山西农业大学
Shanxi Agricultural University

2025秋

计算思维与人工智能基础

实验5　机器学习基础：鸢尾花分类与葡萄酒聚类

本讲提纲

01 机器学习简介

02 Scikit-learn工具箱简介

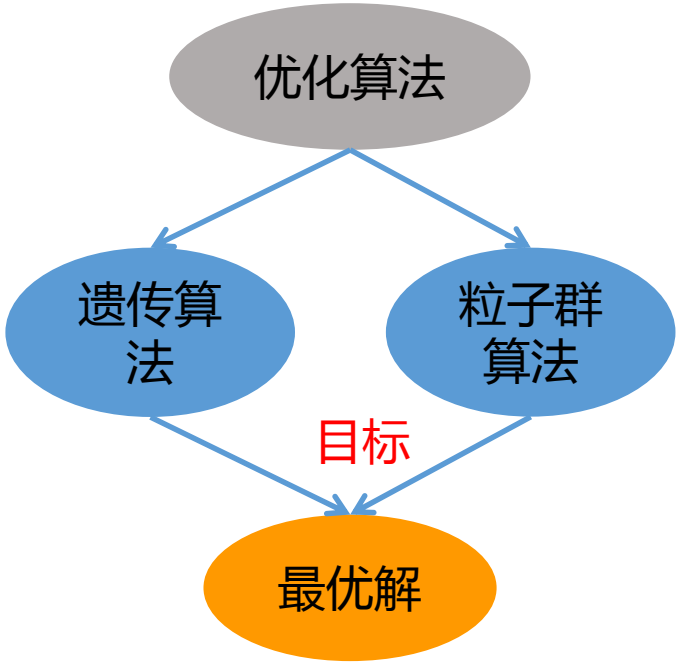
03 监督学习及实践——鸢尾花分类

04 无监督学习及实践——葡萄酒聚类

05 任务发布与解析

上节回顾

优化算法：像“调配方”



- 给定明确目标（如作物最高产量）
- 不断调整氮磷钾比例 → 搜索“最优组合”
- 代表：遗传算法、粒子群算法等

☞ 目标明确，方法 = 搜索

*思考：*是否我们不再直接告诉机器如何找到答案，而是给定海量的数据，让其从数据中“学习”出解决问题的模式和规律呢？

机器学习简介

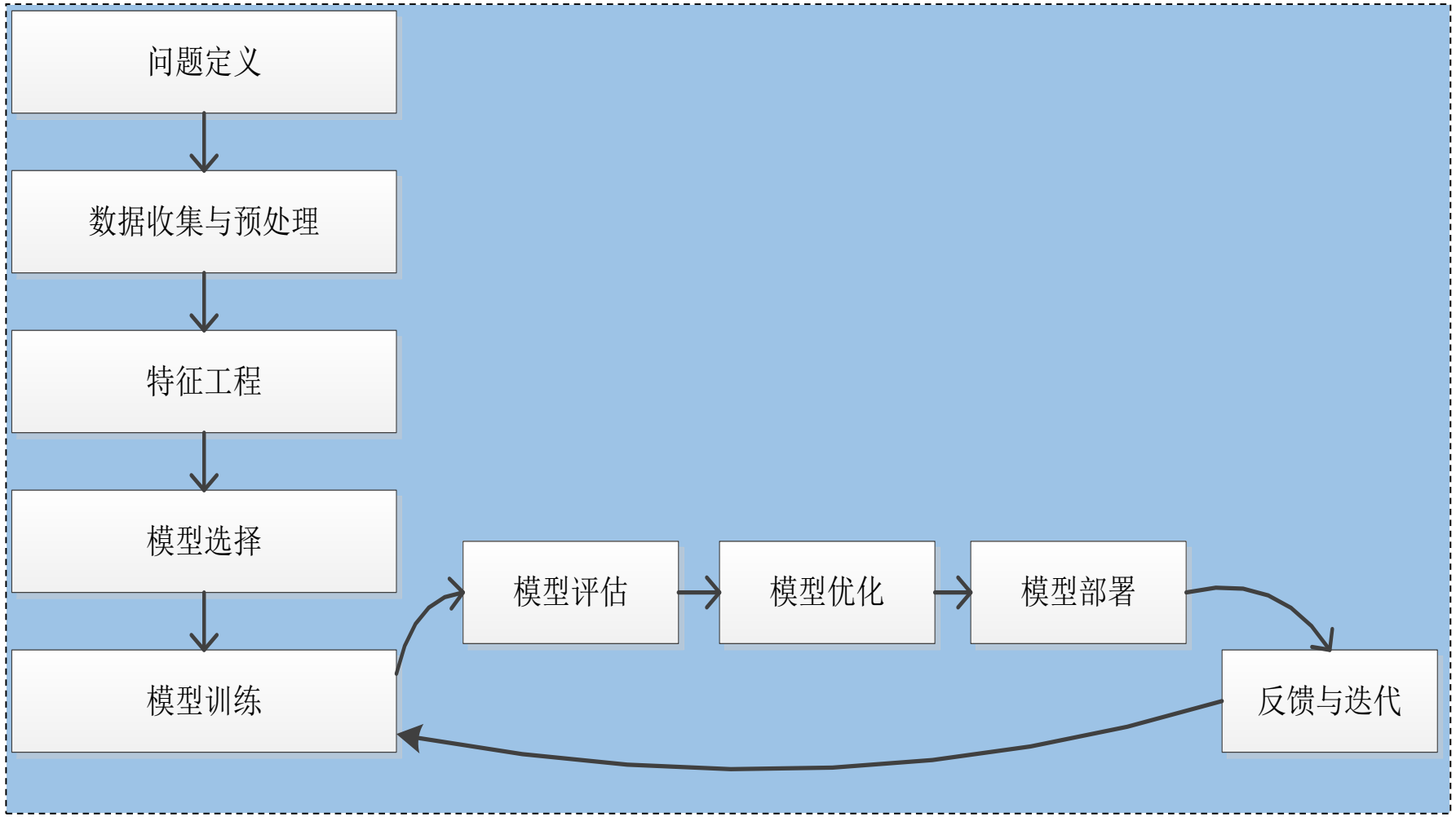
定义：机器学习是人工智能分支，通过算法让计算机从数据中学习规律，无需显式编程，能自主优化模型，应用于预测、分类等任务，是实现智能化的核心技术。

- **机器学习：像“老专家”**

  - 看过成千上万块地
  - 从土壤、气候、作物组合中总结规律
  - 看到新地 → 预测种什么最合适、收成如何

  ☞ 目标隐含，方法 = 学习

机器学习开发流程



机器学习开发流程

预处理-整理数据

在机器学习开发流程中，首先需要对数据进行预处理和整理。这包括数据清洗、缺失值处理、异常值检测等步骤，以确保数据的质量和一致性。

特征工程

特征工程是机器学习中的关键环节。通过选择、提取和转换特征，可以提高模型的预测精度和泛化能力。特征工程包括特征选择、特征缩放、特征组合等操作。

机器学习算法训练

在准备好数据后，需要选择合适的机器学习算法进行训练。训练过程中，算法会不断调整模型参数以最小化损失函数，从而提高模型的预测性能。

模型评估

模型训练完成后，需要对其进行评估以验证其性能。评估指标包括准确率、召回率、F1分数等。通过评估结果，可以判断模型的优劣并进行相应的调整。

反馈与迭代

机器学习开发流程是一个不断迭代和改进的过程。根据评估结果和实际需求，可以对模型进行调优和改进，以提高其性能和适用性。

Scikit-learn工具箱简介

Scikit-learn 是 Python 主流机器学习开源库，主打简单高效的监督 / 无监督学习工具，且兼容 NumPy、Pandas 等库。

其核心价值是降低机器学习门槛，核心优势：

API 统一简洁，所有模型均遵循 “实例化→拟合(fit)→预测(predict)” 流程，学习成本低。

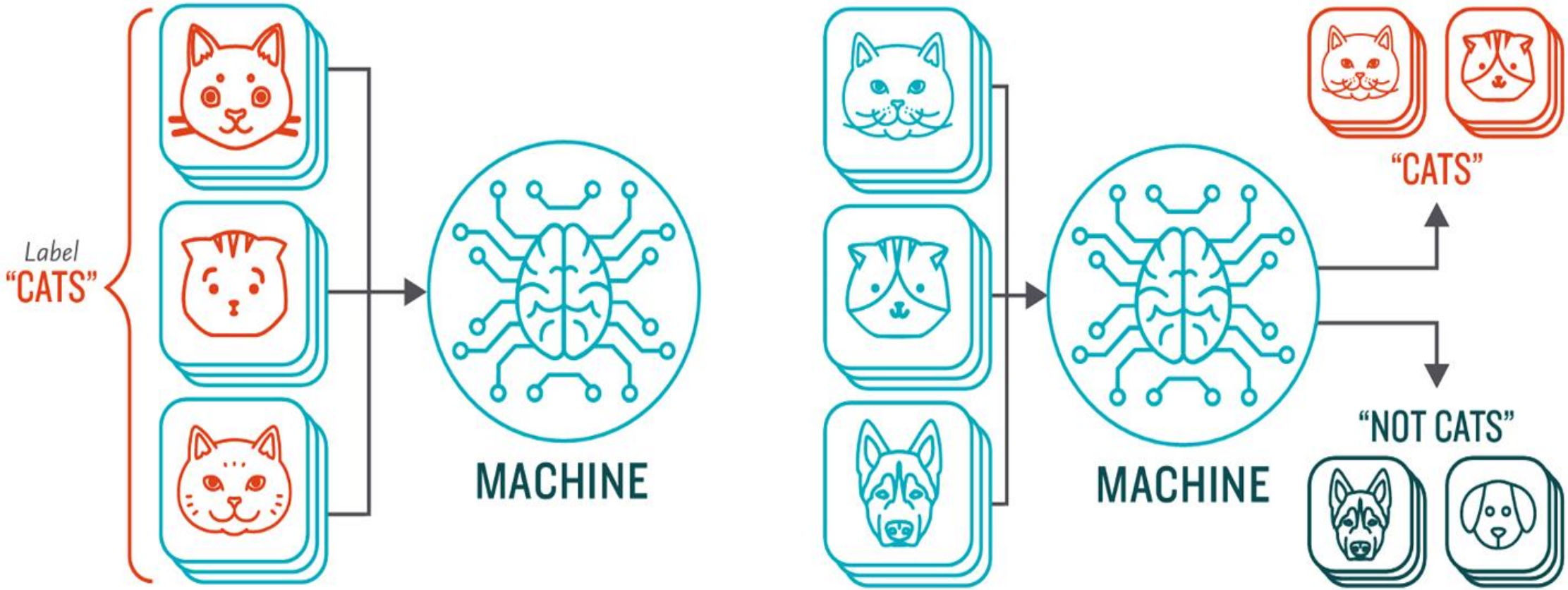
工具链开箱即用，覆盖数据预处理、模型训练、评估调参等全流程，无需拼接多库。

轻量高效，依赖 NumPy 和 SciPy 计算，适合中小型数据集快速实验。

Scikit-learn工具箱简介

模块类别	关键功能	常用工具
数据预处理	数据清洗、标准化、编码	StandardScaler（标准化）、OneHotEncoder（独热编码）、Imputer（缺失值填充）
特征工程	特征选择、降维	SelectKBest（特征选择）、PCA（主成分分析）、TSNE（降维可视化）
模型训练	监督学习、无监督学习	分类：LogisticRegression、RandomForestClassifier； 回归：LinearRegression；聚类：KMeans
模型评估与调参	性能指标、参数优化	accuracy_score（准确率）、cross_val_score（交叉验证）、GridSearchCV（网格搜索调参）

监督学习



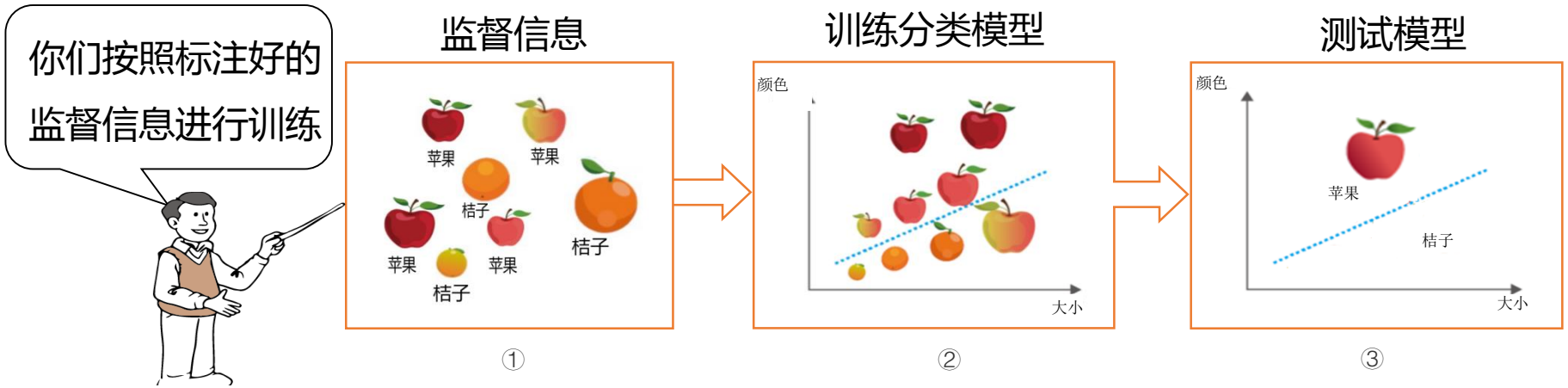
监督学习

下面是监督学习的一个例子，学习的目的是让机器分辨苹果和桔子。

① 收集一些苹果和桔子的图片，并对这些图片进行标注，标明哪些图片是苹果，哪些图片是桔子。这些标注即是监督信息。

② 用这些图片训练一个对苹果和桔子的分类模型。

③ 将一幅没见过的苹果图片送入分类模型，模型分辨出这是一个苹果而不是一个桔子。



监督学习

在机器学习和统计建模中，**拟合**、**过拟合**、**欠拟合**和**泛化**是描述模型性能和适应性的核心概念。

概念	核心定义	关键表现	主要原因
拟合 Fitting	模型调整参数**逼近**训练数据**潜在规律**，兼顾训练拟合与泛化能力	训练误差合理，泛化能力需结合测试数据判断	模型复杂度与数据规律匹配，训练数据质量达标
过拟合 Overfitting	模型**过度学习**训练数据细节（含噪声），忽视普遍规律	训练误差极小，测试误差显著增大	模型复杂度过高，训练数据量不足或噪声过多
欠拟合 Underfitting	模型**未能充分捕捉**训练数据中的真实规律，学习能力不足	训练误差和测试误差均较大，且差距小	模型复杂度太低，特征工程不足（未提取关键特征）
泛化 Generalization	模型将训练数据中学到的规律，应用于未见过的新数据的表现能力	训练数据与测试数据误差稳定且较低	模型捕捉数据普遍规律，未受噪声或复杂度影响

监督学习实践

- 监督学习是指我们给算法一个**数据集**，并且**给定正确答案**。机器通过数据来学习正确答案的计算方法。
- 监督学习类似于老师教学生，把知识直接传授给学生，学生记住了老师讲解的知识，即可用于实践。

像我们小时候学习认字、识图。老师会指着一张花的图片，告诉我们它的各种测量数据，然后明确地告诉我们：“这朵花，叫做‘山鸢尾’”。我们看了很多带有“正确答案”的图片后，再看到一朵新的花，就能根据它的测量数据，自己判断出它的品种。


Iris setosa


Iris versicolor


Iris virginica

监督学习实践

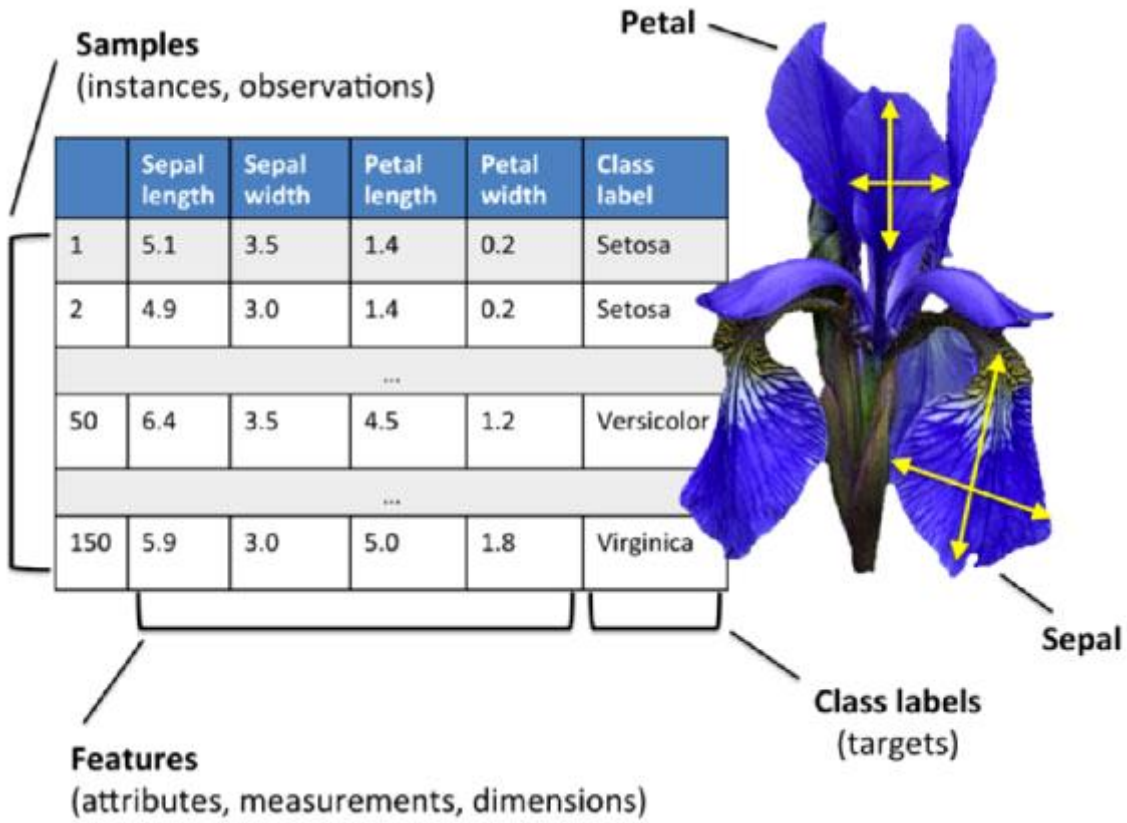
监督学习的训练数据集包括标签数据。图中的每一行都与一个目标或标签相关联。列是不同的特征，行代表不同的数据点，通常称为样本。

鸢尾花分类

数据集：包含山鸢花、变色鸢花和弗吉尼亚鸢尾3种不同鸢尾属的150朵鸢尾花的测量结果。

数据集每行存储一朵花的样本数据，每列存储每种花的度量数据(以cm为单位)，也称为数据集的特征。

带有标签的数据告诉了机器学习算法输人数据与输出结果之间的对应关系。



	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

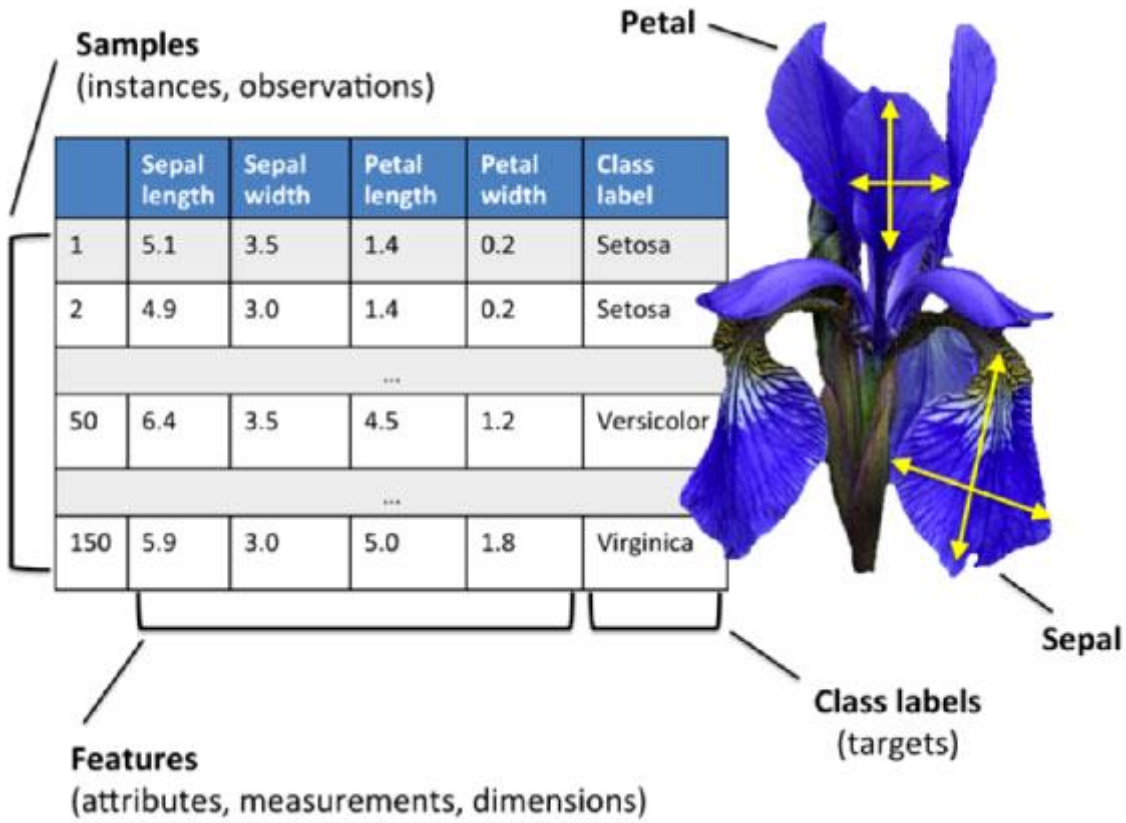
监督学习实践

鸢尾花分类

数据集包括4个属性，分别为**花萼的长**、**花萼的宽**、**花瓣的长**和**花瓣的宽**。

花萼是花冠外面的绿色被叶，在花尚未开放时，保护着花蕾。

鸢尾花分类代码演示讲解
JupyterNotebook



	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

无监督学习

是否可以没有标签的情况下采用机器学习进行训练学习呢？

无监督学习是机器学习的核心范式之一，指利用**无标签（即无已知输出结果）**的训练数据，让模型**自主挖掘数据**内在的结构、规律或隐藏特征，无需人工预先定义类别或目标。

简单来说，模型在“无监督”的情况下自主学习：训练数据只有“特征（输入）”，没有“标签（期望输出）”，模型通过分析数据本身的分布、相似度等信息，完成聚类、降维或异常检测等任务。

无监督学习案例

作物品种聚类

•**应用场景**：在育种研究中，科研人员需对大量未知品系的作物（如小麦、水稻）进行分类。

•**数据特征**：收集不同作物样本的形态特征（株高、穗长、粒重）、生理特征（光合作用速率、抗病性指标）等数据，且不预先标注品种类别。

•**模型作用**：通过聚类算法，自动将特征相似的作物样本归为一类，帮助科研人员快速划分出不同的品种群组，为后续育种筛选提供依据。

土壤肥力等级划分

•**应用场景**：农业生产中，需快速评估一片农田不同区域的土壤肥力，以实现精准施肥。

•**数据特征**：采集土壤样本的理化指标（氮磷钾含量、pH 值、有机质含量），不预先设定肥力等级（如高肥、中肥、低肥）。

•**模型作用**：利用聚类等算法，根据土壤指标的相似度对地块进行分组，自动划分出不同肥力等级的区域，指导农户针对不同区域调整施肥量。

作物病虫害异常检测

•**应用场景**：在大规模种植基地，需及时发现生长异常的作物（可能受病虫害影响），减少损失。

•**数据特征**：通过传感器或图像采集作物的生长数据（叶片颜色、株高增长速度、叶片湿度），大部分样本为正常生长的作物数据。

•**模型作用**：使用异常检测算法，自主学习正常作物的生长数据分布，将明显偏离该分布的样本判定为异常，帮助农户快速定位病虫害发生区域。

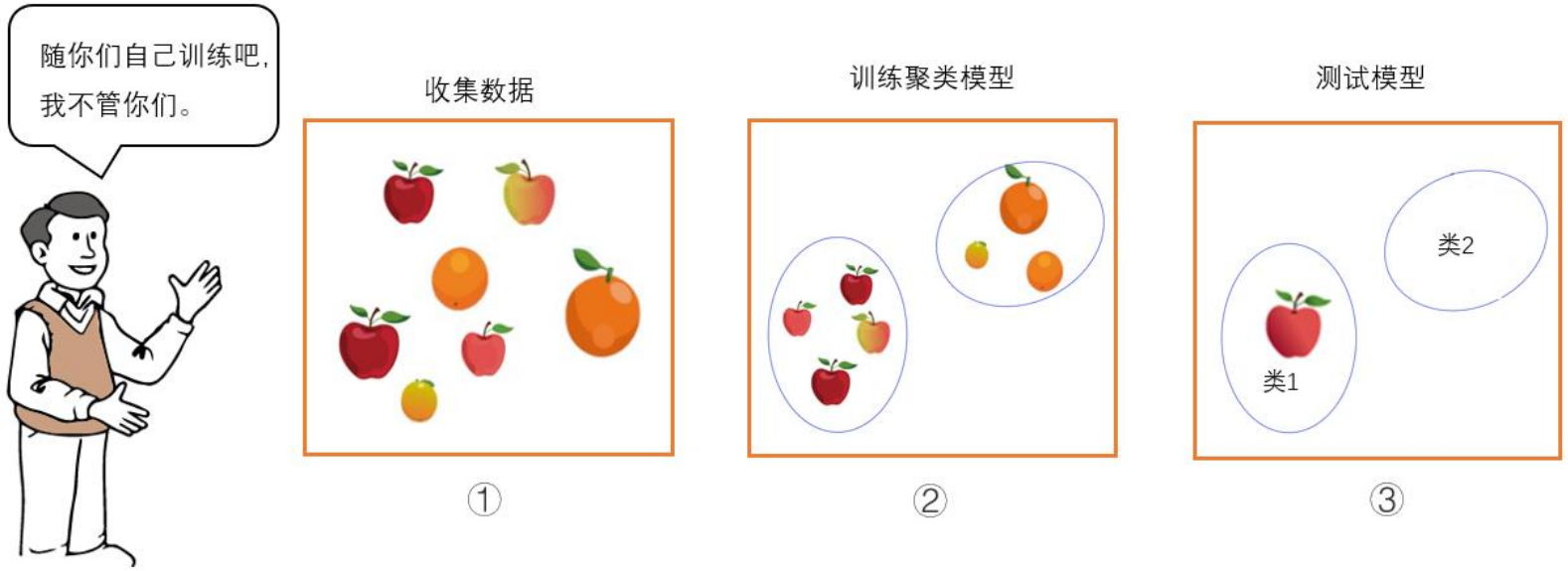
无监督学习案例

将水果进行归类

收集一些水果图片，但并不知道图片中的水果的名字，因此不包含监督信息。

用这些图片训练一个模型，这一模型可以将相似的水果聚成一堆。这一过程称为“聚类”。

将一幅没见过的苹果图片送入模型，模型将归入苹果一类。



无监督学习

聚类 (Clustering): 在没有标签的情况下，将数据根据其相似性自动分组成不同“簇”或“群组”的过程。

簇 (Cluster): 一个数据分组，簇内的数据彼此相似，簇间的数据彼此不相似。

KMeans算法: 一种非常经典的聚类算法。目标是找到K个簇的中心点，使得每个数据点到它所属簇的中心点的距离之和最小。

葡萄酒聚类代码演示讲解
JupyterNotebook

实验内容一：鸢尾花分类 =>监督学习

•通过补全代码，熟悉“加载数据->划分数据->创建模型->训练->预测->评估”这一套监督学习的流程。

•学会做“参数调优”。模型不是一成不变的，比如KNN里的邻居数量K，K值不同结果可能完全不同。需要通过实验，找到那个“最优”的K值。

•完成思考题。

实验内容二：葡萄酒聚类=>无监督学习

- 通过补全代码，熟悉无监督学习的流程。
- 处理多维特征，并理解“数据标准化”为什么在这种情况下至关重要。
- 完整复现“肘部法则”来科学地决定K值。
- 学会一种新的、不依赖可视化的结果分析方法——分组计算均值，来解读每个聚类簇的内在含义。
- 完成思考题。

请同学认真完成实验任务

并保存代码和结果截图