第4章 贝叶斯分类器

导引

- 朴素贝叶斯(naive Bayes) 法是基于贝叶斯定理与特征条件独立假设的分类方法
- 对于给定的训练数据集
  - 模型：首先基于特征条件独立假设，学习输入输出的联合概率分布
    - 因采用的原理是贝叶斯定理，核心的判断依据是后验概率来判断，计算转换为
      - 先验概率$P(y = c_k)$
      - 条件概率$P(X = x|Y = c_k)$
  - 预测：基于此模型【先验概率，条件概率】，对给定的输入，利用贝叶斯定理求出后验概率最大的输出
- 朴素贝叶斯应用于分类时
  - 使用最大后验概率

概率(回顾)

定义：Ω为试验E的样本空间，$B_1, B_2, \cdots, B_n$为E的一组事件，若

$$B_i \cap B_j = \emptyset, i, j = 1,2, \cdots, n$$

$$B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$$

则称$B_1, B_2, \cdots, B_n$为样本空间的一个分划

- 全概公式

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n) = \sum_{i=1}^{n} P(B)P(A|B_i)$$

- 乘法规则

$$P(x, y) = P(x)P(y \mid x)$$

联合概率=边沿概率*条件概率

- 贝叶斯定理

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

1 朴素贝叶斯法的学习与分类

贝叶斯公式，以及条件独立性假设(特征的各个分量独立)

朴素贝叶斯法将实例分到后验概率最大的类中，等价于期望风险最小化

基本方法

输入空间：$\mathcal{X} \subseteq \mathrm{R}^n$，输出空间为类标记集合：$\mathcal{Y} = \{c_1, c_1, \ldots, c_k\}$

联合概率分布$P(X, Y)$

训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$，由联合概率分布独立同分布产生

朴素贝叶斯，通过训练数据集学习联合概率分布$P(X, Y)$

先验概率分布

$$P(y = c_k), k = 1,2, \ldots, K$$

条件概率分布

$$P(X = x|Y = c_k) = P\left(X^{(1)} = x^{(1)}, \ldots, X^{(n)} = x^{(n)} \middle| Y = c_k\right), k = 1,2, \ldots, K$$

朴素贝叶斯：条件独立性假设(特征的各个分量独立)

$$P(X = x|Y = c_k) = P\left(X^{(1)} = x^{(1)}, \ldots, X^{(n)} = x^{(n)} \middle| Y = c_k\right) = \prod_{j=1}^{\mathrm{n}} P\left(X^{(j)} = x^{(j)} \middle| Y = c_k\right)$$

朴素贝叶斯方法用于分类

朴素贝叶斯：条件独立性假设(特征的各个分量独立，朴素得名由来)

$$P(X=x|Y=c_k)=P\left(X^{(1)}=x^{(1)},\dots,X^{(n)}=x^{(n)}\middle|Y=c_k\right)=\prod_{j=1}^{\mathrm{n}}P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)$$

对给定的输入x，通过学习到的模型计算后验概率分布$P(X=x|Y=c_k)$，将后验概率最大的类作为x的类输出

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

$$=\frac{P(Y=c_k)\prod_{j=1}^{\mathrm{n}}P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)}{\sum_k P(Y=c_k)\prod_{j=1}^{\mathrm{n}}P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)}$$

朴素贝叶斯方法用于分类

朴素贝叶斯分类器(最大后验概率)：

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

$$y = f(x) = \arg\max_{c_k} P(Y = c_k | X = x) = \arg\max_{c_k} \frac{P(Y=c_k) \prod_{j=1}^{n} P\left(X^{(j)} = x^{(j)} \middle| Y = c_k\right)}{\sum_k P(Y=c_k) \prod_{j=1}^{n} P\left(X^{(j)} = x^{(j)} \middle| Y = c_k\right)}$$

对于所有c_k分母不变，故上式变为

$$y = f(x) = \arg\max_{c_k} P(Y = c_k) \prod_{j=1}^{n} P\left(X^{(j)} = x^{(j)} \middle| Y = c_k\right)$$

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，等价于期望风险最小化(**结论**)

【证明】选择0-1损失$L(Y,f(X))$，$f(X)$为决策函数，期望风险函$R_{\exp(f)}$

$$R_{\exp(f)} = E\left[L(Y,f(X))\right] = E_X\left(\sum_Y L(c_k,f(X))\,P(Y|X)\right)$$

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

$R_{\exp(f)}$最小化：对$E_X(\quad)$中的$\sum_{k=1}^{K}\left[L(c_k,f(X))P(c_k|X)\right]$的每一个$X=x$都最小化

$$f(x) = \arg\min_{y\in\mathcal{Y}}\sum_{k=1}^{K}[L(c_k,y)P(c_k|X=x)] = \arg\min_{y\in\mathcal{Y}}\sum_{k=1}^{K}P(y\neq c_k|X=x)$$

$$= \arg\min_{y\in\mathcal{Y}}\left(1-P(y=c_k|X=x)\right) = \arg\max_{y\in\mathcal{Y}}P(y=c_k|X=x)$$

即最大化后验概率。

2 朴素贝叶斯法的参数估计

古典概率公式计算经验概率

朴素贝叶斯法的参数估计

朴素贝叶斯法中，学习意味着估计$P(y=c_k)$和$P(X=x|Y=c_k)$

先验概率$P(y=c_k)$的极大似然估计

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

设第j个特征$x^{(j)}$可能取值的集合为：$\{a_{j1},a_{j2},\dots,a_{jS_j}\}$,

条件概率$P(X^{(j)}=a_{jl}|Y=c_k)$的极大似然估计

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

学习与分类算法(Naïve Bayes Algorithm)

输入：训练集$T=\{(x_1,y_1),\dots,(x_N,y_N)\}$，其中$x_i=\left(x_i^{(1)},\cdots,x_i^{(n)}\right)^{\mathrm{T}}$，$x_i^{(j)}\in\left\{a_{j1},\dots,a_{jS_j}\right\}$

输出：x的分类

1. 计算先验概率和条件概率

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

$$P\left(X^{(j)}=a_{jl}\mid Y=c_k\right)=\frac{\sum_{i=1}^{N}I\left(x_i^{(j)}=a_{jl},y_i=c_k\right)}{\sum_{i=1}^{N}I(y_i=c_k)}$$

$$j=1,2,\dots,n;l=1,2,\dots,S_j;k=1,2,\dots,K$$

2. 对于给定的实例$x=\left(x^{(1)},x^{(2)},\cdots,x^{(i)},\cdots,x^{(n)}\right)^{\mathrm{T}}$，计算各个类别的

$$P(Y=c_k)\prod_{j=1}^{\mathrm{n}}P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)$$

3. 确定x的类别：$y=\arg\max_{c_k}P(Y=c_k)\prod_{j=1}^{\mathrm{n}}P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)$

例子

- 4.1
- 4.2

贝叶斯估计

用极大似然估计可能会出现所要估计的概率值为0的情况，从而影响后验概率的计算结果，使分类产生偏差，可采用贝叶斯估计解决

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

式中$\lambda\geqslant 0$。等价于在随机变量各个取值的频数上赋予一个正数 $\lambda>0$

当$\lambda=0$, 极大似然估计;$\lambda=1$, 称为拉普拉斯平滑(Laplacian smoothing)

对任何$l=1,\cdots,S_j,k=1,2,\cdots,K$， 有

$$P_\lambda\left(X^{(j)}=a_{jl}\mid Y=c_k\right)>0$$

$$\sum_{l=1}^{S_j}P_\lambda\left(X^{(j)}=a_{jl}\mid Y=c_k\right)=1$$

表明确为一种概率分布。同样， 先验概率的贝叶斯估计为

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