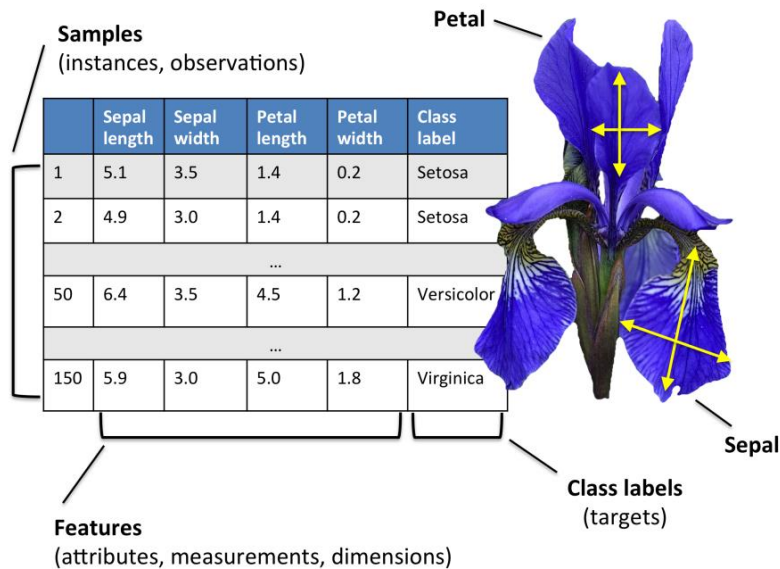


# 实验作业要求

## 一、数据集

Iris 鸢尾花数据集： 包含 3 类分别为山鸢尾 (Iris-setosa)、变色鸢尾 (Iris-versicolor) 和维吉尼亚鸢尾 (Iris-virginica), 共 150 条数据, 每类各 50 个数据, 每条记录都有 4 项特征：花萼长度、花萼宽度、花瓣长度、花瓣宽度，通常可以通过这 4 个特征预测鸢尾花卉属于哪一品种。Iris 数据集内数据示意格式如下：



The diagram illustrates the structure of the Iris dataset. On the left, a table represents the data samples. The table has six columns: an index column, 'Sepal length', 'Sepal width', 'Petal length', 'Petal width', and 'Class label'. The first two rows are for the 'Setosa' class, followed by an ellipsis, then a row for the 'Versicolor' class, another ellipsis, and finally a row for the 'Virginica' class. To the right of the table is an illustration of an Iris flower. Yellow arrows point from the table columns to the corresponding parts of the flower: 'Petal' points to the petals, 'Sepal' points to the sepals, and 'Class labels (targets)' points to the overall flower. Labels for 'Samples (instances, observations)', 'Features (attributes, measurements, dimensions)', and 'Class labels (targets)' are also present with arrows pointing to their respective parts in the diagram.

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

图中第一行数据的意义是：花萼长度、花萼宽度、花瓣长度、花瓣宽度、鸢尾花类别；其中 setosa、versicolor、virginica 分别为三种鸢尾花名。

从第二行开始各列数据的意义：第一列为该条数据的序号；第二列为花萼长度值；第三列为花萼宽度值；第四列为花瓣长度值；第五列为花瓣宽度值；第六列对应是种类（三类鸢尾花分别用 0, 1, 2 表示）。

打印输出示例：

```
5.1, 3.5, 1.4, 0.2, 0
4.9, 3.0, 1.4, 0.2, 0
.....
6.4, 3.5, 4.5, 1.2, 1
.....
5.9, 3.0, 5.0, 1.8, 2
```

## 二、数据集导入

首先要在自己的 Python 环境中下载 sklearn：

```
pip install scikit-learn -i https://pypi.tuna.tsinghua.edu.cn/simple
```

下载数据集：

```
from sklearn.cluster import KMeans
from sklearn import datasets
from sklearn.datasets import load_iris
iris = load_iris()
```

### 三、实验内容（二选一）

#### （一）决策树

以小组为单位，3 人一组，对鸢尾花数据集构造决策树，通过手动划分训练集和测试集来对决策树进行评估，并按照实验报告模版撰写实验报告。

参考代码：

```
# 加载数据并训练决策树
from sklearn.datasets import load_iris
from sklearn import tree

iris = load_iris()
clf = tree.DecisionTreeClassifier()
clf = clf.fit(iris.data, iris.target)

# 可视化决策树
import graphviz

dot_data = tree.export_graphviz(
    clf,
    out_file=None,
    feature_names=iris.feature_names,
    class_names=iris.target_names,
    filled=True,
    rounded=True,
    special_characters=True
)

graph = graphviz.Source(dot_data)
graph.render("iris")
```

常用评估指标：

#### 1. 准确率 (Accuracy)

准确率表示模型预测正确的样本数在总样本中的占比，是最常用的整体性能衡量指标。

$$\text{Accuracy} = \frac{\text{预测正确的样本数}}{\text{总样本数}}$$

#### 2. 精确率 (Precision)

精确率衡量模型预测为某一类别的样本中，有多少是真正属于该类别的。

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}$$

其中：  $TP_i$ : 将类别  $i$  正确预测为  $i$  的数量

$FP_i$ : 将其他类别误判为  $i$  的数量

#### 3. 召回率 (Recall)

召回率衡量模型从某一真实类别中识别出的比例，即模型在该类别上的“漏检率”。

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}$$

其中：  $TP_i$ : 将类别  $i$  正确预测为  $i$  的数量

$FN_i$ : 将类别  $i$  错误预测为其他类别的数量

## （二）K-Means 聚类

以小组为单位，3 人一组，对鸢尾花数据集进行 K-means 聚类，练习如何确定 K 值，并按照实验报告模版撰写实验报告。

常用 K 值评价标准：

### 1. 肘部法则（Elbow Method）

肘部法则通过计算不同 K 值下的 SSE（Sum of Squared Errors）来确定最佳 K 值。SSE 随着 K 的增加而减小，但当 K 增加到一定程度时，SSE 的下降速度会显著减缓，这个点就是“肘部”。

### 2. 轮廓系数（Silhouette Score）

轮廓系数衡量了聚类的紧密性和分离性，取值范围在[-1, 1]之间。值越接近 1，表示聚类效果越好。

根据肘部法则和轮廓系数的结果，选择最佳 K 值，并应用 K-means 聚类。

## 四、实验报告要求（附实验报告模版）

### （一）决策树

1. 展示训练集和测试集的划分方式，并说明每类样本数量的分布情况
2. 展示训练后的决策树结构图和分类模型在测试集上的评估结果
3. 简要分析决策树在鸢尾花数据集上的分类表现

### （二）K-Means 聚类

1. 展示不同 K 值下肘部法则和轮廓系数的图表，并给出你所选择的最佳 K 值以及理由
2. 展示最终的聚类结果
3. 简要分析 K-means 算法在鸢尾花数据集上的表现