

项目 6：数据驱动智能

实训目标

能够使用表格软件完成数据集搭建与可视化图表制作，借助散点图理解 K 近邻分类基本思想，完成样本预测与脏数据干扰实验，理解数据质量对人工智能模型效果带来的直接影响，建立数据清洗的基本观念。

实训环境

Windows 10 及以上系统，计算机安装 Excel 或 WPS 表格软件，设备具备基础文件存储能力。

操作步骤

任务一：构建数据集

- 1. 打开软件。双击桌面 Microsoft Excel 或 WPS 表格图标，启动软件。
- 2. 单击左上角「文件」后，单击「新建」→ 选择「空白工作簿」，创建新表格。单击左上角「保存」，文件命名为：电影题材智能分类实验，保存至桌面。
- 3. 录入表头在对应单元格，输入以下内容：

A1 单元格：电影名称

B1 单元格：打斗镜头（次）（特征 X）

C1 单元格：搞笑镜头（次）（特征 Y）

D1 单元格：电影类型（标签）

录入训练集数据按行依次输入以下 6 组样本，注意检查输入是否错误。

自定义的数据集

电影名称	打斗镜头(次)	搞笑镜头(次)	电影类型
战狼 2	98	2	动作片
碟中谍 6	95	5	动作片

叶问 4	102	3	动作片
泰囧	5	95	喜剧片
西虹市首富	3	92	喜剧片
憨豆特工	10	98	喜剧片

此时，该表是一个标准的训练集。机器通过这 6 条数据，初步掌握了动作片和喜剧片的区别，录入完成后，按 **Ctrl+S** 保存表格。

	A	B	C	D	E	F	G	H	I	J
1	电影名称	打斗镜头(次)	搞笑镜头(次)	电影类型						
2	战狼2	98	2	动作片						
3	碟中谍6	95	5	动作片						
4	叶问4	102	3	动作片						
5	泰囧	5	95	喜剧片						
6	西虹市首富	3	92	喜剧片						
7	憨豆特工	10	98	喜剧片						
8										
9										
10										
11										

录入数据

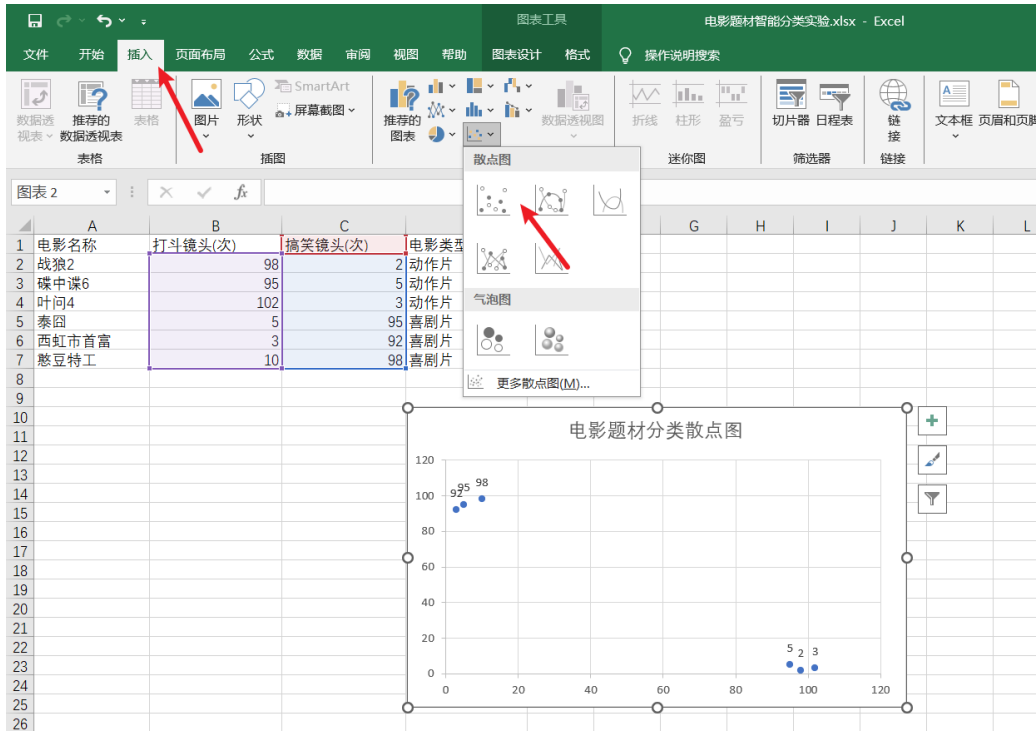
任务二：插入散点图

1. 选中数据区域鼠标拖动选中 B1:C7 (包含表头，只选特征列，不选电影名称、类型)。

	A	B	C	D
1	电影名称	打斗镜头(次)	搞笑镜头(次)	电影类型
2	战狼2	98	2	动作片
3	碟中谍6	95	5	动作片
4	叶问4	102	3	动作片
5	泰囧	5	95	喜剧片
6	西虹市首富	3	92	喜剧片
7	憨豆特工	10	98	喜剧片
8				
9				
10				

选中数据范围

2. 单击顶部菜单栏「插入」→ 找到「图表」→ 选择散点图（默认散点图即可，无需修改样式）。软件自动生成散点图，拖动图表至表格右侧空白区域，方便后续查看，图表标题修改为「电影题材分类散点图」。



生成散点图

3. 观察与分析

动作片：集中在图表右下区域（X 大、Y 小）

喜剧片：集中在图表左上区域（X 小、Y 大）

两类数据形成明显聚类，该散点图直观展示了数据的空间分布特征，这正是 KNN 等算法进行数学分类的基础。

任务三：未知电影的分类

1. 录入未知电影类型数据

在表格第 8 行输入以下数据，不改动原有数据。

A8: 上海堡垒

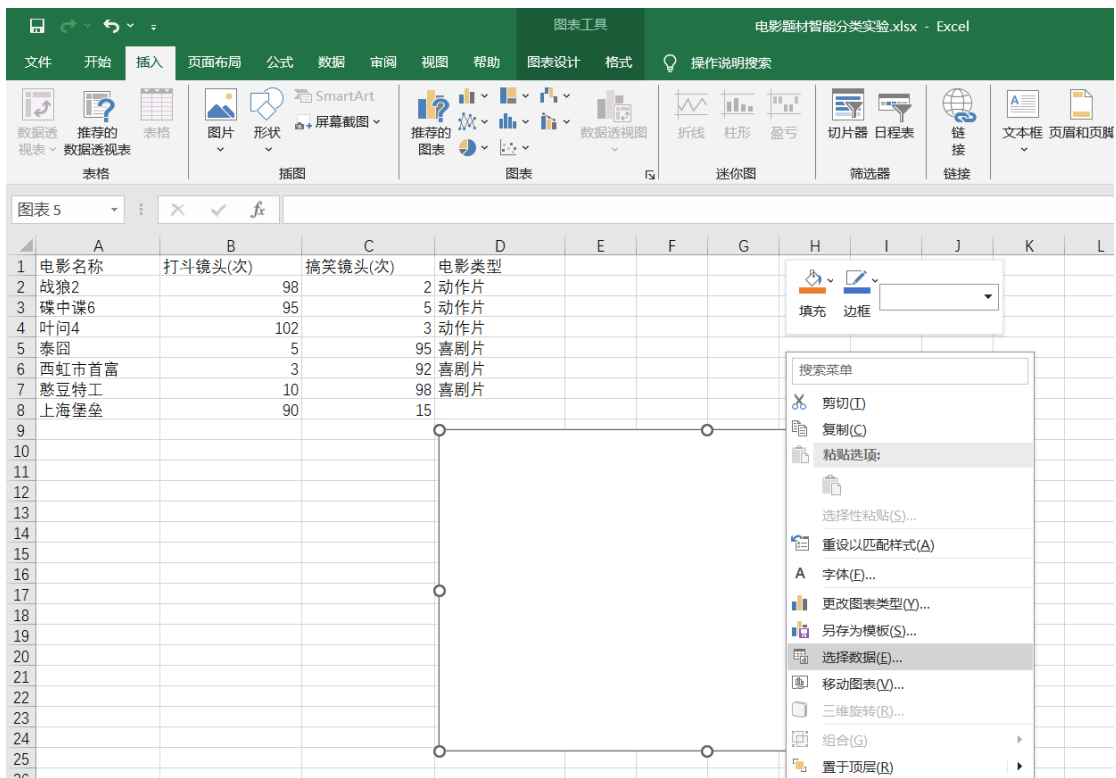
B8: 90

C8: 15

D8: ?

2. 更新散点图数据

右键单击散点图 → 选择「选择数据」。



选择数据

重新选中区域：B1:C8（包含新录入的未知电影特征数据），单击确定。图表自动新增《上海堡垒》对应数据点。图表标题修改为「电影题材分类散点图」。。



更新散点图

根据散点判别观察新点位置，紧邻《战狼 2》、《碟中谍 6》等动作片区域，远离喜剧片区域——数据的位置关系，为分类提供了直接依据。得出结论依据 K-近邻(KNN) 算法：离谁近，就归为哪类。判定《上海堡垒》为：动作片，将“动作片”填入 D8 单

元格，按 Ctrl+S 保存。

任务四：数据干扰测试

1. 录入异常脏数据在表格第 9 行输入错误数据, 模拟人工录错、数据采集失误的场景：

A9：异常电影

B9：100

C9：100

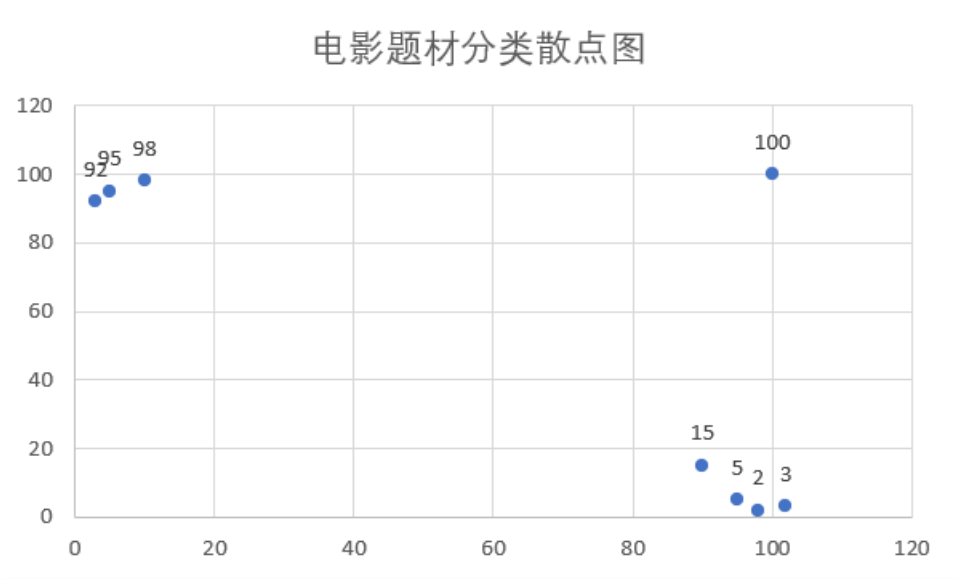
D9：喜剧片

录入异常数据按行依次输入，注意检查输入是否错误。

	A	B	C	D
1	电影名称	打斗镜头(次)	搞笑镜头(次)	电影类型
2	战狼2	98	2	动作片
3	碟中谍6	95	5	动作片
4	叶问4	102	3	动作片
5	泰囧	5	95	喜剧片
6	西虹市首富	3	92	喜剧片
7	憨豆特工	10	98	喜剧片
8	上海堡垒	90	15	动作片
9	异常电影	100	100	喜剧片
10				

新增异常数据

2. 更新图表数据源，右键散点图 → 「选择数据」→ 选中区域 B1:C9 → 「确定」，图表自动新增异常数据点。



散点图更新

3. 观察异常点，异常点出现在图表右上角，远离正常喜剧片区域，喜剧片均为“X 小、Y 大”，该异常点为“X 大、Y 大”，与两类电影的特征规律均不匹配。
4. 若新电影特征为 (95, 95)，机器会因靠近该异常点，误判为喜剧片。得出结论：数据是 AI 的基础，数据质量决定了智能的关键，训练前必须剔除异常值，即前面章节所讲的数据清洗，否则会导致智能决策失效，否则将严重干扰模型的判断逻辑。。

注意事项

训练数据质量直接决定模型效果，存在异常脏数据时容易出现判断偏差；制作图表时准确选择数据源范围，操作完成后及时保存表格文件。