1. 数据集长什么样

Iris 数据集介绍

来源：由英国统计学家 R.A. Fisher 在 1936 年发表的论文中引入，用来展示判别分析方法（正好就是你现在学的 Fisher 判别！）。

规模：

共 150 个样本；

3 个类别，每类 50 个样本。

类别（目标变量 y）：

Setosa（山鸢尾）

Versicolor（杂色鸢尾）

Virginica（维吉尼亚鸢尾）

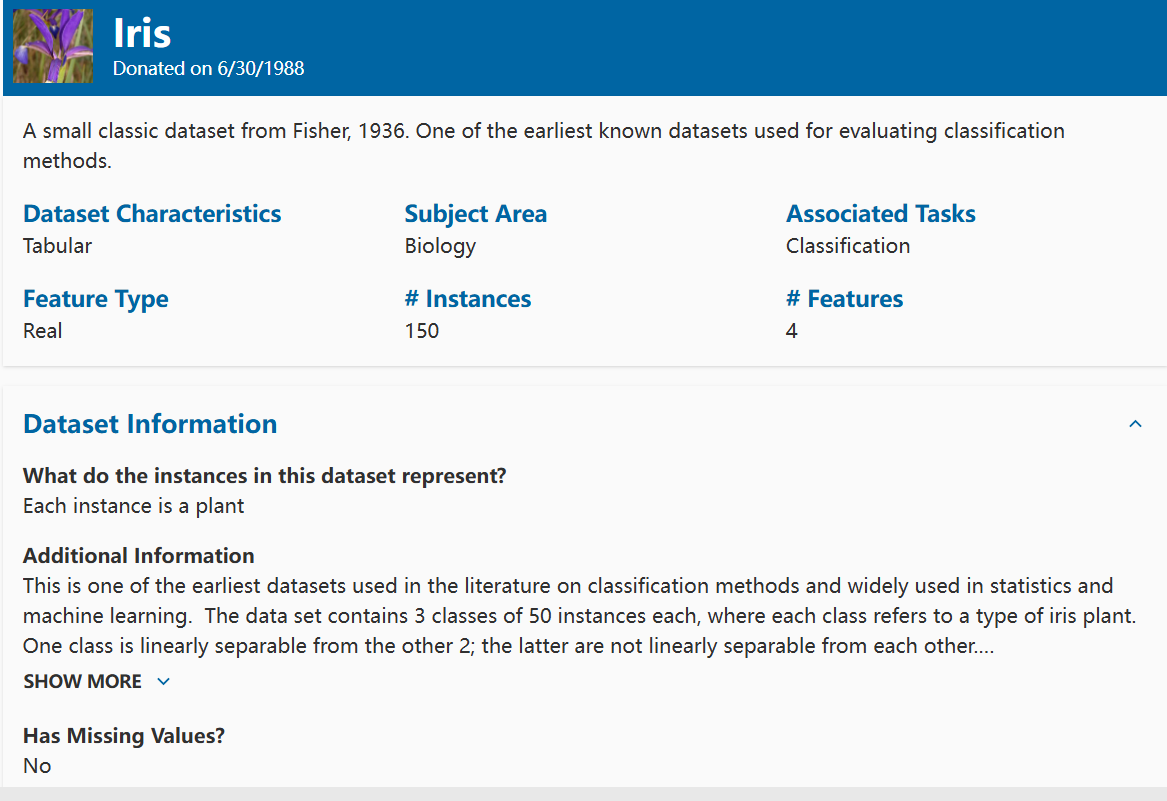
特征（输入 X，共 4 维）：

萼片长度（Sepal length, cm）

萼片宽度（Sepal width, cm）

花瓣长度（Petal length, cm）

花瓣宽度（Petal width, cm）



Sonar 数据集简介

**全名：**Connectionist Bench (Sonar, Mines vs. Rocks) Data Set

**来源：**UCI 机器学习库（id=151），最早由 Gorman 和 Sejnowski (1988) 提供，用于声纳信号分类。

**任务：**二分类问题

**M（Mine）：**声纳探测到的是水下“矿石/金属物体”

**R（Rock）：**声纳探测到的是普通“岩石”

📊 数据规模

**样本数：**208

**特征数：**60（全是数值型特征）

**类别分布：**大约一半是 `R`，一半是 `M`（具体 111 vs 97，略不平衡）

特征说明

每个样本是一个 60 维的向量，表示在不同角度或时间窗口下的声纳返回信号强度（能量值）。

所有特征取值在 `[0, 1]` 之间（已经做过归一化处理）。

没有缺失值。

换句话说：

一行数据 = 一个声纳信号样本

前 60 列 = 声纳信号的 60 个特征值

第 61 列 = 类别标签（M / R）

为什么常用 Sonar？

**1. 高维小样本（208 个样本，但特征有 60 个）：**

非常适合测试降维算法（比如 LDA / PCA）。

也能验证算法在小数据下是否容易过拟合。

**2. 二分类问题：**比 Iris 的三分类简单，但维度更高，挑战在于“如何提取有效的低维表示”。

**3. 历史经典：**很多经典论文和教材都会用 Sonar 验证模型性能。

数据样例（前几行，简化后）

| f1 | f2 | ... | f60 | label |

| ------ | ------ | --- | ------ | ----- |

| 0.0200 | 0.0371 | ... | 0.0079 | R |

| 0.0453 | 0.0523 | ... | 0.0044 | R |

| 0.0262 | 0.0582 | ... | 0.0090 | R |

| 0.0100 | 0.0171 | ... | 0.0031 | M |

注意：

每行有 60 个小数（0~1）。

标签只有 `R` 或 `M`。

与 Iris 的对比

Iris：低维（4D）、多类（3 类）、样本多（150） → 易可视化。

Sonar：高维（60D）、二类（2 类）、样本少（208） → 难以直接画图，需要降维分析。

\* Iris 上 LDA 能很直观地分开三类。

\* Sonar 上需要选择合适的投影维度，看准确率变化。

在机器学习里我们更看重 **样本数量相对维度是否足够**。 所以Iris样本多

