**信息科学技术学院**

**课后作业**

**课程设计名称：** 机器学习

**专 业：**

**学 号：**

**学 生 姓 名：**

**成 绩：**

**批 改 日 期：**

**教 师 签 名：**

**注意事项：**

1. 可相互讨论，但禁止相互抄袭，一旦发现，抄袭方和被抄袭方在本次实验报告都算不及格；
2. 可网络上寻找解决思路，但禁止直接拷贝，若拷贝，直接按抄袭判定不及格；
3. 本课程上机实验推荐使用Python语言，但不限于此语言；
4. 可使用现有的机器学习包，如；Python-SKLearn等；
5. 实验报告请提交PDF版本，文件名称：学号+姓名+实验报告（X）
6. 实验报告提交截止时间为下一次周二上课前（10点），由班长或学委统一收齐交给老师，请班长或学委提供未交报告学生名单，若超过截止时间，一律不算。

# 实验1 红酒分类-决策树

## 1. 实验内容

该实验是依据红酒的属性信息，包括酒精、苹果酸、灰、灰分的碱度、镁、总酚、黃酮类化合物、非黃烷类酚类、原花色素、颜色强度、色调、稀释葡萄酒的0280/OD315、脯氨酸等属性，对红酒进行分类。

1. 数据分析与处理
2. 决策树算法应用

## 2. 实验步骤

1. 数据分析与处理
2. 导入数据

导入datasets模块下的“红酒” 数据集load\_wine。

1. 数据可视化分析

导入matplotlib模块下的pyplot，画出关于各个特征的散点图。

1. 数据划分

导入model\_selection模块下的train\_test\_split，按7:3划分训练集和测试集。

1. 数据归一化

根据数据可视化分析结果，判断是否需要进行归一化操作。

1. 模型训练

导入tree模块下的DecisionTreeClassifier，建立决策树模型。

* 训练fit(X, y)

1. 模型评估

利用估计的模型参数对测试集进行预测，导入metrics模块下的性能度量准则，评估模型预测性能。

* 预测predict(X)
* 准确率accuracy\_score
* 召回率 recall\_score

## 3. 实验要求

分析不同参数对分类结果的影响，包括：

1. 不同的决策树不纯度计算方法criterison，信息熵entropy，基尼指数gini。
2. 控制决策树中的随机选项splitter,取值best、random。
3. 限制决策树的最大深度，超过设定深度的树枝全部剪掉max\_depth。
4. 节点在分之后每个子节点至少包含的训练样本数min\_samples\_leaf
5. 节点允许被分支至少需要包含的训练样本数min\_samples\_split

## 4. 实验代码

## 5. 实验结果

# 实验2 肿瘤分类与预测-神经网络

## 实验内容

采用神经网络方法，对美国威斯康星州的乳腺癌诊断数据集进行分类，实现针对乳腺癌检测的分类器，以判断一个患者的肿瘤是良性还是恶性。本项目是对美国威斯康星州的乳腺癌诊断数据集进行分类，该数据集包括569个病例的数据样本，每个样本具有30个特征值，样本共分为两类：分别是恶性和良性，利用BP神经网络算法来训练模型，并对该模型进行评估和预测。

## 2. 实验步骤

1. 数据处理
2. 导入数据

导入datasets模块下的数据集load\_breast\_cancer。

1. 数据划分

导入model\_selection模块下的train\_test\_split，按8:2比例划分训练集和测试集。

1. 模型训练

导入neural\_network模块下的MLPClassifier，建立BP模型。

* 训练fit(X, y)

1. 模型评估

利用估计的模型参数对测试集进行预测，导入metrics模块下的性能度量准则，评估模型预测性能。

* 预测predict(X)
* 准确率accuracy\_score
* 召回率 recall\_score
* F1值 f1\_score

## 3. 实验要求

进行模型预测，对真实数据和预测数据进行可视化（用 Axes3D 绘制 3d 散点图）。

分析不同参数对分类结果的影响，包括：

1. 网络结构 hidden\_layer\_sizes
2. 优化器solver
3. 激活函数 activation
4. 最大迭代次数 max\_iter

## 4. 实验代码

## 5. 实验结果

# 实验2 红酒分类-支持向量机

## 1. 实验内容

该实验是依据红酒的属性信息，包括酒精、苹果酸、灰、灰分的碱度、镁、总酚、黃酮类化合物、非黃烷类酚类、原花色素、颜色强度、色调、稀释葡萄酒的0280/OD315、脯氨酸等属性，对红酒进行分类。

1. 数据分析与处理
2. 支持向量机算法应用

## 2. 实验步骤

1. 数据分析与处理
2. 导入数据

导入datasets模块下的“红酒” 数据集load\_wine。

1. 数据统计

统计样本在特征值下的平均值、方差、最值指标，以及各类别下的样本个数。

1. 特征相关性分析

绘制特征相关性分析热力图，通过相关性分析的结果，判断是否需要删除部分特征。

1. 数据集划分

导入model\_selection模块下的train\_test\_split，按8:2划分训练集和测试集。

1. 模型训练

导入svm模块下的SVC，建立支持向量机模型。

* 训练fit(X, y)

1. 模型评估

利用估计的模型参数对测试集进行预测，导入metrics模块下的性能度量准则，评估模型预测性能。

* 预测predict(X)
* 准确率accuracy\_score
* 召回率 recall\_score

## 3. 实验要求

分析不同参数对分类结果的影响，包括：

1. 惩罚参数C
2. 核函数类型kernel

## 4. 实验代码

## 5. 实验结果

# 实验1 肿瘤分类与预测-朴素贝叶斯

## 实验内容

采用朴素贝叶斯方法，对美国威斯康星州的乳腺癌诊断数据集进行分类，实现针对乳腺癌检测的分类器，以判断一个患者的肿瘤是良性还是恶性。本项目是对美国威斯康星州的乳腺癌诊断数据集进行分类，该数据集包括569个病例的数据样本，每个样本具有30个特征值，样本共分为两类：分别是恶性和良性，利用高斯朴素贝叶斯算法来训练模型，并对该模型进行评估和预测。

## 2. 实验步骤

1. 数据分析与处理
2. 导入数据

导入datasets模块下的数据集load\_breast\_cancer。

1. 数据可视化分析

导入matplotlib模块下的pyplot，画出关于各个特征的散点图。

1. 数据划分

导入model\_selection模块下的train\_test\_split，按一定比例划分训练集和测试集。

(4)数据归一化

根据数据可视化分析结果，判断是否需要进行归一化操作。

1. 模型训练

导入naive\_bayes模块下的GaussianNB，建立高斯朴素贝叶斯模型。

* 训练fit(X, y)

3）模型评估

利用估计的模型参数对测试集进行预测，导入metrics模块下的性能度量准则，评估模型预测性能。

* 预测predict(X)
* 准确率accuracy\_score
* 召回率 recall\_score
* F1值 f1\_score

## 3. 实验要求

1. 尝试不同的数据集划分比例（如9:1,8:2,7:3,…,1:9），分析对预测结果的影响，画出性能关于不同比例的折线图。
2. 计算分类结果的混淆矩阵，并用热力图显示。

## 4. 实验代码

## 5. 实验结果