朴素贝叶斯方法的学习

姓名

2021 年 10 月 28 日

目录

摘要

本文主要介绍了监督学习中常用到的朴素贝叶斯方法。

1 监督学习

监督学习是指从标注数据中学习预测模型的机器学习问题。标注数据表示输入输出的对应关系，预测模型对给定的输入产生相应的输出。监督学习的本质是学习输入到输出的映射的统计规律。

1.1 变量空间

在监督学习中，将输入和输出所有可能取值的集合分别称为输入空间和输出空间。输入和输出空间可以是有限元素的集合，也可以是整个欧式空间。输入和输出空间可以是同一个空间，也可以是不同的空间；但通常输出空间远远小于输入空间。

1.2 联合概率分布

监督学习假设输入与输出的随机变量 X 和 Y 遵循联合概率分布 $P(X,Y)$。$P(X,Y)$ 表示分布函数，或分布密度函数。注意在学习过程中，假定这一联合概率分布存在，但对学习系统来说，联合概率分布的具体定义是未知的。训练数据与测试数据被看作是依联合概率分布 $P(X,Y)$ 独立同分布产生的。统计学习假设数据存在一定的统计规律，X 和 Y 具有联合概率分布就是监督学习关于数据的基本假设。

1.3 假设空间

监督学习的目的在于学习一个由输入到输出的映射，这一映射由模型来表示。换句话说，学习的目的就在于找到最好的这样的模型。模型属于由输入空间到输出空间的映射的集合，这个集合就是假设空间。假设空间的确定意味着学习范围的确定。监督学习的模型

1.4 问题的形式化

监督学习利用训练数据集学习一个模型，再用模型对测试样本集进行预测。由于在这个过程中需要标注的训练数据集，而标注的训练数据集往往是人工给出的，所以称为监督学习。监督学习氛围学习和预测两个过程，有学习系统与预测系统完成，可以用图 1.1 来描述。

首先给定一个训练数据集

$$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$$

其中 $(x_i, y_i), i = 1, 2, 3, \ldots, N$，成为样本或样本点。$x_i \in \mathbf{X} \subseteq \mathbf{R^n}$ 是输入的观测值，也称为输入或实例，$y_i \in Y$ 是输出的观测值，也称为输出。

监督学习分为学习和预测两个过程，由学习系统与预测系统完成。在学习过程中，学习系统利用给定的训练数据集，通过学习得到一个模型，表示为条件概率分布 $\widehat{P}(Y|X)$ 或决策函数 $Y = \widehat{f}(X)$。条件概率分布 $\widehat{P}(Y|X)$ 或决策函数 $Y = \widehat{f}(X)$ 描述输入与输出随机变量之间的映射关系。在预测过程中，预测系统对于给定的测试样本集中的输入 x_{N+1}，由模型 $y_{N+1} = \arg\max\limits_{y} \widehat{P}(y|x_{N+1})$ 或 $y_{N+1} = \widehat{f}(x_{N+1})$ 给出相应的输出 y_{N+1}。

2 朴素贝叶斯方法

2.1 基本方法

设输入空间 $\mathbf{X} \subseteq \mathbf{R^n}$ 为 n 维向量的集合，输出空间为类标记集合 $Y = \{(c_1), (c_2), \ldots, (c_K)\}$。输入为特征向量 $x \in \mathbf{X}$，输出为类标记 $y \in \mathbf{Y}$。X 是定义在输入空间 $\mathbf{X}$ 上的随机向量，Y 是定义在输出空间 $\mathbf{Y}$ 上的随机变量。$P(X, Y)$ 是 X 和 Y 的联合概率分布。训练数据集

$$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$$

由 $P(X, Y)$ 独立同分布产生。

朴素贝叶斯法通过训练数据集学习联合概率分布 $P(X, Y)$。具体地，学习以下先验概率分布及条件概率分布。先验概率分布

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

条件概率分布

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

于是学习到联合概率分布 $P(X, Y)$。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。事实上，假设 $x^{(j)}$ 可取值有 S_j 个，j=1,2,...,n，Y 可取值有 K 个，那么参数个数为 $K\prod\limits_{j=1}^{n} S_j$ 。

朴素贝叶斯法对条件概率分布做了条件独立性的假设。由于这是一个较强的假设，朴素贝叶斯法也由此得名。具体的，条件独立性假设是

$$\begin{aligned} P(X=x|Y=c_k) &= P(X^{(1)}=x^{(1)},\dots,X^{(n)}=x^{(n)}|Y=c_k) \\ &= \prod_{j=1}^{n} P(X^{(j)}=x^{(j)}|Y=c_k) \end{aligned} \tag{2}$$

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。

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。将后验概率最大的类作为的类输出。后验概率计算根据贝叶斯定理进行:

$$P(Y=c_k|X=x)=\frac{P(X=x|Y=c_k)P(Y=c_k)}{\sum_k P(X=x|Y=c_k)P(Y=c_k)} \tag{3}$$

将式 (2) 代入式 (3)，有

$$P(Y=c_k|X=x)=\frac{P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_k P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}, k=1,2,\dots,K \tag{4}$$

这是朴素贝叶斯法的分类的基本公式。于是，朴素贝叶斯分类器可表示为

$$y=f(x)=\arg\max_y P(Y=c_k|X=x)=\frac{P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_k P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)} \tag{5}$$

注意到，在式 (5) 中分母对所有 c_k 都是相同的，所以，

$$y=\arg\max_{c_k} P(Y=c_k)=\prod_j P(X^{(j)}=x^{(j)}|Y=c_k) \tag{6}$$

2.2 极大似然估计

在朴素贝叶斯法中，学习意味着估计 $P(Y=c_k)$ 和 $P(X^{(j)}=x^{(j)}|Y=c_k)$ 。可以应用极大似然估计法估计相应的概率。先验概率 $P(Y=c_k)$ 的极大似然估计是

$$PY=c_k)=\frac{\sum_{i=1}^{N} I(y_j=c_k)}{N}, k=1,2,\dots,K \tag{7}$$

设第 j 个特征 $x^{(j)}$ 可能的取值的集合为 $\{a_{j1},a_{j2},\dots,a_{jS_j}\}$，条件概率 $P(X^{(j)}=x^{(j)}|Y=c_k)$ 的极大似然估计是

$$P(X^{(j)}=x^{(j)}|Y=c_k)=\frac{\sum_{i=1}^{N} I(x^{(j)}=a_{jl},y_j=c_k)}{\sum_{i=1}^{N} I(y^{(j)}=c_k)} \tag{8}$$

式中 $x_i^{(j)}$，是第 i 个样本的第 j 个特征；a_{jl} 是第 j 个特征可能取的第 l 个值；I 为指示函数。

2.3 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。这时会影响到后验概率的计算结果，使分类产生偏差。解决这一问题的方法是采用贝叶斯估计。具体地，条件概率的贝叶斯估计是

$$P_\lambda(X^{(j)}=x^{(j)}|Y=c_k)=\frac{\sum_{i=1}^{N}I(x^{(j)}=a_{jl},y_j=c_k)+\lambda}{\sum_{i=1}^{N}I(y^{(j)}=c_k)+S_j\lambda} \tag{9}$$

式中 $\lambda\geq 0$。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 $\lambda=0$ 时就是极大似然估计。常取 $\lambda=1$，这时称为拉普拉斯平滑。显然，对任何 $l=1,2,\ldots,S_j$，$k=1,2,\ldots,K$，有

$$P_\lambda(X^{(j)}=x^{(j)}|Y=c_k)>0$$

$$\sum_{i=1}^{S_j}P(X^{(j)}=a_{jl},Y=c_k)=1$$

表明式 (9) 确为一种概率分布。同样，先验概率的贝叶斯估计是

$$P_\lambda(Y=c_k)=\frac{\sum_{i=1}^{N}I(y_j=c_k)+\lambda}{N+K\lambda} \tag{10}$$

References

[1] 统计学习方法（第 2 版）