# 贝叶斯分类器

泽林信息


# 概率

# 贝叶斯定理

- 条件概率：

$$P(A|B) = \frac{P(AB)}{P(B)}$$

- 全概率公式：

$$P(A) = \sum_{j} P(A|B_j)P(B_j)$$

- 贝叶斯公式：

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j} P(A|B_j)P(B_j)}$$

# 贝叶斯定理的含义

- $P(B_i)$称为"先验概率"（Prior probability），即在A事件发生之前，我们对$B_i$事件概率的一个判断。
- $P(B_i|A)$称为"后验概率"（Posterior probability），即在A事件发生之后，我们对$B_i$事件概率的重新评估。
- P(A| $B_i$)/P(A)称为"可能性函数"（Likelyhood），这是一个调整因子，使得预估概率更接近真实概率。

所以，贝叶斯定理可以理解成下面的式子：

后验概率 = 先验概率 X 调整因子

这就是贝叶斯定理的含义。我们先预估一个"先验概率"，然后加入实验结果，看这个实验到底是增强还是削弱了"先验概率"，由此得到更接近事实的"后验概率"。

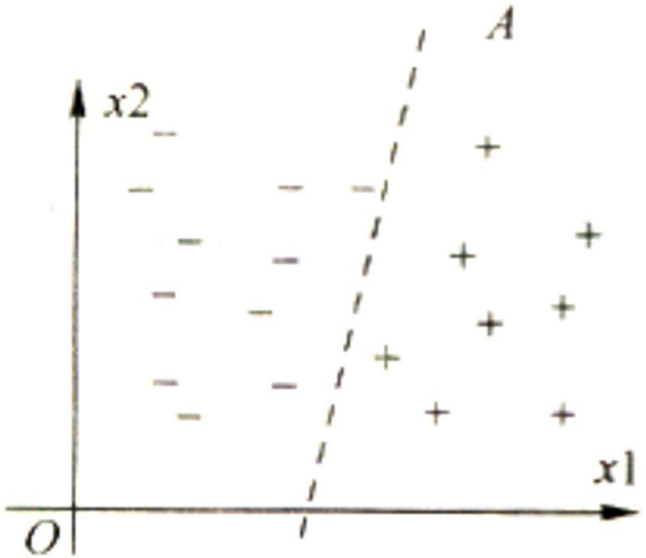
02

PART TWO

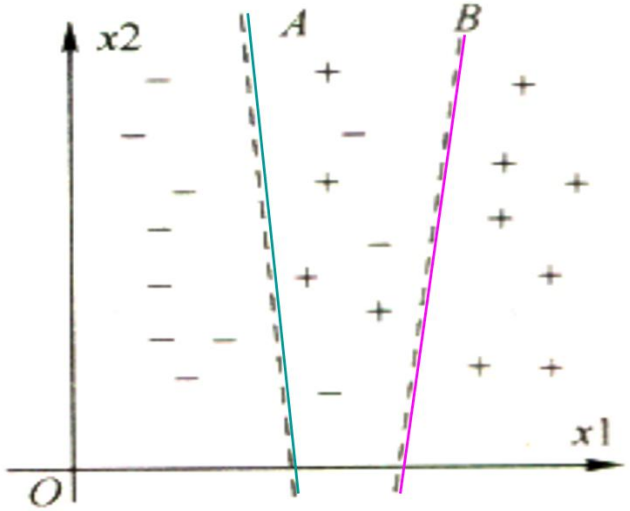
# 贝叶斯决策论

# 贝叶斯决策论

贝叶斯决策论是概率框架下实施决策的基本方法。在分类问题上，贝叶斯决策论考虑的是如何基于这些概率和误判损失来来选择最优的类别。




情况一




情况二

# 贝叶斯决策论

这会引起一个与损失有关联的概率——风险。在做出决策时，要考虑所承担的风险。

- **假设：**有N中可能的类别标记，即 $y = \{c_1, c_2, \cdots, c_N\}$

条件风险 $$\mathrm{R}(c_i|x) = \sum_{j=1}^{N} \lambda_{ij} P(c_j|x)$$

其中，$\lambda_{ij}$是将一个真实的标记$c_j$的样本错误分类为$c_i$所产生的损失。

- **任务：**找到一个判定准则使最小化条件风险。 $h: x \to y$

贝叶斯判定准则 $h^*(x) = \underset{c \in y}{\operatorname{argmin}} R(c|x)$

# 贝叶斯决策论

- **目标：**最小化分类错误率

$$\lambda_{ij} = \begin{cases} 0, \text{如果} i = j \\ 1, \text{如果} i \neq j \end{cases}$$

条件风险 $\mathrm{R}(c|x) = 1 - P(c|x)$

- **分类器：**最小化分类错误率的贝叶斯分类器为

$$h^*(x) = \underset{c \in y}{\operatorname{argmax}} P(c|x) = \underset{c \in y}{\operatorname{argmax}} \frac{P(c)P(x|c)}{P(x)}$$

03

PART THREE

# 朴素贝叶斯分类器

# 朴素贝叶斯分类器

$$h^*(x) = \underset{c \in y}{\operatorname{argmax}} P(c|x) = \underset{c \in y}{\operatorname{argmax}} \frac{P(c)P(x|c)}{P(x)}$$

- **困难：**

  基于贝叶斯的分类器的主要困难在于估计P(x|c)所有属性上的联合概率，难以从有限的样本里面计算出来。

- **方法：**属性条件独立性假设

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)} = \frac{P(c)}{P(x)} \prod_{i=1}^{d} P(x_i|c)$$

- **分类器：**对于所有类别来说，P(x)是相同的。

$$h^*(x) = \underset{c \in y}{\operatorname{argmax}} P(c) \prod_{i=1}^{d} P(x_i|c)$$

# 朴素贝叶斯分类器

$$h^*(x) = \underset{c \in y}{\operatorname{argmax}} P(c) \prod_{i=1}^{d} P(x_i|c)$$

朴素贝叶斯的训练过程就是基于训练集D来估计类先验概率P(c)，并为每个属性(变量)估计条件概率$P(x_i|c)$

- **P(c)**：令$D_c$表示训练集D中第c类样本组成的集合。

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

- **P($x_i$|c)**

离散变量

$$P(x_i|c) = \frac{|D_{c,x_i}|}{|D_c|}$$

表示$D_c$中在第i个变量上取$x_i$数量

连续变量

$$p(x_i|c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right)$$

# 举个例子

# 数据和问题

不帅、性格不好、身高矮、不上进 → 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 |
| 不帅 | 不好 | 矮 | 不上进 | 不嫁 |
| 帅 | 好 | 高 | 不上进 | 嫁 |
| 不帅 | 好 | 高 | 上进 | 嫁 |
| 帅 | 好 | 高 | 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 |

# 计算P(c)

P(不嫁)=6/12=0.5　　P(嫁)=6/12=0.5

| 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 |

# 计算P(帅？|嫁？)

P(不帅|不嫁)=1/6　　P(不帅|嫁)=3/6

P(帅|不嫁)=5/6　　P(帅|嫁)=3/6

| 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 |

# 计算P(性格？|嫁？)

P(不好|不嫁)=3/6　　P(不好|嫁)=1/6

P(好|不嫁)=3/6　　P(好|嫁)=5/6

| 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 |

# 计算P(身高？|嫁？)

P(矮|不嫁)=6/6　P(中|不嫁)=0/6　　P(矮|嫁)=1/6　　P(中|嫁)=2/6

P(高|不嫁)=0/6　　P(高|嫁)=3/6

| 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 |

# 计算P(上进？|嫁？)

P(不上进|不嫁)=3/6

P(上进|不嫁)=3/6

P(不上进|嫁)=1/6

P(上进|嫁)=5/6

| 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 |

# 问题来了

- **问题一 :**不帅、性格不好、身高矮、不上进 → 嫁还是不嫁？

P(不嫁|不帅，不好，矮，不上进)=P(不嫁) P(不帅|不嫁)P(不好|不嫁)P(矮|不嫁)P(不上进|不嫁)

V

=1/2X1/6X3/6X6/6X3/6=1/48

P(嫁|不帅，不好，矮，不上进)=P(嫁) P(不帅|嫁)P(不好|嫁)P(矮|嫁)P(不上进|嫁)

=1/2X3/6X1/6X1/6X1/6=1/124

不嫁

- **问题二 :**不帅、性格不好、身高高、不上进 → 嫁还是不嫁？

P(高|不嫁)=0/6 → 嫁 ?

04

PART FOUR

# 拉普拉斯平滑

# 拉普拉斯平滑

有些变量的值在训练集中没有出现，但在测试集中出现了，这种问题该怎么办？

- 拉普拉斯平滑

为了避免其他属性携带的信息被没有出现的变量值被抹去，在估计概率值时，通常需要进行“平滑”处理。

$$P(c) = \frac{|D_c| + 1}{|D| + \mathrm{N}}$$

类别数量

$$P(x_i|c) = \frac{|D_{c,x_i}| + 1}{|D_c| + \mathrm{N}i}$$

表示$D_c$中在第i个变量上取$x_i$数量

# 遗留问题

- **问题二 :**不帅、性格不好、身高高、不上进 → 嫁还是不嫁？

P(不嫁|不帅，不好，高，不上进)=P(不嫁) P(不帅|不嫁)P(不好|不嫁)P(高|不嫁)P(不上进|不嫁)

=1/2X1/4X4/8X1/9X4/8=1/144

V

P(嫁|不帅，不好，矮，不上进)=P(嫁) P(不帅|嫁)P(不好|嫁)P(高|嫁)P(不上进|嫁)

=1/2X4/8X2/8X1/8X2/8=1/512

不嫁

05

PART FIVE

# 优缺点及应用场景

# 优缺点

优点：

- 既简单又快速，预测表现良好；
- 如果变量独立这个条件成立，相比Logistic回归等其他分类方法，朴素贝叶斯分类器性能更优，且只需少量训练数据；
- 相较于数值变量，朴素贝叶斯分类器在多个分类变量的情况下表现更好。若是数值变量，需要正态分布假设。

缺点:

- 如果分类变量的类别（测试数据集）没有在训练数据集总被观察到，那这个模型会分配一个0（零）概率给它，同时也会无法进行预测。这通常被称为“零频率”。为了解决这个问题，我们可以使用平滑技术，拉普拉斯估计是其中最基础的技术。
- 朴素贝叶斯也被称为bad estimator，所以它的概率输出predict_proba不应被太认真对待。
- 朴素贝叶斯的另一个限制是独立预测的假设。在现实生活中，这几乎是不可能的，各变量间或多或少都会存在相互影响。

# 应用场景

- 实时预测：简单快速
- 文档分类：将文档分词之后有很多特征属性
- 推荐系统：可以结合协同过滤来完成推荐
- 多类预测：因多类预测而闻名

PART SIX

# 项目实战

# 案列分析

## 垃圾邮件分类

- 思路：

2、贝叶斯公式 我们要做的是计算在已知词向量$w=(w_1,w_2,\ldots,w_n)$的条件下求包含该词向量邮件是否为垃圾邮件的概率，即求：$P(s|w),w=(w_1,w_2,\ldots,w_n)$ 其中，$s$表示分类为垃圾邮件 根据贝叶斯公式和全概率公式，$P(s|w_1,w_2,\ldots,w_n)=\frac{P(s,w_1,w_2,\ldots,w_n)}{P(w_1,w_2,\ldots,w_n)}=\frac{P(w_1,w_2,\ldots,w_n|s)P(s)}{P(w_1,w_2,\ldots,w_n)}$

$=\frac{P(w_1,w_2,\ldots,w_n|s)P(s)}{P(w_1,w_2,\ldots,w_n|s)\cdot p(s)+P(w_1,w_2,\ldots,w_n|s')\cdot p(s')}$ ...式1 根据朴素贝叶斯的条件独立假设，并设先验概率$P(s)=P(s')=0.5$,上式可化为：

$=\frac{\prod_{j=1}^{n}P(w_j|s)}{\prod_{j=1}^{n}P(w_j|s)+\prod_{j=1}^{n}P(w_j|s')}$ 再利用贝叶斯$P(w_j|s)=\frac{P(s|w_j)\cdot P(w_j)}{P(s)}$,式子化为$=\frac{\prod_{j=1}^{n}P(s|w_j)}{\prod_{j=1}^{n}P(s|w_j)+\prod_{j=1}^{n}P(s'|w_j)}=\frac{\prod_{j=1}^{n}P(s|w_j)}{\prod_{j=1}^{n}P(s|w_j)+\prod_{j=1}^{n}(1-P(s|w_j))}$ ...式2 至此，我们接下来会用式2来计算概率$P(s|w)$，为什么不用式1而用式2来计算概率，是因为通过式2可以将关于$s'$的部分用$s$表示，方便计算。

- 邮件数据：
  - 7000多封垃圾邮件，7000多封正常邮件
  - 300多封待测试的邮件