数据建模与分析作业

吴泓鹰　3210101890

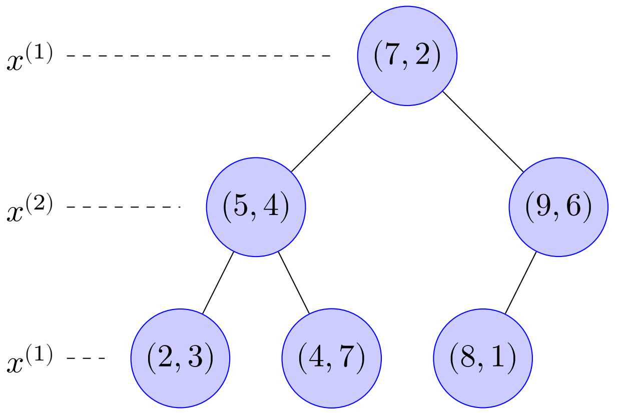
2024 年 3 月 20 日

1　Homework1 感知机

2　Homework2 k-邻近算法, 朴素贝叶斯分类器

2.1　利用例题 3.2 构造的 kd 树 (如下所示) 求点 $x = (3, 4.5)^T$ 的最近邻点.

$$T = \{(2,3)^T, (5,4)^T, (9,6)^T, (4,7)^T, (8,1)^T, (7,2)^T\}$$



首先比较点 $x = (3, 4.5)^T$ 与 $(7, 2)^T$ 的 $x^{(1)}$ 部分，由于 3<7，进入左子树；再与 $(5, 4)^T$ 比较 $x^{(2)}$ 部分，由于 4.5>4，进入右子树，此时到达叶节点，暂将最近邻点记为 $(4, 7)^T$，此时距离为 $d_{min} = d_1 = \sqrt{1^2 + 2.5^2} = 2.69$.

向上返回到 $(5, 4)^T$，由于 $x^{(2)}$ 部分的距离 $0.5 < d_min$，故检查该节点与另一子节点的距离 $d_2 = \sqrt{2^2 + 0.5^2} = 2.06, d_3 = \sqrt{1^2 + 1.5^2} = 1.80$，更新最近邻点为 $(2, 3)^T$，此时距离为 $d_{min} = d_3 = 1.80$；再向上返回到 $(7, 2)^T$，由于 $x^{(1)}$ 部分的距离 $4 > d_{min}$ 且已到达根节点，则搜索结束.

综上所述，最近邻点为 $(2,3)^T$.

2.2 用极大似然估计法推出以下两个朴素贝叶斯法中的概率估计公式.

$$P(Y=c_k)=\frac{\sum_{i=1}^{N} I(y_i=c_k)}{N},\quad k=1,2,...,K$$

$$P(X^{(j)}=a_{jl}|Y=c_k)=\frac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl},y_i=c_k)}{\sum_{i=1}^{N} I(y_i=c_k)};\quad j=1,2,...,n;l=1,2,...,S_j;k=1,2,...,K$$

对第一个公式，设 $P(Y=c_k)=\theta$，进行 N 次实验，有 n 次 $Y=c_k$，即 $\sum_{i=1}^{N} I(y_i=c_k)=n$；则极大似然函数为

$$L(\theta)=P^n(Y=c_k)P^{N-n}(Y\neq c_k)=\theta^n(1-\theta)^{N-n}$$

化为对数形式 $\ln L(\theta)=n\log\theta+(N-n)\log(1-\theta)$，令 $(lnL(\theta))'=n/\theta+(N-n)/(1-\theta)=0$ 即可得到

$$P(Y=c_k)=\theta=\frac{n}{N}=\frac{\sum_{i=1}^{N} I(y_i=c_k)}{N}$$

第二个公式的证明类似，设 $P(X^{(j)}=a_{jl}|Y=c_k)=\theta$，进行 N 次实验，有 n 次 $Y=c_k$，有 m 次 $X^{(j)}=a_{jl},Y=c_k$，即 $\sum_{i=1}^{N} I(y_i=c_k)=n,\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl},y_i=c_k)=m$; 则极大似然函数为

$$L(\theta)=P^m(X^{(j)}=a_{jl}|Y=c_k)P^{n-m}(X^{(j)}\neq a_{jl}|Y=c_k)=\theta^m(1-\theta)^{n-m}$$

化为对数形式 $\ln L(\theta)=m\log\theta+(n-m)\log(1-\theta)$，令 $(lnL(\theta))'=m/\theta+(n-m)/(1-\theta)=0$ 即可得到

$$P(X^{(j)}=a_{jl}|Y=c_k)=\theta=\frac{m}{n}=\frac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl},y_i=c_k)}{\sum_{i=1}^{N} I(y_i=c_k)}$$

2.3 用贝叶斯估计法推出以下两个朴素贝叶斯法中的概率估计公式.

$$P_\lambda(X^{(j)}=a_{jl}|Y=c_k)=\frac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl},y_i=c_k)+\lambda}{\sum_{i=1}^{N} I(y_i=c_k)+S_j\lambda}$$

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

类似上一题，设进行 N 次实验，有 n_k 次 $Y=c_k$，其中有 m_l 次 $X^{(j)}=a_{jl},Y=c_k$，即 $\sum_{i=1}^{N} I(y_i=c_k)=n_k,\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl},y_i=c_k)=m_l$;

先证明第二个公式, 设 $P_\lambda(Y=c_k)=\theta_k$ 满足参数为 α_k 的狄利克雷分布为先验分布, 即

$$P(\theta_1,\theta_2,\dots,\theta_K|\alpha_1,\alpha_2,\dots,\alpha_K)\propto\prod_{i=1}^{K}\theta_i^{\alpha_i-1}$$

又类似上一题的极大似然估计, 可以得到

$$P(N|\theta_1,\theta_2,\dots,\theta_K)=\prod_{i=1}^{K}\theta_i^{n_i}$$

从而做贝叶斯估计有

$$P(\theta_1,\theta_2,\dots,\theta_K|N)\propto P(N|\theta_1,\theta_2,\dots,\theta_K)P(\theta_1,\theta_2,\dots,\theta_K)\propto\prod_{i=1}^{K}\theta_i^{n_i}\prod_{i=1}^{K}\theta_i^{\alpha_i-1}=\prod_{i=1}^{K}\theta_i^{n_i+\alpha_i-1}$$

这表明后验分布也满足狄利克雷分布, 且参数为 $n_k+\alpha_k-1$, 从而 $P_\lambda(Y=c_k)=\theta_k$ 的期望为

$$E(\theta_k)=\frac{n_k+\alpha_k}{\sum_{i=1}^{K}(n_i+\alpha_i)}=\frac{n_k+\alpha_k}{N+\sum_{i=1}^{K}\alpha_i}$$

取 $\alpha_k=\lambda(k=1,2,\dots,K)$ 即可得到所求的公式

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

对于第一个公式的证明也类似, 设 $P_\lambda(X^{(j)}=a_{jl}|Y=c_k)=\theta_l$ 满足参数为 α_l 的狄利克雷分布为先验分布, 即

$$P(\theta_1,\theta_2,\dots,\theta_{S_j}|\alpha_1,\alpha_2,\dots,\alpha_{S_j})\propto\prod_{i=1}^{S_j}\theta_i^{\alpha_i-1}$$

又

$$P(n_k|\theta_1,\theta_2,\dots,\theta_{S_j})=\prod_{i=1}^{S_j}\theta_i^{m_i}$$

从而做贝叶斯估计有

$$P(\theta_1,\theta_2,\dots,\theta_{S_j}|n_k)\propto P(n_k|\theta_1,\theta_2,\dots,\theta_{S_j})P(\theta_1,\theta_2,\dots,\theta_{S_j})\propto\prod_{i=1}^{S_j}\theta_i^{m_i}\prod_{i=1}^{S_j}\theta_i^{\alpha_i-1}=\prod_{i=1}^{S_j}\theta_i^{m_i+\alpha_i-1}$$

这表明后验分布也满足狄利克雷分布, 且参数为 $m_l+\alpha_l-1$, 从而 $P_\lambda(X^{(j)}=a_{jl}|Y=c_k)=\theta_l$ 的期望为

$$E(\theta_l)=\frac{m_l+\alpha_l}{\sum_{i=1}^{S_j}(m_i+\alpha_i)}=\frac{m_l+\alpha_l}{n_k+\sum_{i=1}^{S_j}\alpha_i}$$

取 $\alpha_l=\lambda(l=1,2,\dots,S_j)$ 即可得到所求的公式

$$P_\lambda(X^{(j)}=a_{jl}|Y=c_k)=E(\theta_k)=\frac{\sum_{i=1}^{N}I(x_i^{(j)}=a_{jl},y_i=c_k)+\lambda}{\sum_{i=1}^{N}I(y_i=c_k)+S_j\lambda}$$