高斯朴素贝叶斯算法实现数据分类

肖劲松 何立功 * 舒地灵 农高海
（百色学院信息工程学院，广西 百色 533000）

摘要：朴素贝叶斯分类是机器学习中非常重要的一种分类算法，基于贝叶斯定理和特征条件独立假设，可以用于文本分类、垃圾邮件过滤、信息检索等应用场景。本文较为详细的给出了高斯朴素贝叶斯分类的理论过程，并试着采用鸢尾花数据集进行模型训练与分类。根据实验结果，表明该算法能够实现数据分类，分类准确度较高，且分类准确与数据集特征，训练模型数据量大小等因素有关。随着数据特征的增加，或者训练集数据增加，分类准确率有所提高。

关键词：高斯朴素贝叶斯；鸢尾花数据集；训练与分类

朴素贝叶斯算法是一种基于贝叶斯定理的分类算法，因其简单、高效而广受欢迎。在处理大规模数据集时，朴素贝叶斯算法具有运行速度快、内存消耗小的优点。此外，该算法还具有良好的可解释性，能够提供每个类别的概率和特征之间的关系，使得用户可以更直观地理解分类结果。朴素贝叶斯算法是一种概率分类算法，其基本思想是，对于给定的数据集，首先根据特征建立分类模型，然后利用贝叶斯定理计算每个类别的概率，最后将样本分配给概率最大的类别。核心在于假设数据以及特征之间相互独立，从而简化了计算和解释。本文旨在探讨朴素贝叶斯算法的原理、实现方法和应用场景，旨在帮助读者深入理解朴素贝叶斯算法，并提供相关应用案例供参考。

一、朴素贝叶斯算法

由贝叶斯定理

$$P(A \mid B)=\frac{P(B \mid A)P(A)}{P(B)} \qquad (1)$$

知，后验概率 $P(A \mid B)$ 是通过似然概率 $P(B \mid A)$ 与先验概率 $P(A)$ 的乘积，再除以全概率

$$P(B)=\sum_{j=1}^{N}P(B \mid A_j)P(A_j) \qquad (2)$$

得到的，其中式通常为一常数。因此，在中可知 $P(A \mid B)$ 正比与 $P(B \mid A)*P(A)$，即

$$P(A \mid B) \propto P(B \mid A)P(A) \qquad (3)$$

在概率论上 $P(A \mid B)$ 的意思是“在 B 发生的情况下，A 发生的概率”，二对于文本分类则是通过特征来判断属于哪一种类别，可以转换成：A 为分类标签，B 为一系列的特征属性。实际计算的过程中便是每个样本在当前的特征值为 B 的条件下，属于类别 A 的概率，然后从所有的概率中取最大值便认为是当前的分类。因此，只需要计算出 $P(B \mid A)$ 与 $P(A)$ 即可。

假设训练集为 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，其中 $x_i=(x_i^1,x_i^2,\cdots,x_i^n)$ 为 n 维特征向量，x_i^j 指的是第 i 个样本的第 j 个特征值，$y_i=(c_1,c_2\cdots,c_K)$ 为类标签。且由“朴素”二字知数据之间相互严格独立。则似然概率

$$\begin{aligned}P(X^j=x^j \mid Y=c_k)&=P(X^j=x_1^j,X^j=x_2^j,\cdots,X^j=x_n^j \mid Y=c_k)\\&=\prod_{i=1}^{n}P(X^j=x_i^j \mid Y=c_k)\end{aligned} \qquad (4)$$

先验概率

$$P(Y=c_k)=\frac{\sum_{i=1}^{N}(y_i=c_k)}{m},\qquad k=1,2,\cdots,K \qquad (5)$$

结合式，可得分类后验概率

$$P(Y=c_k \mid X^j=x^j)=P(Y=c_k)\prod_{i=1}^{n}P(X^j=x_i^j \mid Y=c_k) \qquad (6)$$

进而求最大值为

$$\widehat{c_k}=\arg\max_{c_k}P(Y=c_k \mid X^j=x^j)=\arg\max_{c_k}P(Y=c_k)\prod_{i=1}^{n}P(X^j=x_i^j \mid Y=c_k) \qquad (7)$$

当 $\widehat{c_k}$ 取最大值时，便可知类别为 c_k。

二、高斯朴素贝叶斯

所谓高斯贝叶斯指的是，假定样本每个特征维度的条件概率均服从高斯分布，进而再根据贝叶斯公式计算新样本在某个特征分布下，其属于各个类别的后验概率，最后通过极大化后验概率来确定样本的所属类别。

假定数据样本在各个类别下，每个特征变量的条件概率均服从高斯分布，其概率密度函数为

$$P(X^j=x_i^j \mid Y=c_k)=\frac{1}{\sqrt{2\pi}\sigma_{cj}}\exp\left(-\frac{(x_i^j-\mu_{cj})^2}{2\sigma_{cj}^2}\right) \qquad (8)$$

σ_{cj}^2 和 μ_{cj} 分布表示在类别 $Y=c_k$ 下特征 x_i^j 对应的标准差和期望。在计算得到每个特征维度的似然概率后，再进行极大化后验概率计算。即将公式带入公式后，得

$$\begin{aligned}\widehat{c_k}&=\arg\max_{c_k}\log\left\{P(Y=c_k)\prod_{i=1}^{n}P(X^j=x_i^j \mid Y=c_k)\right\}\\&=\arg\max_{c_k}\log\left\{P(Y=c_k)\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_{cj}}\exp\left(-\frac{(x_i^j-\mu_{cj})^2}{2\sigma_{cj}^2}\right)\right\} \qquad (9)\\&=\arg\max_{c_k}\left\{\log\left[P(Y=c_k)\right]-\frac{1}{2}\sum_{i=1}^{n}\left[\log(2\pi\sigma_{cj})+\frac{(x_i^j-\mu_{cj})^2}{\sigma_{cj}^2}\right]\right\}\end{aligned}$$

由此便可求得高斯朴素贝叶斯最大后验概率，也就是完成最终分类。

三、实验仿真

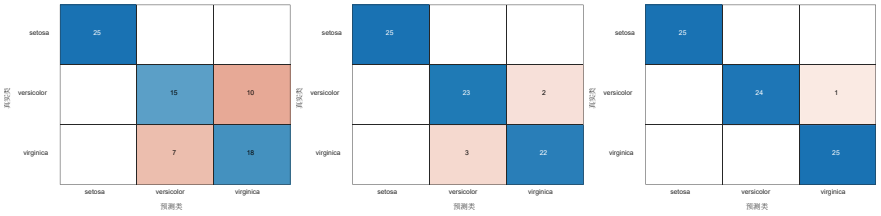
采用 MATLAB 编程语言，应用软件自带的鸢尾花数据集，该数据集共有 150 个数据，分为 setosa、versicolor 以及 virginica 三类，每类 50 个数据，每个数据包含 4 个属性，分别为花萼长度、花萼宽度、花瓣长度及花瓣宽度。

在实验的过程中，分三组实验进行，每组实验训练集数据与测试集数据个数如表 1：

表 1 分组实验数据个数情况

	第一组	第二组	第三组
训练集数据个数	75	105	135
测试集数据个数	75	45	15

每组实验又根据取不同的数据集特征，分 a、b、c 三次实验进行：*a.* 取花萼长度和花萼宽度；*b.* 取花瓣长度及花瓣宽度；*c.* 取花萼长度、花萼宽度、花瓣长度及花瓣宽度。每组仿真分类结果给出相应的混淆矩阵，混淆矩阵对角线上数据为分类正确数据，其他数据为分类错误数据，其中横向为真实类，纵向为预测类。三组实验分类结果分别如图 1、图 2、图 3。

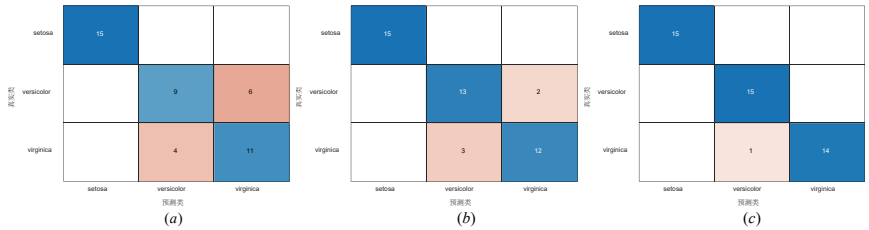


(a) (b) (c)

图 1 第一组仿真实验分类结果混淆矩阵

(a) 取花萼特征；(b) 取花瓣特征；(c) 取花萼与花瓣特征

从图 1 的实验结果可以看出，使用花瓣的特征进行分类的结果，较好与使用花萼的特征分类结果。而同时取花萼和花瓣的特征进行分类，要比单独取花萼或者花瓣的特征进行分类的结果更好。分类错误主要是 versicolor 和 virginica 两类，最差的情况是 10 个 versicolor 类，被预测为 virginica，7 个 virginica 类被预测为 versicolor 类。最好的情况只有 1 个 versicolor 类被预测为 virginica 类。

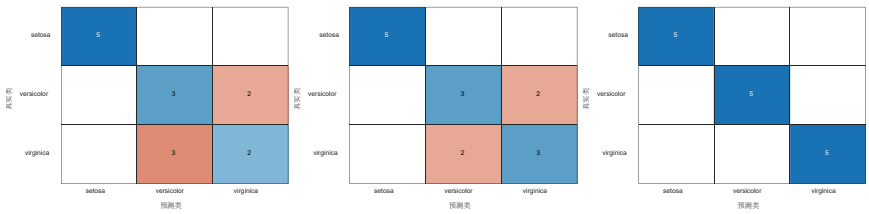


(a) (b) (c)

图 2 第二组仿真实验分类结果混淆矩阵

(a) 取花萼特征；(b) 取花瓣特征；(c) 取花萼与花瓣特征

从图 2 可以看到与图 1 类似的效果。但通过对比两次实验的结果 (a)，发现随着训练集的增加，分类的准确率明显有所提高。此时最差的情况是 6 个 versicolor 类，被预测为 virginica，4 个 virginica 类被预测为 versicolor 类。最好的情况只有 1 个 virginica 类被预测为 versicolor 类。



(a) (b) (c)

图 3 第三组仿真实验分类结果混淆矩阵

(a) 取花萼特征；(b) 取花瓣特征；(c) 取花萼与花瓣特征

由图 3 实验结果对比前两次实验，发现分类的准确率更加高了，特别是本次实验结果 (c)，分类结果完全正确。此时最差的情况是 2 个 versicolor 类，被预测为 virginica，3 个 virginica 类被预测为 versicolor 类。

分类结果准确率如表 2：

表 2 分类结果准确率情况

	第一组			第二组			第三组		
	a	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
准确度 %	77.33	93.33	98.67	77.78	88.89	97.78	66.67	73.33	100

根据表 2，分类准确率有以下两方面的规律：首先，每组实验在数据特征方面，取花瓣特征或者同时取花瓣与花萼特征进行分类的准确率，整体高于只取花萼特征分类。分类准确率最高的是同时取花瓣与花萼特征。说明特征属性越多，分类效果越好，也符合了解特征越多，判断越准确的直观认识。其次，在模型训练过程中，训练数据越多，分类结果也越准确。

四、结论

通过对高斯朴素贝叶斯分类算法的原理进行详细的剖析和实验验证，可以看出，朴素贝叶斯可以进行分类，且分类准确度较高，当随着特征类型的增加，或者训练数据的增加，分类准确度越来越高，在一些情况下甚至接近百分百。但是本文测试数据有限，不一定真时的反应分类情况。

整体上基于高斯模型的朴素贝叶斯算法相对简便，对分类结果的解释更容易理解。对小规模的数据反映好，时间复杂度小，计算速度快，准确率较高，能处理较多分类项目，适合增量式训练，随着样本量的逐渐增大贝叶斯会提高准确性。

参考文献：

[1] 郭秀娟，李庆凯，孟庆楠等 . 基于朴素贝叶斯算法分析鸢尾花数据集分类 [J]. 工业和信息化教育，2022（06）：82-84+91.

[2] 何伟 . 基于朴素贝叶斯的文本分类算法研究 [D]. 南京：南京邮电大学，2020.

[3] 陶阳明 . 经典人工智能算法综述 [J]. 软件导刊，2020，19（3）：276-280.

项目名称：新工科背景下以 OBE 理念为导向电气类专业多元化创新型人才培养模式的探索与实践；编号：2022JGA327

⋆通信作者：何立功，百色学院