数据挖掘实验报告

成绩:

姓名	沈建鑫	学号	18051624
实验时间	11.13	实验地点	四教北 404
实验名称	分类技术		

一、实验过程与问题

请列举实验过程中遇到的问题与对应的解决方案。

1.在虚拟环境上安装注册 Graphviz 时要重启一下 anaconda 才能生效。

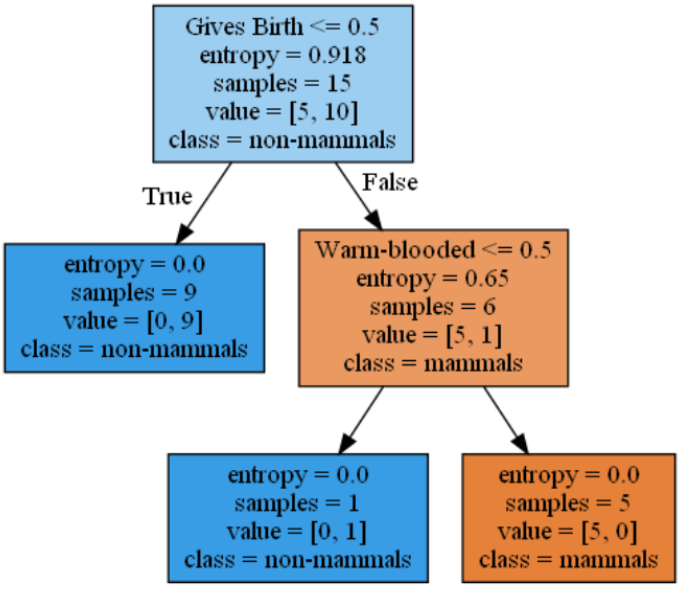
2.在查阅 numpy 文档找生成不同分布数据的函数时没有仔细看，导致一开始输出的形状都不对，经过几次试验之后得到了正确结果。

二、实验结果与分析

请回答实验代码文件（ipynb 文件）中的思考题

1. 观察使用不同配置的决策树的运行结果并进行分析

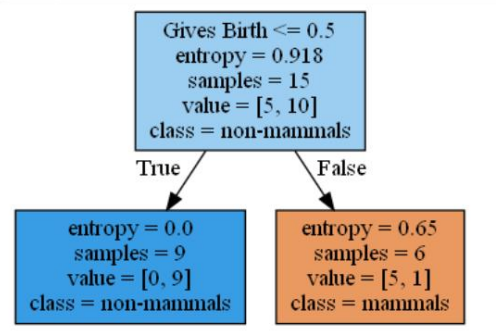
答：max_depth=3



	Name	Predicted Class
0	gila monster	non-mammals
1	platypus	non-mammals
2	owl	non-mammals
3	dolphin	mammals

正确率 0.75

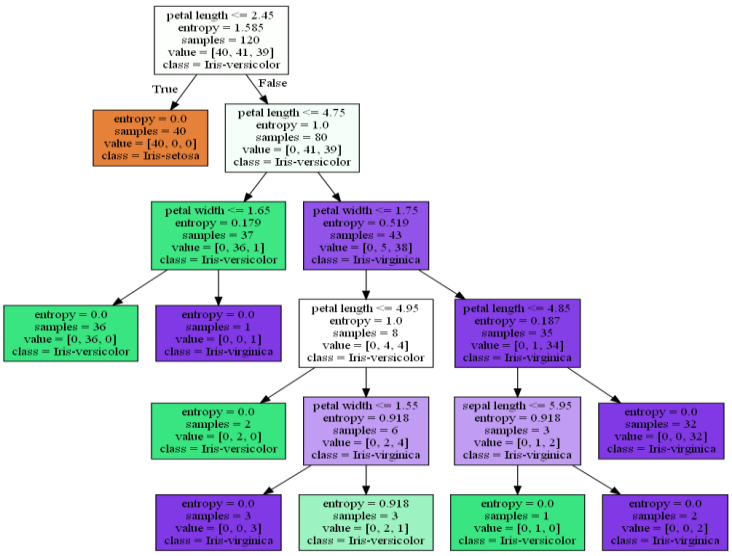
Max_depth=1:



分析：实验尝试了 max_depth=1，2，3，4 的情况，发现除了 1 的决策树和其他不一样，其他结果都一样，并且在决策树上，max_depth 为 1 的树其实就是其他情况的一个子树。可见，该数据和标签条件下的决策树最优深度即为 2，在第一步时通过 Gives Birth 来进行划分，在第二步时通过 Warm-blooded 来进行划分。

2. 观察决策树在鸢尾花数据上的表现并进行分析

答：




```
testY = test_data['class']
testX = test_data.drop(['class'],axis=1)

predY = clf.predict(testX)
predictions = pd.Series(predY,name='Predicted Class')
```

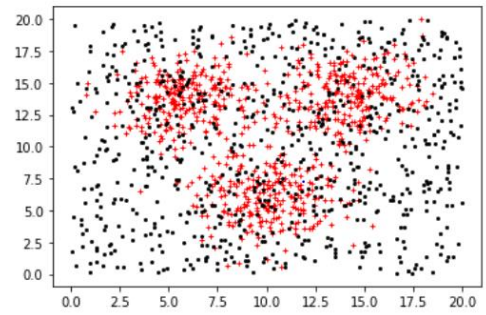
```
from sklearn.metrics import accuracy_score

print('Accuracy on test data is %.2f' % (accuracy_score(testY, predY)))
```

```
Accuracy on test data is 1.00
```

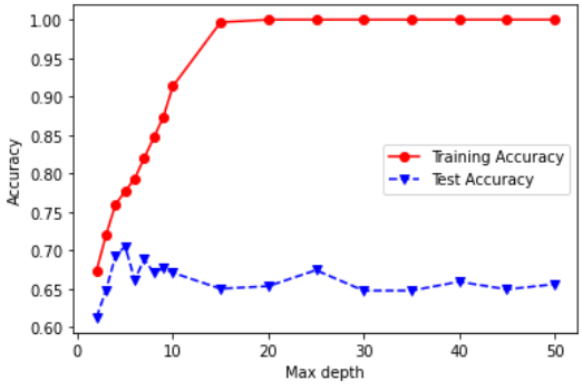
经过实验，发现 depth_max 取 5 时决策树构建最为全面，经过训练后，在测试集上正确率可达到 100%。

3. 观察不同规模的测试数据对决策树的过拟合情况是否有影响，如果有，请分析其规律

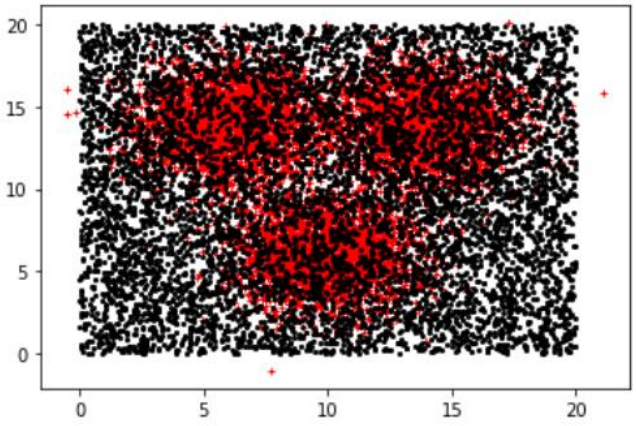
答：数据数量：1500：



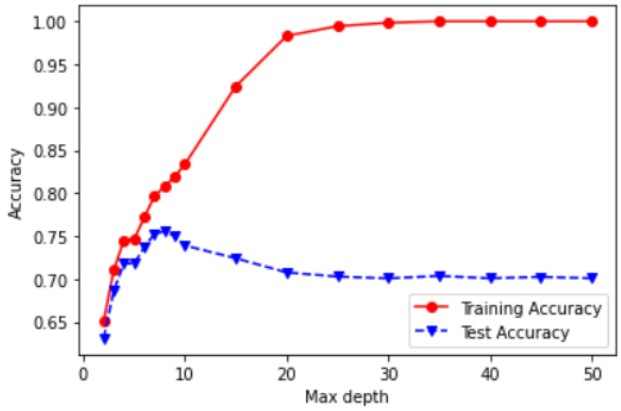

```
Text(0, 0.5, 'Accuracy')
```



数据数量：15000




```
Text(0, 0.5, 'Accuracy')
```



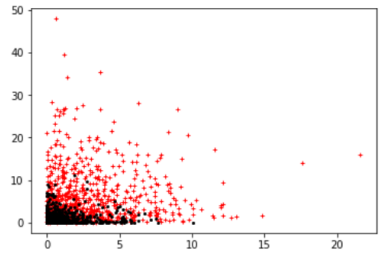
在数据规模 1500、6000、9000、150000 时都做了实验，由实验结果可得测试数据规模会影响决策树的过拟合情况。并且在一定数据规模范围内，数据规模越大，其泛化能力越好，即防止过拟合能力越强，而到了一定数据范围之后就基本不再变化。

4. 观察决策树在使用其它数据分布构造的数据集上的表现并进行分析。

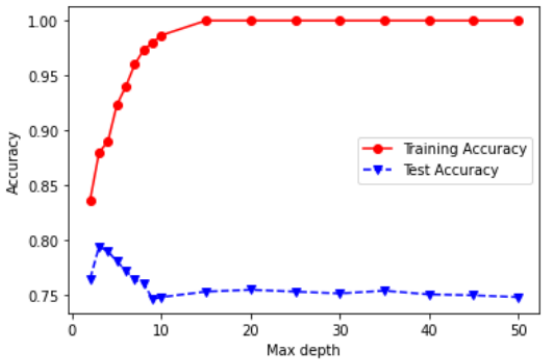
答：

```
X = np.random.exponential([3,6],size=[int(N/2),2])#指数分布
# X = np.random.chisquare(1,size=[int(N/2),2])#卡方发布
X = np.concatenate((X, np.random.chisquare(1,size=[int(N/2),2])))

Y = np.concatenate((np.ones(int(N/2)),np.zeros(int(N/2))))
plt.plot(X[:int(N/2),0],X[:int(N/2),1],'r+',X[int(N/2):,0],X[int(N/2):,1],'k.',ms=4)
```

```
[<matplotlib.lines.Line2D at 0x1e992e34828>,
 <matplotlib.lines.Line2D at 0x1e992e341d0>]
```



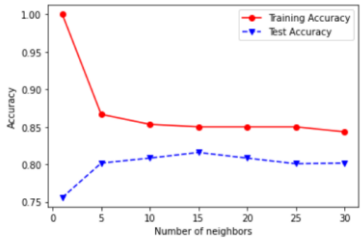
上图中尝试了指数分布和卡方分布的样本数据。



可以看到，当使用大众化的分布函数之后，得到的结果基本还是可以的，说明决策树的泛用性还是较广的。但是，经过一些其他的实验后，还是发现了一些问题，就比如当我拿数据分布偏斜的样本集（也就是当多数数据集中在曲线的一端，而少数数据在曲线的另一端时），就会发现决策树对少数类别样本的分类精度很低。通过查阅资料发现，决策树对数据分布甚至缺失是十分宽容的，不容易受到极值的影响，但是其最大的缺点就是容易过拟合。这与我们的实验结果也基本相符，如上我用一般分布函数尝试时基本没什么问题，但当数据分布偏斜时就会出现对少数样本分类效果不好的问题，很明显是过拟合了大类别数据。

5. 观察并分析 k 与训练和测试误差的关系

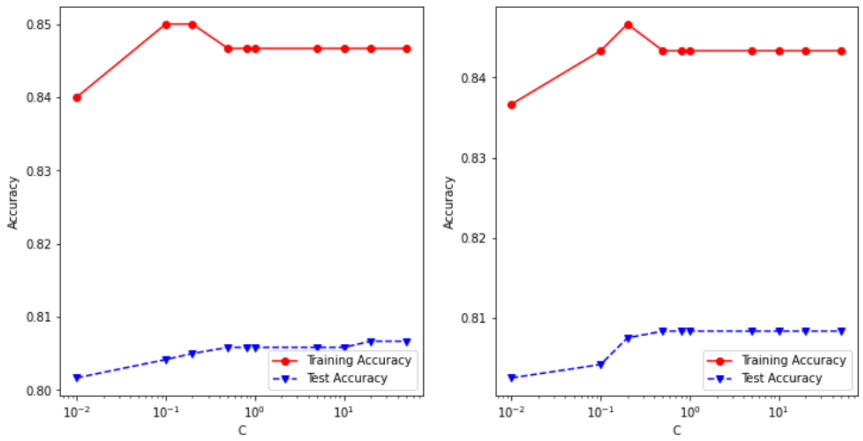
答：



在 k 值很小时，模型十分复杂，容易受到异常点的影响，非常容易产生过拟合的情况，从图中我们可以看到在 k 值取 1 时训练集上的正确率能够达到 100%，但在测试集上却只有 75%左右，这是很明显的过拟合现象。而在 k 值过大时，模型比较简单，受到样本不平衡的影响大，就容易产生欠拟合的情况，所以随着 k 值的增大，训练误差是一直下降，而测试误差是先上升后下降，从图中我们可以看出在该数据集中 k 值取 15 左右能够得到最好的测试误差。

6. 观察并尝试分析分类结果

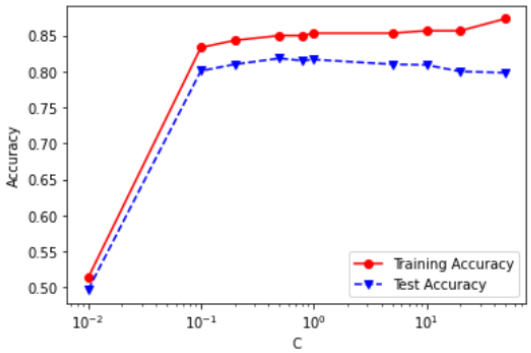
答：



因为超参数越小，即正则化越强，也就是防止过拟合能力越强。所以图中在 10^-2 处正则项的惩罚力度过大，导致模型未能较好地训练，而在 10 处则是因为正则项惩罚力度不足，但是泛化性得到保证，因此测试误差基本稳定在最高值左右，但是训练误差会略有下降。所以我们要使得惩罚力度刚刚好，有点类似机器学习里的学习率（也就是步长）问题，在图中可以发现 10^0 左右是比较好的。

7. 比较 SVM 和 Nonlinear SVM，并分析其特点

答：



非线性支持向量机对比线性支持向量机解决了非线性分布数据的分类问题，也就是解决了非凸问题。它利用核函数来找到数据在高维空间的近似线性分布，从而得到更好的泛化性。但是其核函数往往不能够通用，难以查找。

8. 通过查阅资料，学习三种集成方法，并对比分析其异同

答：

装袋：Bagging 的主要思想是“减少一个估计方差的一种方式就是对多个估计进行平均”。通俗一点讲就是把几个算法算出的结果，通过投票得出最终的结果。装袋法由于多次采样，每个样本被选中的概率相同，因此噪声数据的影响下降，所以装袋法不太容易受到过拟合的影响。因此我们可以看到装袋法对模型结果有一定提升。当然，提升程度与原模型的结果和数据质量有关。如果分类回归树的高数设置为 3 或 5，原算法本身的效果就会比较好，装袋法就没有提升的空间。所以装袋法还是有一定的数据局限。

增强：Boosting 指的是通过算法集合将弱学习器转换为强学习器。Boosting 的主要原则是训练一系列的弱学习器，所谓弱学习器是指仅比随机猜测好一点点的模型，例如较小的决策树，训练的方式是利用加权的数据。在训练的早期对于错分数据给予较大的权重。在实验中使用的是 Adaboosting，即自适应 boosting。

随机森林：随机森林是 Bagging 算法的进化版，它的思想仍然是 Bagging。 RF 使用了 CART 决策树作为弱学习器。在使用决策树的基础上，RF 对决策树的建立做了改进，对于普通的决策树，会在节点上所有的 n 个样本特征中选择一个最优的特征来做决策树的左右子树划分，但是 RF 通过随机选择节点上的一部分样本特征，这个数字小于 n，假设为 nsub，然后在这些随机选择的 nsub 个样本特征中，选择一个最优的特征来做决策树的左右子树划分。这样进一步增强了模型的泛化能力。

总的来讲，这三种集成学习的方法都是为了将多个学习器组合在一起，从而使得多个学习器互帮互助，把一个学习任务完成的更好。但是三种方法各有优缺点，在不

同数据的情况下要根据它们各自的特点来选择最为合适的方法。

三、意见和建议

（如有，请写出对本次实验的具体意见和建议，包括但不限于教学内容、实验内容、教学 PPT 等）