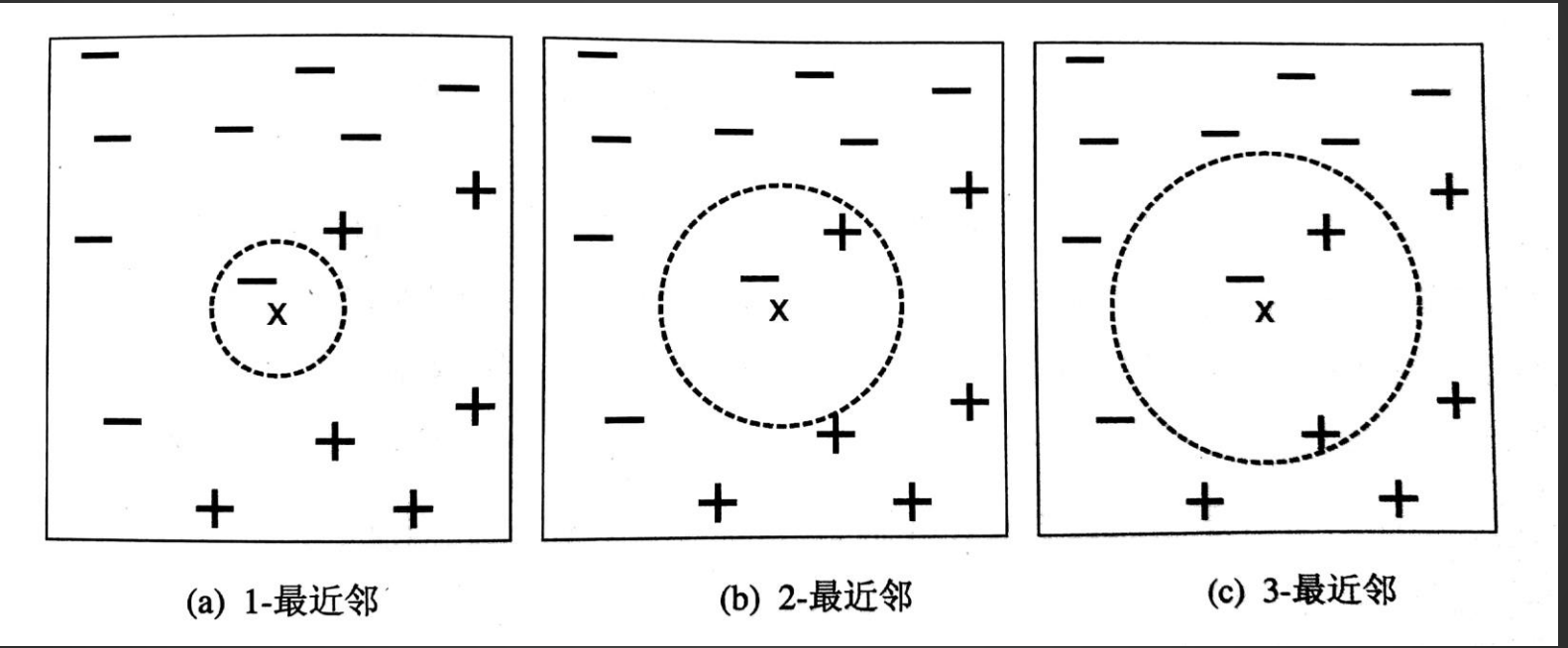
K近邻算法

K近邻算法基本原理

- K近邻，k-Nearest Neighbor，KNN



图片摘自《数据挖掘导论》

KNN实例

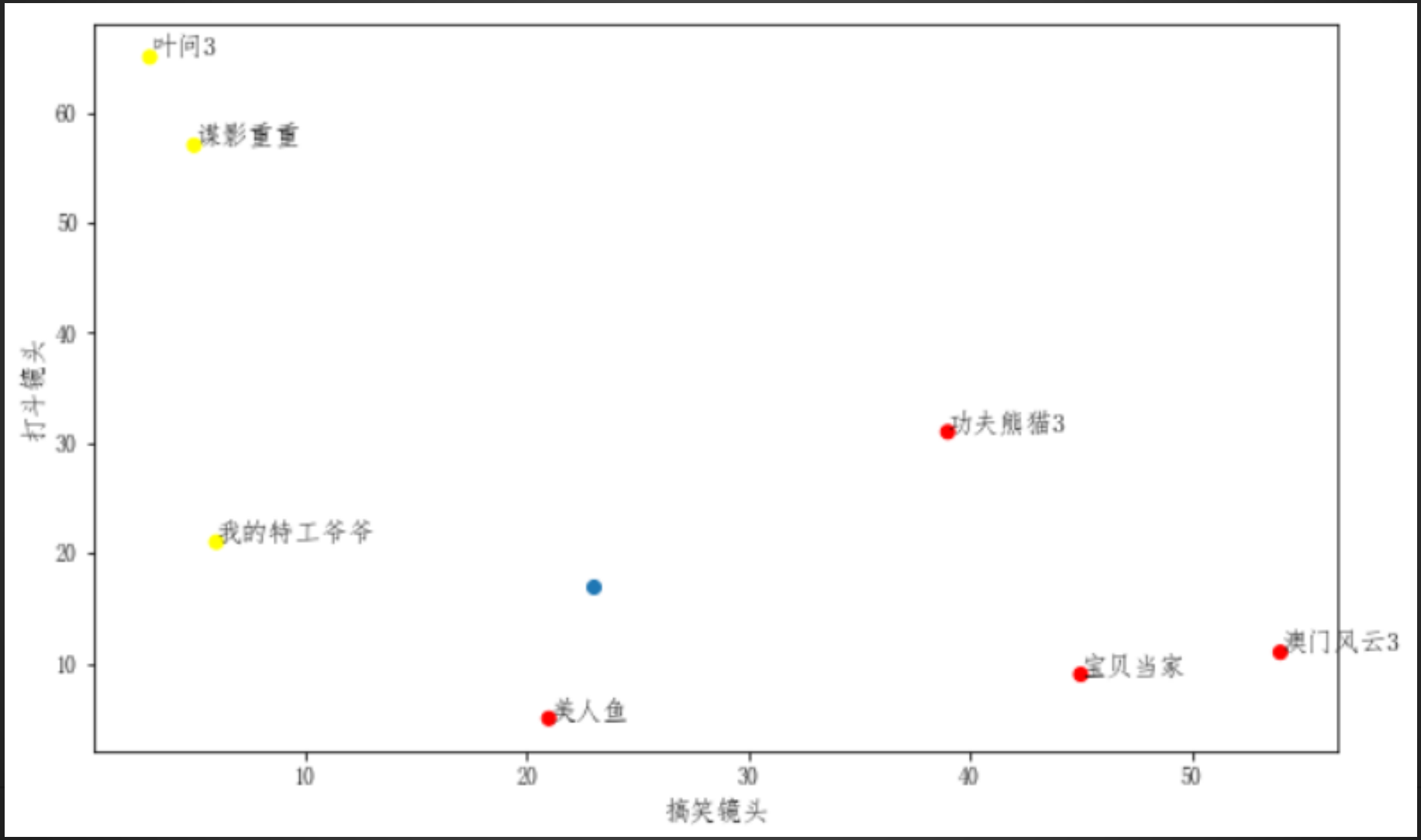
- 任务目标：对电影按照题材进行分类

- 样本数据：7个
- 特征：搞笑镜头、打斗镜头
- 类别：动作片、喜剧片

- 未知电影：《唐人街探案》
- 搞笑镜头：23，打斗镜头：17
- 问题：这部电影属于动作片还是喜剧片?

	电影名称	搞笑镜头	打斗镜头	电影类型
0	谍影重重	5	57	动作片
1	叶问3	3	65	动作片
2	我的特工爷爷	6	21	动作片
3	宝贝当家	45	9	喜剧片
4	美人鱼	21	5	喜剧片
5	澳门风云3	54	11	喜剧片
6	功夫熊猫3	39	31	喜剧片

散点图


叶问3
谍影重重
我的特工爷爷
功夫熊猫3
美人鱼
宝贝当家
澳门风云3
60
50
40
30
20
10
10
20
30
40
50
打斗镜头
搞笑镜头

距离的度量

- 欧式距离

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

- 例如，计算这个未知电影和电影《美人鱼》之间的距离，计算公式为：

$$d = \sqrt{(23 - 21)^2 + (5 - 17)^2} = 12.165525$$

- 按照类似的方式可以将未知样本点和已知样本点的距离都计算出来。

	电影名称	搞笑镜头	打斗镜头	电影类型
0	谍影重重	5	57	动作片
1	叶问3	3	65	动作片
2	我的特工爷爷	6	21	动作片
3	宝贝当家	45	9	喜剧片
4	美人鱼	21	5	喜剧片
5	澳门风云3	54	11	喜剧片
6	功夫熊猫3	39	31	喜剧片

未知电影：《唐人街探案》

搞笑镜头：23，打斗镜头：17

给出结论

- 假定k=3，这三部电影中有2部是喜剧片，1部是动作片，因此我们判定未知电影是喜剧片。

	电影名称	搞笑镜头	打斗镜头	电影类型	距离
4	美人鱼	21	5	喜剧片	12.165525
2	我的特工爷爷	6	21	动作片	17.464249
6	功夫熊猫3	39	31	喜剧片	21.260292
3	宝贝当家	45	9	喜剧片	23.409400
5	澳门风云3	54	11	喜剧片	31.575307
0	谍影重重	5	57	动作片	43.863424
1	叶问3	3	65	动作片	52.000000

代码实现：KNN对电影分类

- 现在需要将前面的理论用代码实现！！！
- 在Jupyter Notebook中写代码!