计算机模拟实验报告

武子越 3170104155

1. 实验内容

构建贝叶斯分类器，针对鸢尾花的诸多参数，将 iris.csv 的数据进行分类，其中 60%的数据用于进行先验统计，即进行训练集的构建，40%的数据进行测试。

2. 算法原理与实验基本思路

1） 训练集与测试集的划分：首先将数据按行打乱，将前面的 60%用作训练集计算先验，后面 40%用于精度的测试。

2） 计算每个 target 的先验概率（即这种类别在数据集中出现的概率），利用贝叶斯定理：$\mathrm{p}(\mathrm{C_i}|X) = P(X|C_i)P(C_i)/P(X)$，分别计算分到每个类别的后验概率。

3） 对于每个新的数据点，选择具有最大后验概率的类别。

$$\mathrm{p}(\mathrm{C_i}|\mathrm{X}) \propto \mathrm{p}(\mathrm{X}|\mathrm{C_i})p(C_i)\ \ i = 1,2,3$$

$$\arg\max_i\ p(C_i|X)\ \ i = 1,2,3$$

其中$\mathrm{p}(\mathrm{X}|\mathrm{C_i}) = \prod_{k=1}^{n} p(x_k|C_i)$，这里 X 是特征向量，X=(X1,X2,X3,X4)对应花的四个特征。

4） 将测试集中的数据分类结果与实际类别进行对比，统计分类正确的数量并计算正确率。

$$\mathrm{cre} = \mathrm{sum(right)/sum(all)}$$

3. 实验结果分析

函数输出结果：Accuracy: 0.5833

我们可以看到，这个贝叶斯分类器的分类准确率为 0.5833，并不是特别高。可能会有如下几种原因：

1） 对于类别的划分

数据本质上是数值型特征而不是类别特征，这里构建分类器的时候将每 0.1 当成一类，而现实中，每种花花瓣或花萼的长度可能服从一个正态分布，并不能完全按照数值上的等分划分开来。另外 0.1 的划分结果在处理上比较方便，不过由于训练集中只有 90 个数据，因此实际每个区间的样本量较少，计算先验概率时受到数据点的影响较大。

2） 模型本身的问题

朴素贝叶斯分类器的假设默认各个特征之间是独立的，然而实际上花瓣和花萼的大小可能存在着某种相关关系，在植物学的分类中这些特征不一定能够当做独立特征进行处理。而贝叶斯分类器默认特征独立，因此可能会与实际的效果产生一定的偏差。

4. **改变训练集与测试集划分**，并分别进行十次测试，统计出的正确率变化如下：

编号	先验训练集比例	测试集比例	平均正确率
1	50%	50%	0.4667
2	60%	40%	0.5920
3	70%	30%	0.6659
4	80%	20%	0.6231

可以看出，训练集比例在 70%左右的时候预测的正确率是比较高的，当用于计算先验的数据过少时，每个类别往往会有很少的数据甚至没有数据（导致概率为 0），对于最终预测结果影响较大，而测试集比例过小则难以准确衡量算法的正确率。