《数据仓库与数据挖掘》

分类算法：贝叶斯学习

北京邮电大学
计算机应用中心
王小茹

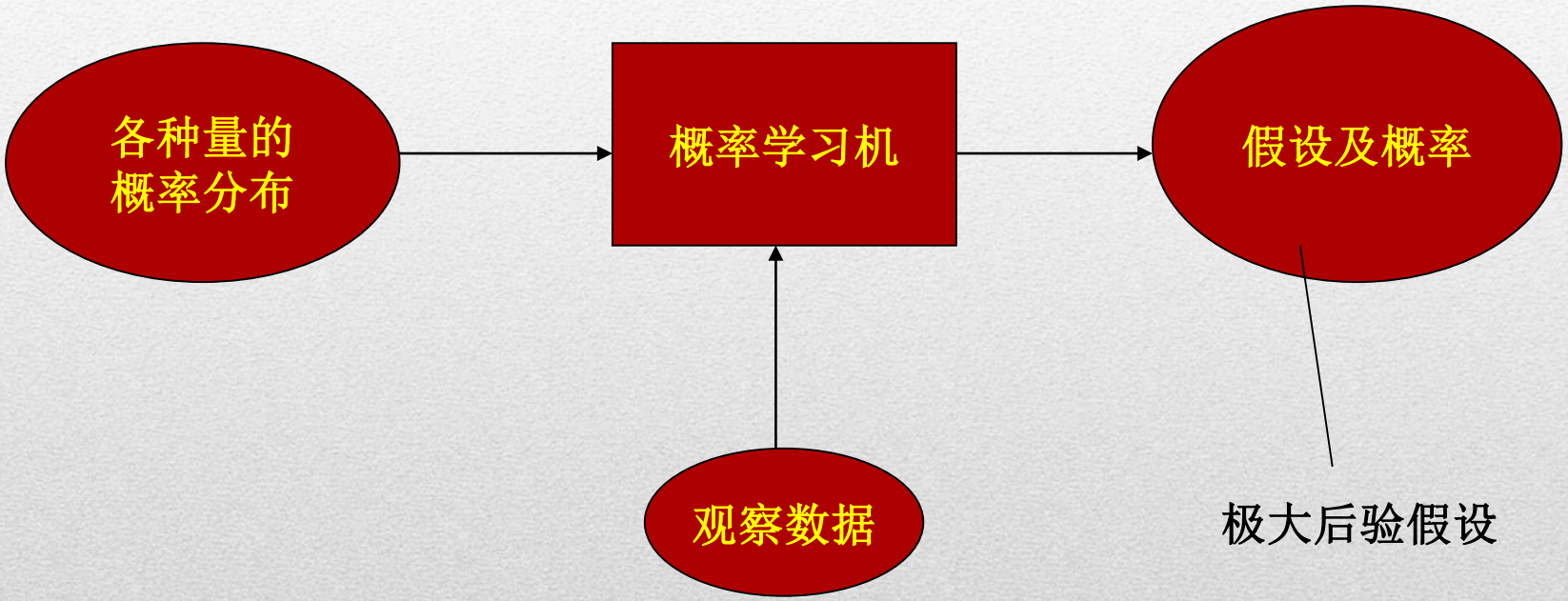
主要内容

- 贝叶斯法则
- 贝叶斯法则和概念学习
- 朴素贝叶斯分类器
- 实例：文本分类

贝叶斯法则

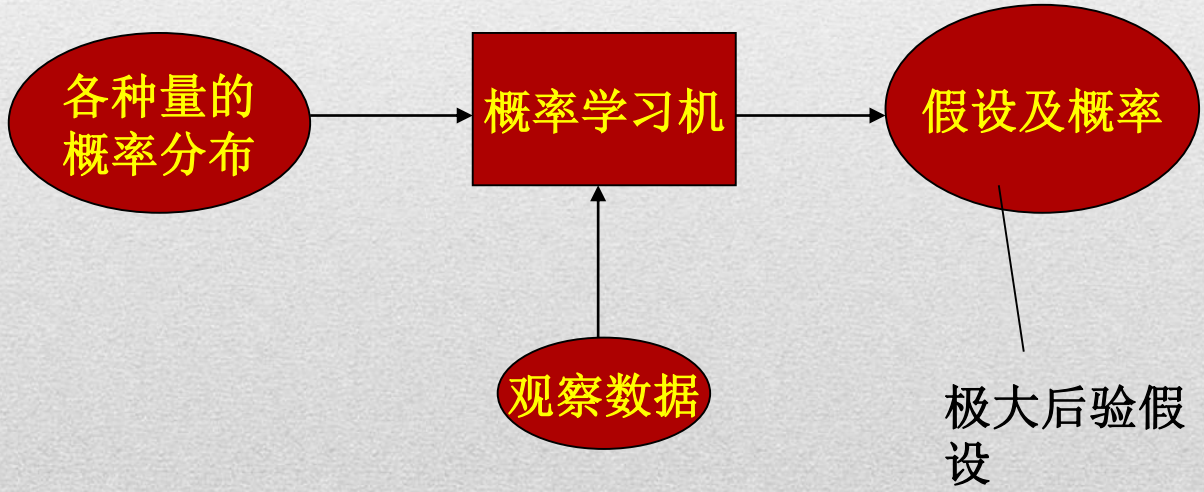
- 贝叶斯推理提供了推理的一种概率手段
- 两个基本假设
  - a）待考查的量遵循某概率分布
  - b）可根据这些概率及已观察到的数据进行推理，以作出最优的决策
- 贝叶斯推理对机器学习十分重要
  - 它为衡量多个假设的置信度提供了定量的方法
    - 为直接操作概率的学习算法提供了基础
    - 为其他算法的分析提供了理论框架

概率学习系统的一般框架



贝叶斯法则

- 机器学习的任务：在给定训练数据D时，确定假设空间H中的最佳假设
  - 最佳假设：在给定数据D以及H中不同假设的先验概率的有关知识下的最可能假设
- 贝叶斯理论提供了一种基于假设的先验概率、以及观察到的数据本身计算假设概率的方法



贝叶斯法则

贝叶斯法则

- 基本术语
  - D ：训练数据
  - H ： 假设空间
  - h ： 假设
  - P(h)：假设h的先验概率(Prior Probability)
    - 即没有训练数据前假设h拥有的初始概率
  - P(D)：训练数据的先验概率
    - 即在没有确定某一假设成立时D的概率
  - P(D|h)：似然度，在假设h成立的情况下，观察到D的概率
  - P(h|D)：后验概率，给定训练数据D时h成立的概率

贝叶斯法则

- 贝叶斯定理

$$P(h|D) = \frac{P(D|h)\,P(h)}{P(D)}$$

- 后验概率正比于P(h)和P(D|h)
- 反比于P(D)
  - D独立于h出现的概率越大，则D对h的支持度越小

- 贝叶斯公式是贝叶斯学习的基础，它提供了根据先验概率P(h)、P(D)以及观察概率P(D|h)，计算后验概率P(h|D)的方法

贝叶斯法则

- 极大后验（maximum a posteriori，MAP）假设
  - 给定数据D和H中假设的先验概率，具有最大后验概率的假设h

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(h|D)$$

$$= \underset{h \in H}{\operatorname{argmax}} \frac{P(D|h)\,P(h)}{P(D)}$$

$$= \underset{h \in H}{\operatorname{argmax}} P(D|h)\,P(h)$$

不依赖假设**h**为一个常量

贝叶斯法则

- 极大似然假设(Maximum Liklihood, ML)
  - 当H中的假设具有相同的先验概率时，给定h，使P(D|h)最大的假设h_{ML}：

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} P(h|D)$$

$$= \underset{h \in H}{\operatorname{argmax}} \frac{P(D|h)\,P(h)}{P(D)}$$

$$= \underset{h \in H}{\operatorname{argmax}} P(D|h)\,P(h)$$

$$= \underset{h \in H}{\operatorname{argmax}} P(D|h)$$

不依赖假设**h**为一个常量

贝叶斯法则

计算$P(h|D)$的方法

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

- 贝叶斯公式
- 极大后验（maximum a posteriori, MAP）假设

给定数据D时可能性最大的假设$h\in H$

计算公式

$$\begin{aligned}h_{MAP} &\equiv \arg\max_{h\in H} P(h|D)\\ &= \arg\max_{h\in H}\frac{P(D|h)P(h)}{P(D)}\\ &= \arg\max_{h\in H} P(D|h)P(h)\end{aligned}$$

$P(D)$不依赖于h

- 极大似然（maximum likelihood, ML）假设

使$P(D|h)$（给定h时数据D的似然度）最大的假设

计算公式

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

贝叶斯法则

- 举例：一个天气估计问题
  - 两个假设H： h1={晴天}、h2={非晴天}
  - 可观察到的数据：温度高+和温度低-
  - 先验知识p(h)
    - 北京晴天的概率0.99： P(h1)=0.99
    - 非晴天0.01： P(h2)=0.01
  - 观察到的概率P(D|h)：
    - P(温度高 | 晴天) = 0.85
    - P(温度低 | 非晴天) = 0.93
  - 问题：现在观察到温度低，判断是否非晴天？

贝叶斯法则

- 极大后验计算
  - P(非晴天 | 温度低)
    $\propto$P(温度低|非晴天)*P(温度低)
    = 0.93*0.01 = 0.0093
  - P(晴天 | 温度低)
    $\propto$ P(温度低|晴天)*P(温度高)
    = 0.15*0.99 = 0.1485
  - 答案：晴天

贝叶斯法则

- 极大似然计算
  - P(非晴天 | 温度低)
    $\propto$ P(温度低|非晴天)
    = 0.93
  - P(晴天 | 温度低)
    $\propto$ P(温度低|晴天)
    = 0.15
  - 答案：非晴天

- 另一个例子：医疗诊断问题
  - 两个假设H：h1={病人有癌症} h2={病人无癌症}
  - 可观察数据为化验结果 ：⊕（正）和⊖（负）
  - 先验知识：P(cancer)=0.008
  - 观察数据： $P(\oplus \mid cancer) = 0.98$ $P(\Theta \mid \neg cancer) = 0.97$
- 问题：新来的病人的检查结果为⊕，应如何诊断？

贝叶斯法则

贝叶斯法则—示例

可选的假设：病人有癌症；病人无癌症

$$P(cancer)=0.008 \qquad P(\neg cancer)=0.992$$
$$P(\oplus \mid cancer)=0.98 \qquad P(\ominus \mid cancer)=0.02$$
$$P(\oplus \mid \neg cancer)=0.03 \qquad P(\ominus \mid \neg cancer)=0.97$$

假定有一病人，化验测试⊕，该病人有癌症否？

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

$$P(\oplus \mid cancer)P(cancer)=(0.98)\cdot(0.008)=0.0078$$
$$P(\oplus \mid \neg cancer)P(\neg cancer)=(0.03)\cdot(0.992)=0.0298$$

因此，$h_{MAP}=\neg cancer$

注：(1)贝叶斯推理的结果很大地依赖于先验概率；

(2)并没有完全地被接受或拒绝假设，而只是在观察到较多的数据后假设的可能性增大或减小了。

基本概率公式表

- 乘法规则：
  $P(A \wedge B)=P(A|B)P(B)=P(B|A)P(A)$
- 加法规则：$P(A \vee B)=P(A)+P(B)-P(A \wedge B)$
- 贝叶斯法则：$P(h|D)=P(D|h)P(h)/P(D)$
- 全概率法则：如果事件$A_1 \ldots A_n$互斥，且满足 $\sum_{i=1}^{n} P(A_i)=1$，则

$$P(B)=\sum_{i=1}^{n} P(B|A_i)P(A_i)$$

- 贝叶斯法则
- 贝叶斯法则和概念学习
- 朴素贝叶斯分类器
- 实例：文本分类

主要内容

贝叶斯法则和概念学习

- 贝叶斯法则为计算给定训练数据下任一假设的后验概率提供了原则性方法
- 概念学习问题
  - 定义在实例空间X上的有限的假设空间H，任务是学习某个目标概念$c: X \rightarrow \{0,1\}$
- **Brute-Force MAP学习算法**
  - 对于H中每个假设h，计算后验概率

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

  - 输出有最高后验概率的假设

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

  - Brute-Force算法需要计算每个假设的后验概率，计算复杂度较高

MAP假设和一致学习器

- 假定
  - 训练数据D是无噪声的，即$d_i=c(x_i)$
  - 目标概念c包含在假设空间H中
  - 每个假设的概率相同
- 从而
  - 由于所有假设的概率之和是1，因此
  - 由于训练数据无噪声，那么给定假设h时，与h一致的D的概率为1，不一致的概率为0，因此 $P(h)=\frac{1}{|H|}$

$$P(D|h)=\begin{cases}1 & \forall d_i, d_i=h(x_i)\\ 0 & otherwise\end{cases}$$

MAP假设和一致学习器

- 考虑Brute-Force MAP算法
  - h与D不一致

$$P(h|D) = \frac{0 \cdot P(h)}{P(D)} = 0$$

  - h与D一致

$$P(h|D) = \frac{1 \cdot \frac{1}{|H|}}{P(D)} = \frac{\frac{1}{|H|}}{P(D)}$$

  - 每个一致的假设都是MAP假设

MAP假设和一致学习器

- P(D)的计算

$$P(D) = \sum_{h_i \in H} P(D|h_i)P(h_i)$$

$$= \sum_{h_i \in VS_{H,D}} 1 \times \frac{1}{|H|} + \sum_{h_i \notin VS_{H,D}} 0 \times \frac{1}{|H|}$$

$$= \sum_{h_i \in VS_{H,D}} 1 \times \frac{1}{|H|}$$

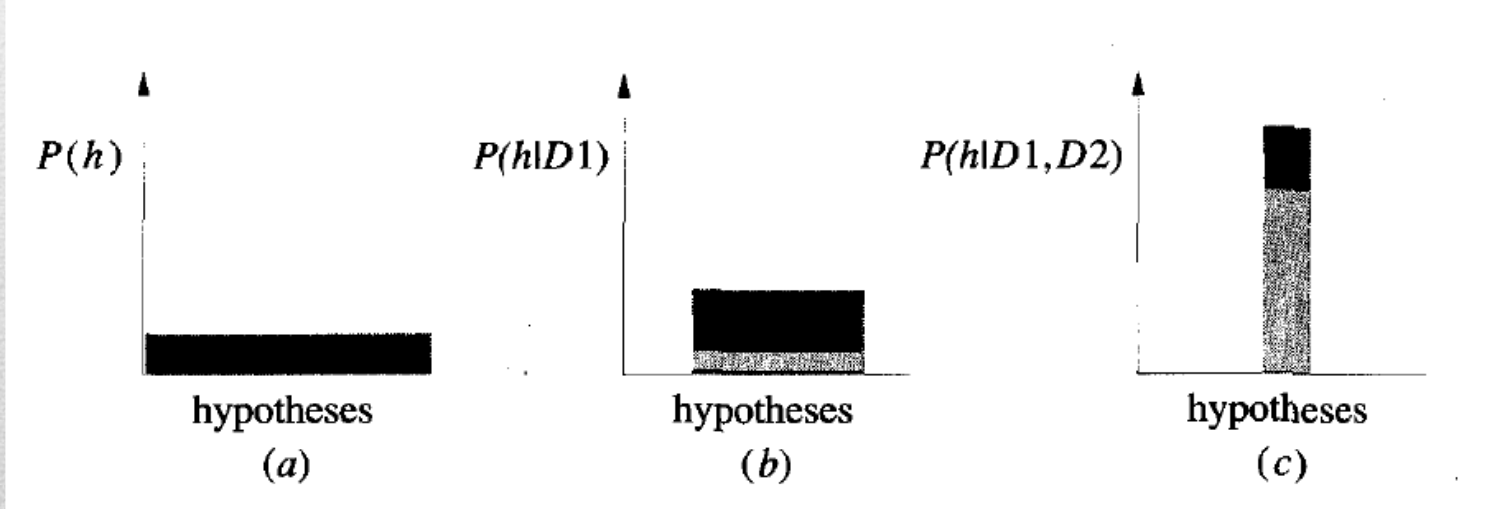
$$= \frac{|VS_{H,D}|}{|H|}$$

- 因此

$$P(h|D) = \frac{\frac{1}{|H|}}{P(D)} = \frac{\frac{1}{|H|}}{\frac{|VS_{H,D}|}{|H|}} = \frac{1}{|VS_{H,D}|}$$

MAP假设和一致学习器

- 假设的概率演化情况
  - 初始时所有假设具有相同的概率
  - 当训练数据逐步出现后，不一致假设的概率变为0，而整个概率的和为1，它们均匀分布到剩余的一致假设中



MAP假设和一致学习器

- 一致学习器
  - 如果某个学习器输出的假设在训练样例上为0错误率，则称为一致学习器
- 如果H上有均匀的先验概率，且训练数据是确定性和无噪声的，任意一致学习器将输出一个MAP假设
- Find-S算法按照特殊到一般的顺序搜索架设空间H，并输出一个极大特殊的一致假设，因此可知在上面定义的P(h)和P(D|h)概率分布下，它输出MAP假设
- 更一般地，对于先验概率偏袒于更特殊假设的任何概率分布，Find-S输出的假设都是MAP假设
- 因此：贝叶斯框架提出了一种刻画学习算法（如Find-S算法）行为的方法，即使该学习算法不进行概率操作

- 贝叶斯学习的重要性
  - 能够计算显式的假设概率，是解决相应学习问题的最有实际价值的方法之一
  - 为理解其它学习算法提供了一种有效的手段，而这些算法不一定直接操作概率数据
- 特点
  - 观察到的每个训练样例可以增量式地降低或升高某假设的估计概率
    - 其他算法会在某个假设与任一样例不一致时完全去掉该假设
  - 先验知识可以与观察数据一起决定假设的最终概率
  - 贝叶斯方法可允许假设做出不确定性的预测
    - 比如今天是晴天的概率为93%

小结

- 难点
  - 需要概率的初始知识
    - 当这概率预先未知时，可以基于背景知识、预先准备好的数据以及关于基准分布的假定来估计这些概率
  - 一般情况下确定贝叶斯最优假设的计算代价比较大（同候选假设的数量成线性关系）

小结

- 贝叶斯法则
- 贝叶斯法则和概念学习
- 朴素贝叶斯分类器
- 实例：文本分类

主要内容

- 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？

贝叶斯最优分类器

- 贝叶斯最优分类器
  - 如果新实例的可能分类可取某集合V中的任一值v_j，那么概率$P(v_j|D)$表示新实例分类为v_j的概率

$$P(v_j \mid D) = \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D)$$

  - 新实例的最优分类为使$P(v_j|D)$最大的v_j值，为：

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D)$$

贝叶斯最优分类器

贝叶斯最优分类器

- 例子
  - 假设空间H：包含三个假设h_1，h_2，h_3
  - 给定数据D时，三个假设的后验概率分别是
    - $P(h_1|D)=0.4$，$P(h_2|D)=0.3$，$P(h_3|D)=0.3$
  - 若一新实例x被h_1分类为正，被h_2和h_3分类为反
  - 问题：给出x的分类？

贝叶斯最优分类器

- 例子
  - 已知：
    - 新实例的可能分类集合为V={+, −}
    - $P(h_1|D)=0.4,\ P(-|h_1)=0,\ P(+|h_1)=1$
    - $P(h_2|D)=0.3,\ P(-|h_2)=1,\ P(+|h_2)=0$
    - $P(h_3|D)=0.3,\ P(-|h_3)=1,\ P(+|h_2)=0$
  - 因此
    - $\sum_{h_i \in H} P(+|h_i)P(h_i|D)=0.4 \qquad \sum_{h_i \in H} P(-|h_i)P(h_i|D)=0.6$
    - $\arg\max_{v_j \in \{+,-\}, h_i \in H} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)=-$

- 贝叶斯最优分类器能从给定训练数据中获得最好的性能，但算法的开销很大

贝叶斯最优分类器

- 学习任务
  - 给定训练集合D={(x_i, y_i)}，其中每个实例x=< $a_1, \ldots, a_n$ >，由属性值的合取描述，$y_i \in V$
  - 学习一个分类函数f(x)
  - 对新给定的实例x_{new}=<$a_1, \ldots, a_n$>，得到最可能的目标值v_{MAP}

$$v_{MAP} = \arg\max_{v_j} P(v_j \mid a_1, \ldots, a_n)$$

朴素贝叶斯分类器

朴素贝叶斯分类器

- 使用贝叶斯公式

$$v_{MAP} = \arg\max_{v_j \in V} \frac{P(a_1,...,a_n \mid v_j)P(v_j)}{P(a_1,...,a_n)}$$

$$= \arg\max_{v_j \in V} P(a_1,...,a_n \mid v_j)P(v_j)$$

- 基于训练数据估计p(v)和p(x|v)
  - 估计$P(v_j)$很容易
    - 计算每个目标值v_j出现在训练数据中的频率
  - 估计$P(a_1, \ldots a_n \mid v_j)$遇到数据稀疏问题
    - 除非有一个非常大的训练数据集，否则无法获得可靠的估计

朴素贝叶斯分类器

- 独立性假设
  - 在给定目标值时，属性值之间相互条件独立，即

$$P(a_1,\ldots,a_n \mid v_j) = \prod_i P(a_i \mid v_j)$$

- 最终的朴素贝叶斯分类器

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j)$$

朴素贝叶斯分类器

- 从训练数据中估计不同$P(a_i|v_j)$项的数量比要估计$P(a_1,...,a_n|v_j)$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v_{NB}等于MAP分类，否则是近似
- 朴素贝叶斯分类器与其他学习方法的区别
  - 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

- 例子

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

- 给新实例分类：

<Outlook=sunny, Temperature=cool, Humidity=high, Wind=strong>

朴素贝叶斯分类器

朴素贝叶斯分类器

- 朴素贝叶斯分类器

$$v_{NB} = \arg\max_{v_j \in \{yes, no\}} P(v_j) \prod_i P(a_i \mid v_j)$$

$$= \arg\max_{v_j \in \{yes, no\}} P(v_j) P(sunny \mid v_j) P(cool \mid v_j) P(high \mid v_j) P(strong \mid v_j)$$

- 根据上计算出上式需要的概率值
  - P(yes)=9/14=0.64
  - P(no)=5/14=0.36
  - P(strong|yes)=3/9=0.33
  - P(strong|no)=3/5=0.60
  - ...
- 求v_{NB}
  - P(yes)P(sunny|yes)P(cool|yes)P(high|yes)P(strong|yes)=0.0053
  - P(no)P(sunny|no)P(cool|no)P(high|no)P(strong|no)=0.0206
  - v_{NB}=no

- 估计概率
  - 我们通过在全部事件基础上观察某事件出现的比例来估计概率
  - 当样本很小时，采用平滑技术，m-估计

$$\frac{n_c + mp}{n + m}$$

  - p是将要确定的概率的先验估计
    - 在缺少其他信息时，选择p的一种典型的方法是均匀概率，比如某属性有k个可能值，那么p=1/k
  - m是一个称为等效样本大小的常量
    - m被称为等效样本大小的原因是将n个实际的观察样本扩大，加上m个按p分布的虚拟样本

朴素贝叶斯分类器

大纲

- 贝叶斯法则
- 贝叶斯法则和概念学习
- 朴素贝叶斯分类器
- 实例：文本分类

文本分类

- 问题框架
  - 实例空间X包含了所有的文本文档
  - 给定某未知目标函数f(x)的一组训练样例$D=\{(x_i, f(x_i))\}$，f(x)的值来自某有限集合V
  - 任务是从训练样例中学习，预测后续文本文档的目标值

- 应用实例
  - 垃圾邮件
  - 我感兴趣的电子新闻稿
  - 讨论机器学习的万维网页

文本分类

- 设计朴素贝叶斯分类器的两个主要问题：
  - 文档的表示
    - 怎样将任意文档表示为属性值的形式
  - 概率的估计
    - 如何估计朴素贝叶斯分类器所需的概率

文本分类

- 假定：
  - 我们共有1000个训练文档
    - 其中700个分类为dislike，300个分类为like，
  - 现在要对下面的新文档进行分类：
    - This is an example document for the naive Bayes classifier. This document contains only one paragraph, or two sentences.

文本分类

- 文档的表示
  - 给定一个文本文档，对每个单词的位置定义一个属性，该属性的值为在此位置上找到的英文单词
- 例如：以下文档
  - This is an example document for the naive Bayes classifier. This document contains only one paragraph, or two sentences
  - 可表示为
    - 19个属性
    - 每个属性的值
      - a1=this, a2=is, a3=an, …

文本分类

- 概率估计

$$v_{NB} = \arg\max_{v_j \in \{like, dislike\}} P(v_j) \prod_{i=1}^{19} P(a_i \mid v_j)$$

$$= \arg\max_{v_j \in \{like, dislike\}} P(v_j) P(a_1 = "this" \mid v_j) ... P(a_{19} = "sentences" \mid v_j)$$

- 此处贝叶斯分类器隐含的独立性假设并不成立。通常，某个位置上出现某个单词的概率与前后位置上出现的单词是相关的
- 但是在实践中，朴素贝叶斯学习器在许多文本分类问题中性能非常好

- 概率项$P(v_i)$的估计
  - P(vi)的估计：P(dislike)=0.7，p(like)=0.3
- 概率项$P(a_i=w_k|v_i)$的估计
  - 后一项含三个参数
    - 位置 i，单词 k，假设 j
    - 需要估计的概率项总数： 2 * 19 * 50000 ≈ 2000000
  - 引入新的假设以减少需要估计的概率项的数量
    - 假定单词w_k出现的概率独立于单词所在的位置，即$P(a_i=w_k|v_i)=P(w_k|v_j)$
    - 新的需要估计的概率项总数： 2 * 50000 ≈ 100000

文本分类

文本分类

- 概率项$P(a_i=w_k|v_i)$的估计
  - 采纳m-估计方法

$$P(w_k \mid v_j) = \frac{n_k + mp}{n+m} = \frac{n_k + 1}{n+|Vocabulary|}$$

  - n 是在类别v_j中词出现的位置总数（或者说词频总数）
  - n_k是词w_k出现的位置总数（或者说词频）
  - $m = |Vocabulary|$
  - $p = 1/|Vocabulary|$

Learn_naive_Bayes_text(*Examples*, *V*)

Examples 为一组文本文档以及它们的目标值。*V* 为所有可能目标值的集合。此函数作用是学习概率项 $P(w_k|v_j)$，它描述了从类别 v_j 中的一个文档中随机抽取的一个单词为英文单词 w_k 的概率。该函数也学习类别的先验概率 $P(v_j)$。

1.收集 *Examples* 中所有的单词、标点符号以及其他记号

- *Vocabulary*←在 *Examples* 中任意文本文档中出现的所有单词及记号的集合

2.计算所需要的概率项 $P(v_j)$和 $P(w_k|v_j)$

- 对 *V* 中每个目标值 v_j
  - $docs_j$←*Examples* 中目标值为 v_j 的文档子集
  - $P(v_j) \leftarrow \dfrac{|docs_j|}{|Examples|}$
  - $Text_j$←将 $docs_j$ 中所有成员连接起来建立的单个文档
  - n←在 $Text_j$ 中不同单词位置的总数
  - 对 *Vocabulary* 中每个单词 w_k
    - n_k←单词 w_k 出现在 $Text_j$ 中的次数
    - $P(w_k|v_j) \leftarrow \dfrac{n_k+1}{n+|Vocabulary|}$

文本分类的朴素贝叶斯算法

Classify_naive_Bayes_text(*Doc*)

对文档 *Doc* 返回其估计的目标值。a_i 代表在 *Doc* 中的第 i 个位置上出现的单词。

- *positions* ← 在 *Doc* 中包含的能在 *Vocabulary* 中找到的记号的所有单词位置
- 返回 $v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_{i \in positions} P(a_i \mid v_j)$

文本分类的朴素贝叶斯算法

实验结果

- Joachims将此算法用于新闻组文章的分类
  - 每一篇文章的分类是该文章所属的新闻组名称
  - 20个新闻组，每个新闻组有1000篇文章，共2万个文档
  - 2/3作为训练样例，1/3进行性能测量
  - 词汇表不包含最常用词（比如the、of）和罕见词（数据集中出现次数少于3）
  - 结果
    - 那么随机猜测的分类精确度为5%。由程序获得的精确度为89%

- Lang用此算法学习目标概念“我感兴趣的新闻组文章”
  - NewsWeeder系统，让用户阅读新闻组文章并为其评分，然后使用这些评分的文章作为训练样例，来预测后续文章哪些是用户感兴趣的
  - 每天向用户展示前10%的自动评分文章，它建立的文章序列中包含的用户感兴趣的文章比通常高3～4倍
  - 例如，若一个用户对通常的文章有16%感兴趣，其对于NewsWeeder 推荐的文章有59%感兴趣。

实验结果