 - + 对于目标值，数据各属性之间相互条件独立。
 - + 即， a_1, a_2, \dots, a_n 的联合概率等于每个单独属性的概率乘积：

$$P(a_1, a_2, \dots, a_n \mid h_i) = \prod_j P(a_j \mid h_i)$$

朴素贝叶斯方法

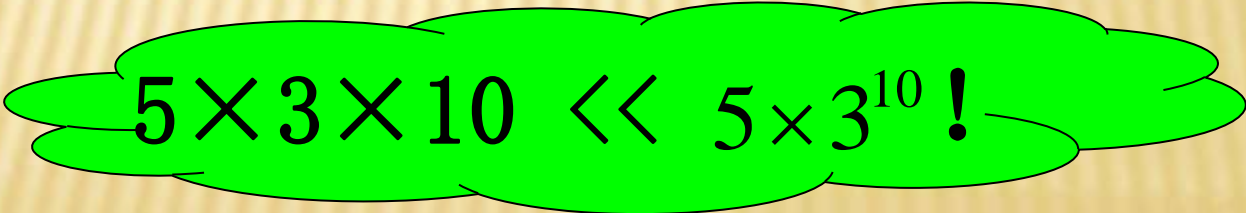
- ✗ 将上页的式子带入上面求 h_{MAP} 的公式中，就得到朴素贝叶斯分类器所用的方法：

$$h_{NB} = \arg \max_{h_i \in H} P(h_i) \prod_j P(a_j | h_i)$$

- ✗ 其中 h_{NB} 表示朴素贝叶斯分类器输出的目标值。

- ✗ 仍假设一个实例有10个属性，每个属性有3个可能取值，而目标集合中有5个候选目标。朴素贝叶斯分类器中需要从训练数据中估计的 $P(a_j|h_i)$ 项的数量是 $5 \times 3 \times 10$ 。

•
•
•


$$5 \times 3 \times 10 \ll 5 \times 3^{10} !$$

- ✘ 朴素贝叶斯学习的主要过程在于计算训练样例中不同数据组合的出现频率，统计出 $P(h_i)$ 和 $P(a_j|h_i)$ 。
- ✘ 算法比较简单，是一种很有效的机器学习方法。

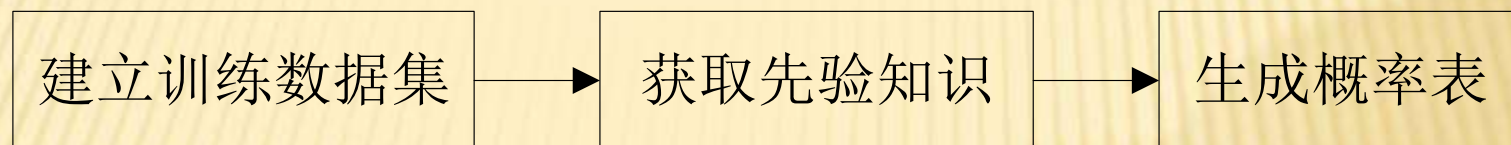
- ✘ 当各属性条件独立性满足时，朴素贝叶斯分类结果等于MAP分类。
- ✘ 这一假定一定程度上限制了朴素贝叶斯方法的适用范围。
- ✘ 但是在实际应用中，许多领域在违背这种假定的条件下，朴素贝叶斯学习也表现出**相当的健壮性和高效性**。

6.3 贝叶斯学习

- ✖ 6.3.1 贝叶斯法则
- ✖ 6.3.2 朴素贝叶斯方法
- ✖ 6.3.3 贝叶斯网络
- ✖ 6.3.4 EM算法
- ✖ 6.3.5 用贝叶斯方法过滤垃圾邮件

6.3.5 用贝叶斯方法过滤垃圾邮件

✘ 朴素贝叶斯方法的学习过程

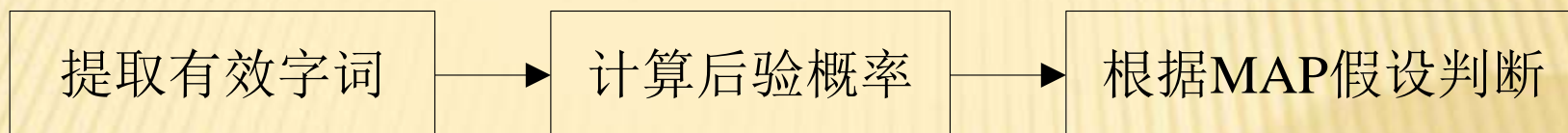


- + 收集大量垃圾邮件和非垃圾邮件，建立垃圾邮件集和非垃圾邮件集。
- + 提取邮件主题和邮件内容中的有效字词 w_i ，例如“内幕”、“真相”等等。然后统计其出现次数，即在该训练集上的词频 $TF(w_i)$ 。
- + 对垃圾邮件集和非垃圾邮件集中所有邮件执行第二步。
- + 对垃圾邮件集和非垃圾邮件集分别建立哈希表 W_{spam} 和 W_{valid} ，存储从有效字词到其词频的映射关系。
- + 计算每个有效字词在垃圾邮件集 (W_{spam}) 上出现的概率 $P(w_i | C=spam)$ 和在非垃圾邮件集 (W_{valid}) 上出现的概率 $P(w_i | C=valid)$

$$P(w_i | C) = \frac{TF(w_i)}{\sum_{w_i \in W} TF(w_i)}$$

- + 在垃圾邮件集和非垃圾邮件集上的学习过程结束，获得在垃圾邮件集和非垃圾邮件集上每个有效字词的出現概率。

✗ 用朴素贝叶斯方法判定一封邮件的过程



- + 对于一封邮件提取其所有有效字词 t_1, t_2, \dots, t_n 。
- + 从哈希表 W_{spam} 和 W_{valid} 中分别提取不同类别中上述有效字词的的概率 $P(t_i | C=\text{spam})$ 和 $P(t_i | C=\text{valid})$ 。
- + 依据朴素贝叶斯方法计算该邮件为垃圾邮件的概率 $P(C=\text{spam} | t_1, t_2, \dots, t_n)$ 和为非垃圾邮件的概率 $P(C=\text{valid} | t_1, t_2, \dots, t_n)$

$$\begin{aligned} &P(C = \text{spam} | t_1, t_2, \dots, t_n) \\ &= \frac{P(C = \text{spam}) \times P(t_1 | C = \text{spam}) \times P(t_2 | C = \text{spam}) \times \dots \times P(t_n | C = \text{spam})}{P(t_1, t_2, \dots, t_n)} \end{aligned}$$

$$\begin{aligned} &P(C = \text{valid} | t_1, t_2, \dots, t_n) \\ &= \frac{P(C = \text{valid}) \times P(t_1 | C = \text{valid}) \times P(t_2 | C = \text{valid}) \times \dots \times P(t_n | C = \text{valid})}{P(t_1, t_2, \dots, t_n)} \end{aligned}$$

- + 如果 $P(C=\text{spam} | t_1, t_2, \dots, t_n) > P(C=\text{valid} | t_1, t_2, \dots, t_n)$ 则该邮件为垃圾邮件，否则该邮件不是垃圾邮件。判定过程结束。

M-估计方法

× 问题

+ 某个词频为0的时候，实际概率不应该为0

× 思想：

+ 把原先n个实际观察扩大，加上m个按照p分布的虚拟样本。

$$P(w_i | C) = \frac{TF(w_i) + mp}{\sum_{w_i \in W} TF(w_i) + m}$$

+ 其中p是先验估计概率。

+ m是一个表示等效样本大小的常量。

× 估计p最常用的方法就是假定均匀分布的先验概率。

+ 若属性（即训练样例）有k个可能取值，那么 $p=1/k$ 。

× m最常见的取值就是所有不同有效字词的个数，即词汇表的大小。

× 此时若采用均匀分布的先验概率，则 $mp=1$ 。所以上式变为：

$$P(w_i | C) = \frac{TF(w_i) + 1}{\sum_{w_i \in W} TF(w_i) + |W|}$$

本章待续.....