Naive Bayes Algorithm

Shen Hengheng

2017

本节主要讲解三个基本算法的之一: **朴素贝叶斯算法**, 该算法主要是依据**贝叶斯规则/公式/推理**，它的主要作用在于，它和其他将要讲述的 k **近邻算法**，**感知机算法**都一样，你的算法设计出来之后，至少你的算法的 **performance** 要比这三类算法的效果要好! 另外贝叶斯学习算法基于一些统计上的假设而成立的, 因此贝叶斯学习算法又称为统计学习的算法必须学习的算法. 本节主要有以下部分内容:

- Bayes' Rule
- Applying Baye' s Rule to Classification
- The Posterior Probability :$\mathrm{P}(C|x) = \frac{\mathrm{P}(C)\mathrm{P}(x|C)}{\mathrm{P}(x)}$
- Extend to Multi-class classification
- Naive Bayes for Classification
- Zero-frequency Problem

0.1 贝叶斯定理/公式/规则

假设 $\{B_1, B_2, ..., B_k\}$ 是 S 集合的**划分**，并且对于 $i = 0, 1, ..., k$ 有 $\mathrm{P}(B_i) > 0$，那么

$$\mathrm{P}(B_j|A) = \frac{\mathrm{P}(A|B_j)\mathrm{P}(B_j)}{\sum_{i=1}^{k}\mathrm{P}(A|B_i)\mathrm{P}(B_i)}$$

0.2 在分类问题上应用贝叶斯分类器

0.2.1 例子：信用卡评估：高风险/低风险

信用卡评估是这样的一个过程，银行通过对用户的历史交易信息来进行选择的，如果该用户在过去的交易中按时缴纳了本金和利息，并且银行从中获益，那么这些用户是属于**低风险的**，如果他们没有按时缴纳本金和利息或者拒绝还款的话，这类用户比银行定义为**高风险的用户**.

我们现在的任务就是来学习如何对用户进行分类**高风险** vs **低风险**. 那么我们现在有以下用户的每年的收入信息和消费信息，我们们用两个**随机变量** X_1 和 X_2 来表示. 用户的信用度可以用伯努利随机变量 $\mathrm{C} = 1$ 和 $\mathrm{C} = 0$ 来表示。其中 $C = 0$ 表示这是一个低风险的用户，$C = 0$ 表示这是一个高风险的用户。

当一个新的用户来进行贷款时，他提供他的每年的收入和消费，即 $X_1 = x_1$ 和 $X_2 = x_2$. 如果我们知道信用度 C 的概率受限于观测值 $X = [x_1, x_2]$，并且有

$$\text{if } \mathrm{P}(C = 1|[x_1, x_2]) > 0.5, \textit{ then } \mathrm{C} = 1$$

$$\text{if } \mathrm{P}(C = 1|[x_1, x_2]) \leq 0.5, \textit{ then } \mathrm{C} = 0$$

我们估计的误差建立在 $1-max\mathrm{P}(C=1|[x_1,x_2]),\mathrm{P}(C=0|[x_1,x_2])<0.5$，一般地，有 $P(\mathrm{C}=1|[x_1,x_2])+\mathrm{P}(C=0|[x_1,x_2])=1$。

0.2.2 后验概率 $\mathrm{P}(C|x)=\frac{\mathrm{P}(C)\mathrm{P}(x|C)}{\mathrm{P}(x)}$

其中，对于标题的概率公式又叫做贝叶斯公式，它主要思想是利用先验来估计后验，即 $P(C=1)$ 叫做 C=1 的先验概率，比如信用卡评估问题上，我们已经有了大量的历史客户数据，$P(C=1)$ 表示在这些数据中高风险的人占总人数的比例，肯定地，这是我们知道的！所以被称作先验概率。

1. $P(x|C)$ 被称作类的似然，它是条件概率，意思为一个事件在已知属于 C 类的情况下，事件 x 发生的概率,
2. $P(x)$ 表示事件 x 发生的概率，不管它是正类还是负类,
3. $P(C=0|x)+P(C=1|x)=1$，来源于事件的划分。

另外还有以下性质:

1. $P(X1,X2)$ 表示随机变量 X_1 和 X_2 的联合概率
2. 在假设前提下,两个随机变量是条件概率独立,有 $P(X1,X2|C)=P(X1|C)P(X2|C)$，这是贝叶斯规则重要的假设。尽管假非常简单，但是在实际应用中十分关键。

现在我们将问题扩展到多类分类问题上，即 $C_i,\ i=1,2,...,K$. 比如，在数字识别问题上，输入是一个 2 值化的图像，目标总共有 10 类，即 0-9，我们可以认为这 K 类看作是输入空间的一个划分。那么贝叶斯分类器是将最高的后验概率的类看作是目标类，也就是 $\mathrm{P}(C_i|x)=\max_k\mathrm{P}(C_k|x)$. 同样地，对于多类分类问题，我们可以将所有的属性/特征看作是一个随机变量，那么对于贝叶斯分类器将该样本分为 No.1 类的概率为 $Pr(y=1|x)=Pr(y=1|\mathbf{X}_1=x_1,\mathbf{X}_2=x_2,...,\mathbf{X}_n=x_n)$.

但是朴素贝叶斯分类七有两个不合理的假设:

1. 每个属性/特征的重要性是一样的
2. 所有属性/特征都是条件概率独立的，有

$$Pr(y=1|x)=\frac{1}{Pr(\mathbf{X}=x)}\Pi_{i=1}^{n}Pr(\mathbf{X}_i=x_i|y=1)$$

0.3 天气数据的例子