

Lecture 7: 贝叶斯决策论

朴素贝叶斯

概念: 朴素贝叶斯

- 概念:

- 假设训练集: $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ 由 $P(X, Y)$ 独立同分布产生;
- y 是样本的类别, $y_i = c_k$, 其中 $k = 1 \dots K$
- 朴素贝叶斯法就是要通过训练数据集学习联合概率分布 $P(X, Y)$, 进而通过样本 x 、计算其所属的类别, 即: $P(y = c_k | x)$
- 具体地, 需要学习以下先验概率分布及条件概率分布:
 - 先验概率分布: $P(Y = c_k), k = 1, 2, \dots, K$
 - 条件概率分布:
$$P(X = x | Y = c_k) = P(X^{(1)} = x^{(1)}, \dots, X^{(n)} = x^{(n)} | Y = c_k), k = 1, 2, \dots, K$$
 - 补充:
条件概率分布 $P(X = x | Y = c_k)$ 有指数级数量的参数, 其估计实际是不可行的。
- 之所以叫朴素贝叶斯法, 是因为它假定不同的数据之间、满足条件独立性假设, 即:

$$\begin{aligned} P(X = x | Y = c_k) &= P(X^{(1)} = x^{(1)}, \dots, X^{(n)} = x^{(n)} | Y = c_k) \\ &= \prod_{j=1}^n P(X^{(j)} = x^{(j)} | Y = c_k) \end{aligned}$$

- 朴素贝叶斯法实际上学习到生成数据的机制, 所以属于生成模型;
- 条件独立假设等于是说用于分类的特征在类确定的条件下都是条件独立的。这一假设使朴素贝叶斯法变得简单, 但有时会牺牲一定的分类准确率;
- 概念: 计算所属类别
 - 朴素贝叶斯法分类时, 对给定的输入 x , 通过学习到的模型计算后验概率分布 $P(Y = c_k | X = x)$, 将后验概率最大的类作为 x 的类输出;
 - 后验概率计算根据贝叶斯定理进行:

$$P(Y = c_k | X = x) = \frac{P(X = x | Y = c_k)P(Y = c_k)}{\sum_k P(X = x | Y = c_k)P(Y = c_k)}$$

- 分母上的每一个值、对于每一个类别而言都是一样的;
- 由于假设不同样本是条件独立, 根据假设导出贝叶斯定理中, 得到朴素贝叶斯的表示:

$$P(Y = c_k | X = x) = \frac{P(Y = c_k) \prod_j P(X^{(j)} = x^{(j)} | Y = c_k)}{\sum_k P(Y = c_k) \prod_j P(X^{(j)} = x^{(j)} | Y = c_k)}, k = 1, 2, \dots, K$$

- 最后, 因为是求后验概率最大的类作为 x 的类输出, 因此求最大值:

$$y = f(x) = \operatorname{argmax}_{c_k} \frac{P(Y = c_k) \prod_j P(X^{(j)} = x^{(j)} | Y = c_k)}{\sum_k P(Y = c_k) \prod_j P(X^{(j)} = x^{(j)} | Y = c_k)}$$

概念：后验概率最大化的含义

- 朴素贝叶斯法将实例分到后验概率最大的类中。这等价于期望风险最小化。假设选择 0-1 损失函数：

$$L(Y, f(X)) = \begin{cases} 1, & Y \neq f(X) \\ 0, & Y = f(X) \end{cases}$$

- 其期望风险函数为：

$$R_{\text{exp}}(f) = E[L(Y, f(X))]$$

- 期望是对联合分布 $P(X, Y)$ 取的。由此取条件期望：

$$R_{\text{exp}}(f) = E_X \sum_{k=1}^K [L(c_k, f(X))]P(c_k|X)$$

- 求其极小化：

$$\begin{aligned} f(x) &= \operatorname{argmin}_{y \in Y} \sum_{k=1}^K L(c_k, y)P(c_k|X=x) \\ &= \operatorname{argmin}_{y \in Y} \sum_{k=1}^K P(y \neq c_k|X=x) \\ &= \operatorname{argmin}_{y \in Y} (1 - P(y = c_k|X=x)) \\ &= \operatorname{argmax}_{y \in Y} P(y = c_k|X=x) \end{aligned}$$

两种模型

概念：判别式模型与生成式模型

- 引入：
 - $P(c|m)$ 在现实中通常难以直接获得从这个角度来看，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；
 - 两种基本策略：

判别式 (Discriminative) 模型

思路：直接对 $P(c|x)$ 建模

代表：

- 决策树
- BP 神经网络
- SVM

生成式 (Generative) 模型

思路：先对联合概率分布 $P(x, c)$ 建模，再由此获得 $P(c|x)$

$$P(c|x) = \frac{P(x, c)}{P(x)}$$

代表：贝叶斯分类器

- 概念：判别式
 - 将两类分开；

- 概念：生成式
 - 任何一个拿到的数据集，都可以认为其实对其的一个采样 x ，然后这个采样的样本因为是同一个数据集，总是默认认为不同样本之间是独立同分布的；
 - 如果可以把原本的数据集的分布恢复出来（因为我们的采样，可以理解为就是 $P(x, y)$ ，即从样本中取出的概率），把原本的分布复原出来，此时即知道样本的 $P(x)$ 概率，又知道从分布中取出的联合概率 $P(x, y)$ ，此时即可以把 $P(y|x)$ 求出来；

概念：频率主义和贝叶斯主义

- 假设：当前数据满足一个高斯分布
- 概念：频率主义
 - 代表：支持向量机；
 - 方法：统计学习 $\rightarrow$ 点估计
 - 它认为，高斯分布当中的期望和方差是确定存在的一个东西，当前需要做的事情、就是把期望和方差的参数估计出来，参数得到后、也就把分布得了出来，而当前的样本，就是从这个分布当中采样、摘出来的；
 - 参数是这个分布中实实在在存在的東西，可以认为其实参数空间实在存在的一个点；
- 概念：贝叶斯主义
 - 代表：贝叶斯学习
 - 方法：分布估计
 - 高斯分布的里面的参数不是实在存在的東西，只不过是刚好这次观测到了，感觉是这个分布；其实这个参数分身、是这样一个分布；
 - 参数本身不是一个点，参数本身也是一个分布，世界上没有一个东西是客观存在下来的，都是你观察到之后、你觉得它是这个样子；
 - 因此需要对期望和方差再进行假设、假设其满足某一个分布；

最大似然估计

概念：贝叶斯方法与最大似然估计

- 概念：
 - 先假设某种概率分布形式，再基于训练样例对参数进行估计；
 - 极大似然法是一种求解问题的方法论，是一个框架，基本思路就是：先把似然函数写出来，然后考虑每一个样本都是独立随机事件，那么就可以考虑每一个样本都出现、这个似然最大，通过连乘来做，并通过对数求和来做，得到使其连乘最大的参数；
- 分析：
 - 假设在一个数据集 D 中，抽取样本 x_1, x_2, x_3 ，因为假设抽取的样本是独立同分布的，因此可以将数据集 D 中抽取样本的概率连乘，得到同时抽取 x_1, x_2, x_3 概率；
 - 因此此时可以用连乘符号来表达；

先假设某种概率分布形式，再基于训练样例对参数进行估计

假定 $P(x | c)$ 具有确定的概率分布形式，且被参数 θ_c 唯一确定，则任务就是利用训练集 D 来估计参数 θ_c

θ_c 对于训练集 D 中第 c 类样本组成的集合 D_c 的似然(Likelihood)为

$$P(D_c | \theta_c) = \prod_{x \in D_c} P(x | \theta_c)$$

连乘易造成下溢，因此通常使用对数似然 (Log-Likelihood)

$$LL(\theta_c) = \log P(D_c | \theta_c) = \sum_{x \in D_c} \log P(x | \theta_c)$$

于是， θ_c 的极大似然估计为 $\hat{\theta}_c = \arg \max_{\theta_c} LL(\theta_c)$

朴素贝叶斯分类器

概念：朴素贝叶斯分类器

- 概念：

$$P(c | x) = \frac{P(c) P(x | c)}{P(x)}$$

主要障碍：所有属性上的联合概率难以从有限训练样本估计获得

组合爆炸；样本稀疏

基本思路：假定属性相互独立？

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

d 为属性数， x_i 为 x 在第 i 个属性上的取值

$P(x)$ 对所有类别相同，于是

$$h_{nb}(x) = \arg \max_{c \in \mathcal{Y}} P(c) \prod_{i=1}^d P(x_i | c)$$

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

□ 估计 $P(x|c)$:

- 对离散属性, 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

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

- 对连续属性, 考虑概率密度函数, 假定 $p(x_i | c) \sim \mathcal{N}(\mu_{c,i}, \sigma_{c,i}^2)$

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

• 举例:

(青绿; 稍蜷; 浊响; 清晰) - 好瓜 or 坏瓜?

编号	色泽	根蒂	敲声	纹理	好瓜	
1	青绿	蜷缩	浊响	清晰	是	$P(\text{青绿} \text{好瓜}) = 3/8$
2	乌黑	蜷缩	沉闷	清晰	是	$P(\text{青绿} \text{坏瓜}) = 3/9$
3	乌黑	蜷缩	浊响	清晰	是	$P(\text{稍蜷} \text{好瓜}) = 3/8$
4	青绿	蜷缩	沉闷	清晰	是	$P(\text{稍蜷} \text{坏瓜}) = 4/9$
5	浅白	蜷缩	浊响	清晰	是	$P(\text{浊响} \text{好瓜}) = 6/8$
6	青绿	稍蜷	浊响	清晰	是	$P(\text{浊响} \text{坏瓜}) = 4/9$
7	乌黑	稍蜷	浊响	稍糊	是	$P(\text{清晰} \text{好瓜}) = 7/8$
8	乌黑	稍蜷	浊响	清晰	是	$P(\text{清晰} \text{坏瓜}) = 2/9$
9	乌黑	稍蜷	沉闷	稍糊	否	$P(\text{青绿} \text{好瓜}) P(\text{稍蜷} \text{好瓜}) P(\text{浊响} \text{好瓜})$
10	青绿	硬挺	清脆	清晰	否	$P(\text{清晰} \text{好瓜}) P(\text{好瓜}=\text{yes}) = 3/8 \times 3/8$
11	浅白	硬挺	清脆	模糊	否	$\times 6/8 \times 7/8 \times 8/17$
12	浅白	蜷缩	浊响	模糊	否	
13	青绿	稍蜷	浊响	稍糊	否	$P(\text{青绿} \text{坏瓜}) P(\text{稍蜷} \text{坏瓜}) P(\text{浊响} \text{坏瓜})$
14	浅白	稍蜷	沉闷	稍糊	否	$P(\text{清晰} \text{坏瓜}) P(\text{好瓜}=\text{no}) = 3/9 \times 4/9 \times$
15	乌黑	稍蜷	浊响	清晰	否	$4/9 \times 2/9 \times 9/17$
16	浅白	蜷缩	浊响	模糊	否	
17	青绿	蜷缩	沉闷	稍糊	否	

把其似然算出来

- 好瓜的连乘比坏瓜的连乘大, 因此认为是好瓜;