第4章. 朴素贝叶斯法

提前学习部分

贝叶斯公式：$P(A|B)=\frac{P(B|A)\cdot P(A)}{P(B)}=P(A)\cdot\frac{P(B|A)}{P(B)}$

$$P(\text{规律}|\text{现象})=\frac{P(\text{现象}|\text{规律})\,P(\text{规律})}{P(\text{现象})}$$

思考模式

① 先验分布 $\pi(\theta)$ + 样本信息 $X \Rightarrow$ 后验分布 $\pi(\theta|X)$

意味：新观察到的样本信息将修正人们对事物的认知。

↓

从 $\pi(\theta)$ 修正为 后验分布 $\pi(\theta|X)$

② 边缘概率（即先验概率）

理解为：在联合概率中，把最终结果中那些不需要的事件通过合并成它们的全概率，从而消去它们。

（离散型随机变量 → 求和
连续 ····· → 积分）

③ 条件概率（即后验概率）← 指我们观测到的样本的分布，也就是似然函数。

④ 联合概率 表示2个事件共同发生的概率

⑤ 后验概率 = 先验概率 × 调整因子

↓ $P(A|B)$ ↓ $P(A)$ ↓ $\frac{P(B|A)}{P(B)}$

先预估一个“先验概率”，然后加入实验结果，看是增强还是削弱了“先验概率”

宏观上来看

决策 $Y=f(X)$

条件概率 $\max P(Y|X)$

生成 $P(Y|X)=\frac{P(X,Y)}{P(X)}$

判别 $Y=f(X)$, $P(Y|X)$

$Y=f(X) \Rightarrow P(Y|X) \Rightarrow P(Y|X)=\frac{P(X,Y)}{P(X)}$

$Y=f(X)$：不考虑X,Y随机性，感知机

$P(Y|X)$：考虑Y随机性

$P(Y|X)=\frac{P(X,Y)}{P(X)}$：考虑X与Y随机性，朴素贝叶斯公式

~~贝叶斯估计(4.2.3)~~ 推导：用极大似然法估计贝叶斯参数 为什么 $P(Y=C_k)=\frac{m}{N}$

① $Y \in \{C_1, C_2, \cdots, C_K\}$ 意为Y取 C_i 时概率为 θ_i

$\theta_1 \quad \theta_2 \quad \cdots \quad \theta_K$ $\quad \sum_{i=1}^{K} \theta_i = 1$

θ (向量)

② $P(Y=y|\theta) = \theta_1^{I(y=C_1)} \times \theta_2^{I(y=C_2)} \times \cdots \times \theta_K^{I(y=C_K)}$

意为给定 θ_1，问你 y 是不是等于 C_1 $I(y=C_1)$ 是一个事件函数

③ 在极大似然估计中，拿到 $y_1, y_2, \cdots y_N$ 样本

根据样本写出样本的概率分布 y_i 是独立的

$P(y_1, y_2, \cdots y_N|\theta) = P(y_1|\theta) \times P(y_2|\theta) \cdots P(y_N|\theta)$

$= \theta_1^{m_1} \cdot \theta_2^{m_2} \cdots \theta_K^{m_K}$

出现结果是 C_1 的概率是 m_1 次数 $\quad \therefore m_1 + \cdots + m_K = N$

④ 最大化这个联合概率分布

$\max P(y_1, y_2 \cdots y_N|\theta)$

$= \max \ln P(\sim) = m_1 \ln\theta_1 + m_2 \ln\theta_2 + \cdots + m_K \ln\theta_K$

要求使最大的 θ 的向量 还有一个约束 s.t. $\theta_1 + \cdots + \theta_K = 1$

⑤ ∴要引入拉格朗日乘子（带约束的优化问题）

$\Rightarrow \max_{\theta_1, \cdots \theta_K} m_1 \ln\theta_1 + m_2 \ln\theta_2 + \cdots + m_K \ln\theta_K + \lambda(\theta_1 + \cdots + \theta_K - 1)$

分别对 $\theta_1 \cdots \theta_K$ 求导 $\frac{m_1}{\theta_1} + \lambda = 0 \Rightarrow \theta_1 = -\frac{m_1}{\lambda}$ 得 $\frac{m_1}{N}$

$\vdots$

$\frac{m_K}{\theta_K} + \lambda = 0 \Rightarrow \theta_K = -\frac{m_K}{\lambda}$

根据 $\theta_1 + \cdots + \theta_K = 1$ $\quad \therefore$ 即 $\frac{m_1 + \cdots + m_K}{\lambda} = 1 \Rightarrow \lambda = -(m_1 + \cdots + m_K)$

代入

贝叶斯估计 (4.2.3)

例子：抛硬币 $y \in$ [正, 反]

推导：为什么用贝叶斯估计

朴素贝叶斯参数时

$$P_\lambda(Y=c_k)=\frac{\sum_{i=1}^{N} I(y_i=c_k)+\lambda}{N+K\lambda}$$

① 设多项分布时

$y \in \{c_1, \cdots, c_K\}$

$\theta_1, \cdots, \theta_K$

先验分布 → $p(\theta)=\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\theta_1^{\alpha_1-1}\cdots\theta_K^{\alpha_K-1}$

Dirichlet (狄立克雷分布)

类比

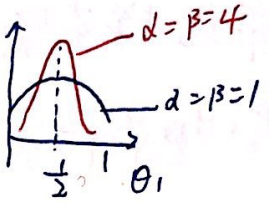
二项分布的先验信息：$p(\theta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_1^{\alpha-1}\theta_2^{\beta-1}$

(Beta分布)

② 直观解释

根据右图，我们倾向认为先验

开始时 $\alpha_1=\alpha_2=\cdots\alpha_K=\alpha$



③ 拿到样本后，算后验

即 $p(\theta|y_1,\cdots,y_N)$ 根据贝叶斯公式

$=\frac{p(\theta_1, y_1, \cdots y_N)}{p(y_1,\cdots,y_N)}$ — 联合分布 / 边缘概率，与θ无关

$\propto p(\theta)\cdot p(y_1,\cdots,y_N|\theta)$ → 前面似然函数已求出

$\propto \theta_1^{\alpha-1}\theta_2^{\alpha-1}\cdots\theta_K^{\alpha-1}\cdot\theta_1^{m_1}\cdot\theta_1^{m_2}\cdots\theta_K^{m_K}$

$\propto \theta_1^{m_1+\alpha-1}\theta_2^{m_2+\alpha-1}\times\cdots\times\theta_K^{m_K+\alpha-1}$ (狄立克雷分布)

即 $p(\theta|y_1,y_2\cdots,y_N)=\frac{\Gamma(m_1+\alpha+m_2+\alpha+\cdots+m_K+\alpha)}{\Gamma(m_1+\alpha)\cdot\Gamma(m_2+\alpha)\times\cdots\times\Gamma(m_K+\alpha)}\times\theta_1^{\cdots}\theta_2^{\cdots}\cdots\theta_K^{\cdots}$

(LDA)

④ 最大化 $\theta_1^{m_1+\alpha-1}\ \theta_2^{m_2+\alpha-1}\ \cdots\ \theta_k^{m_k+\alpha-1}$

$$\Rightarrow \theta_1 = \frac{m_1+\alpha-1}{m_1+\alpha-1+m_2+\alpha-1+\cdots+m_k+\alpha-1}$$

$$= \frac{m_1+\alpha-1}{N+K\alpha-K} \qquad \lambda=\alpha-1$$

$$= \frac{m_1+\alpha-1}{N+K(\alpha-1)}$$

推导：为什么后验概率最大化即期望风险最小化

① 在分类问题中的损失函数

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

（Y：真实，$f(X)$：预测）

把决策函数与条件概率分布联合起来

$$f(x) = \arg \max_{C_k} P(Y = C_k | X)$$

期望风险就是损失函数的期望

$$\min E\, L(Y, f(X))$$

$$= \sum_Y \sum_X [L(Y, f(X)) P(X, Y)]$$

$$= \sum_Y \sum_X L(Y, f(X)) P(Y|X) \cdot P(X)$$

$$= \sum_X \left[\sum_Y L(Y, f(X)) P(Y|X) \right] P(X)$$

最小化上面式子即最小化

$$\min \sum_Y L(Y, f(X)) P(Y|X)$$

$$\Rightarrow \min \sum_{C_k} L(Y = C_k, f(X)) P(Y = C_k | X)$$

事情函数 I：$f(X)$ 为 1

$$\Rightarrow \min \sum_{C_k} I(f(X) \neq C_k) P(Y = C_k | X)$$

$$\Rightarrow \min \sum_{C_k} [1 - I(f(X) = C_k)] P(Y = C_k | X)$$

条件概率之和为 1

$$\Rightarrow \min \sum_{C_k} P(Y = C_k | X) - \sum_{C_k} I(f(X) = C_k) \cdot P(Y = C_k | X)$$

$$\Rightarrow \min \left[1 - \sum_{C_k} I(f(X) = C_k) P(Y = C_k | X) \right]$$

等价于 $\max \sum_{C_k} I(f(X) = C_k) \cdot P(Y = C_k | X)$

$$\max \sum_{C_k} I(f(x)=C_k)\cdot P(Y=C_k|X)$$

★等价于 从 中找到一个 C_k，使 $P(Y=C_k|X)$ 概率最大且使 $I(f(x)=C_k)$ 事情概率为1

因为 $f(x)$ 一次只能取一个值.

即等价于

$$f(x)=\arg\max_{C_k} P(Y=C_k|X)$$

∴得证

总结 朴素贝叶斯法

生成模型

$$P(Y=C_k|X=x)=\frac{P(Y=C_k)\cdot P(X=x|Y=C_k)}{P(X=x)}$$

模型假设-条件独立性（朴素）

$$P(X=x|Y=C_k)=\prod_{i=1}^{n}P(X^{(i)}=x^{(i)}|Y=C_k)$$

预测准则：后验概率最大

$$y=\arg\max_{C_k}P(Y=C_k|X=x)$$