CDA 数据分析师
CERTIFIED DATA ANALYST

精英班系列课程

KNN

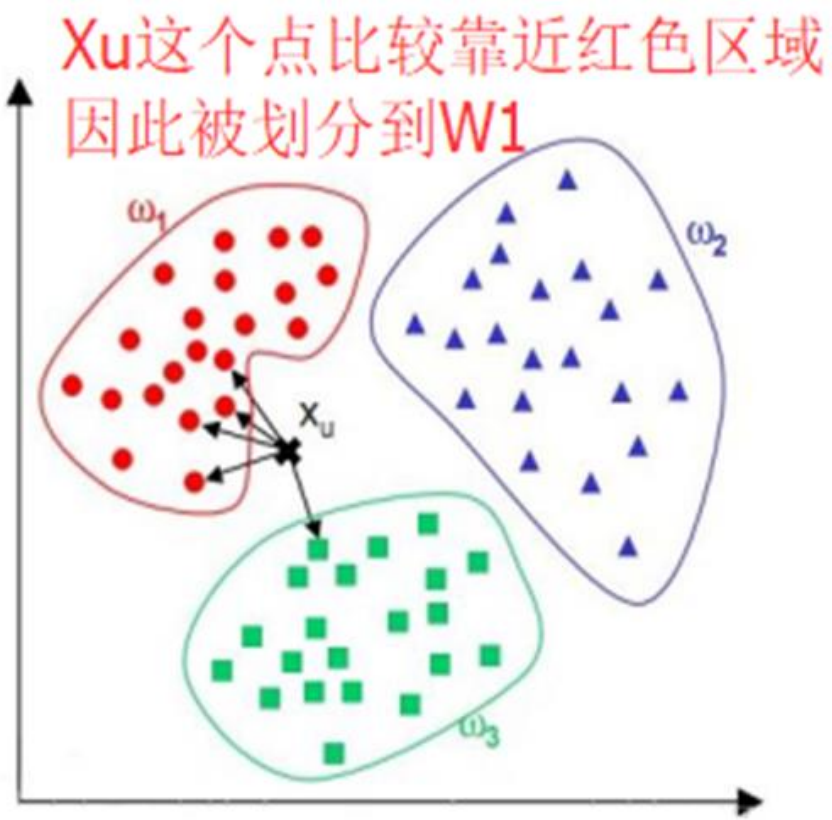
KNN

- K最近邻(kNN, k-Nearest Neighbor):
  - ✓ k近邻法(k-nearest neighbor, k-NN)是1967年由Cover T和Hart P提出的一种基本分类与回归方法。K最近邻分类算法是数据挖掘分类技术中最简单的方法之一。所谓K最近邻，就是k个最近的邻居的意思，每个样本都可以用它最接近的k个邻居来代表
  - ✓ kNN算法的核心思想是如果一个样本在特征空间中的k个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性

KNN

- K最近邻(kNN，k-Nearest Neighbor)：
  - ✓ 工作原理：存在一个样本数据集合，也称作为训练样本集，并且样本集中每个数据都存在标签，即每一个数据与所属分类的一一对应关系。输入没有标签的新数据后，将新的数据的每个特征与样本集中数据对应的特征进行比较，然后算法提取样本最相似数据(最近邻)的分类标签
  - ✓ 一般来说，只选择样本数据集中前k个最相似的数据，这就是k-近邻算法中k的出处，通常k是不大于20的整数。最后，选择k个最相似数据中出现次数最多的分类，作为新数据的分类

- K最近邻(kNN, k-Nearest Neighbor)：



- 距离度量：

欧氏距离

欧氏距离是最常用的一种距离度量方法，源于欧式空间中两点的距离。其计算方法如下：

$$d = \sqrt{\sum_{k=1}^{n}(x_{1k} - x_{2k})^2}$$

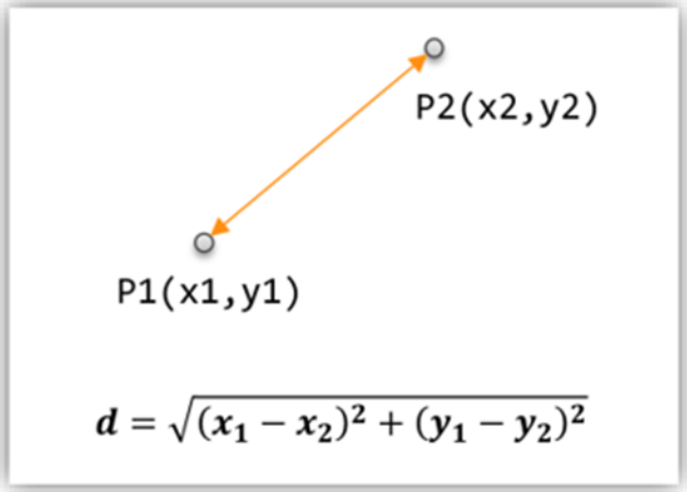


图. 二维空间中欧式距离的计算

- 距离度量：

曼哈顿距离

曼哈顿距离也称作“城市街区距离”，类似于在城市之中驾车行驶，从一个十字路口到另外一个十字楼口的距离。其计算方法如下：

$$d = \sum_{k=1}^{n} |x_{1k} - x_{2k}|^2$$

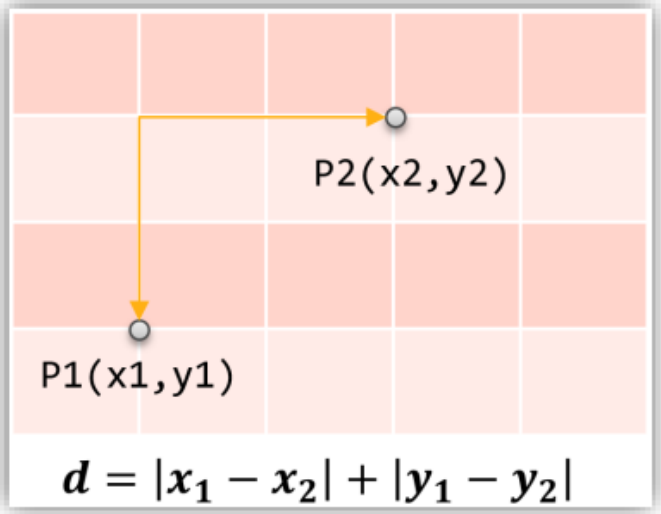


图. 二维空间中曼哈顿距离的计算

- 距离度量：

马氏距离

马氏距离表示数据的协方差距离，是一种尺度无关的度量方式。也就是说马氏距离会先将样本点的各个属性标准化，再计算样本间的距离。其计算方式如下：（s是协方差矩阵，如图）

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T s^{-1} (x_i - x_j)}$$

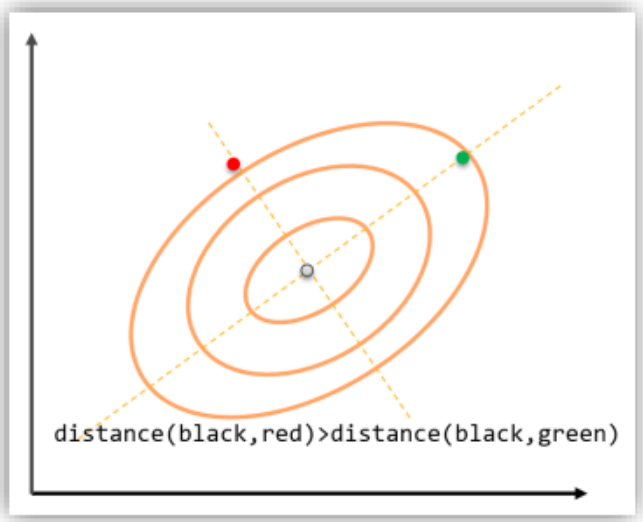


图. 二维空间中的马氏距离

- 距离度量：

夹角余弦

余弦相似度用向量空间中两个向量夹角的余弦值作为衡量两个样本差异的大小。余弦值越接近1，说明两个向量夹角越接近0度，表明两个向量越相似。其计算方法如下：

$$cos(\theta) = \frac{\sum_{k=1}^{n} x_{1k}x_{2k}}{\sqrt{\sum_{k=1}^{n} x_{1k}^2}\sqrt{\sum_{k=1}^{n} x_{2k}^2}}$$

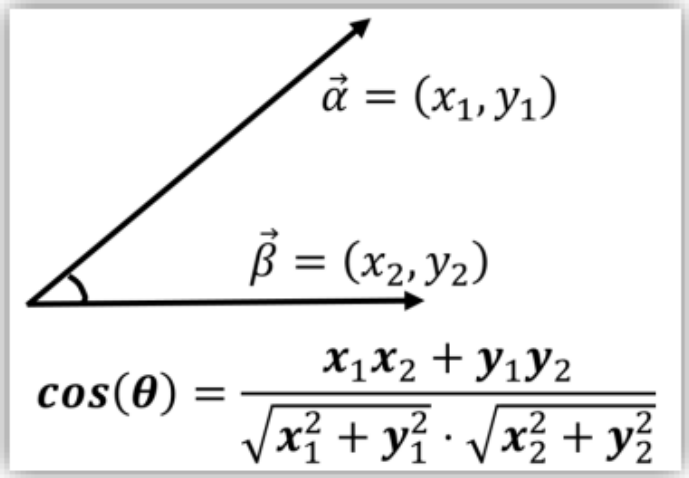


图. 二维空间中的夹角余弦

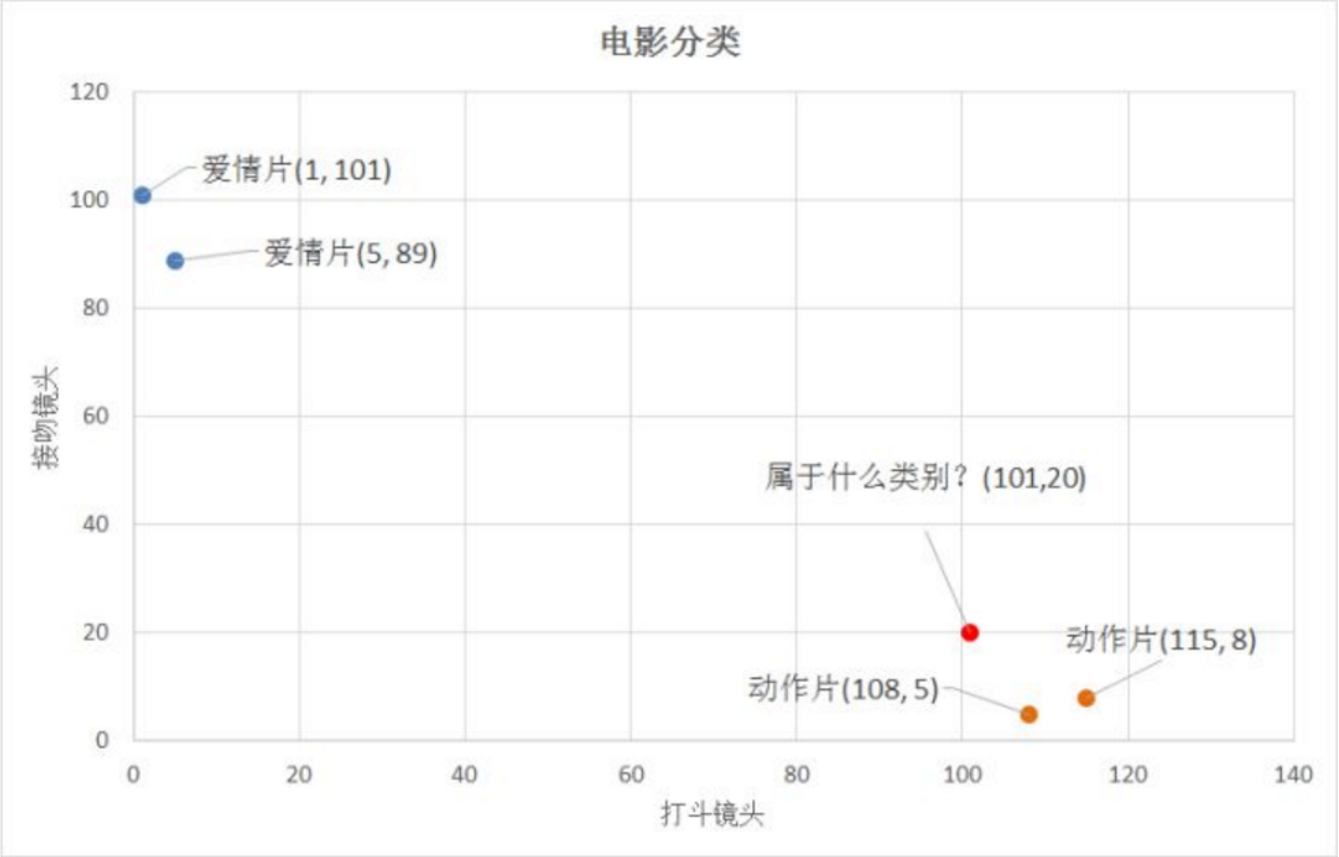
使用knn分类电影是爱情片还是动作片：

电影名称	打斗镜头	接吻镜头	电影类型
电影1	1	101	爱情片
电影2	5	89	爱情片
电影3	108	5	动作片
电影4	115	8	动作片

表1.1 每部电影的打斗镜头数、接吻镜头数以及电影类型

- ✓ 数据集特征：打斗镜头数和接吻镜头数
- ✓ 分类标签：爱情片，动作片

- 电影所属类别距离度量：


电影分类
120
100
80
60
40
20
0
接吻镜头
爱情片(1, 101)
爱情片(5, 89)
属于什么类别？(101,20)
动作片(115, 8)
动作片(108, 5)
0
20
40
60
80
100
120
140
打斗镜头

- 距离度量：

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

- 通过计算，可得：
  - ✓ (101, 20)->动作片(108, 5)的距离约为16.55
  - ✓ (101, 20)->动作片(115, 8)的距离约为18.44
  - ✓ (101, 20)->爱情片(5, 89)的距离约为118.22
  - ✓ (101, 20)->爱情片(1, 101)的距离约为128.69
- 可知，红色圆点标记的电影到动作片 (108, 5)的距离最近，为16.55。如果算法直接根据这个结果，判断该红色圆点标记的电影为动作片，这个算法就是最近邻算法，而非k-近邻算法

- 判决过程：
  - ✓ k取3，按距离依次排序的三个点分别是动作片(108,5)、动作片(115,8)、爱情片(5,89)。在这三个点中，动作片出现的频率为三分之二，爱情片出现的频率为三分之一，所以该红色圆点标记的电影为动作片。这个判别过程就是k-近邻算法

k-近邻算法步骤：

- ✓ 计算已知类别数据集中的点与当前点之间的距离；
- ✓ 按照距离递增次序排序；
- ✓ 选取与当前点距离最小的k个点；
- ✓ 确定前k个点所在类别的出现频率；
- ✓ 返回前k个点所出现频率最高的类别作为当前点的预测分类

- k-近邻算法代码实现电影分类

k-近邻算法实战：约会网站配对效果判定

- ✓ 海伦女士一直使用在线约会网站寻找适合自己的约会对象。尽管约会网站会推荐不同的人选，但她并不是喜欢每一个人。经过一番总结，她发现自己交往过的人可以进行如下分类：
  - 不喜欢的人
  - 魅力一般的人
  - 极具魅力的人

- k-近邻算法实战：约会网站配对效果判定
  - ✓ 海伦收集约会数据已经有了一段时间，她把这些数据存放在文本文件datingTestSet.txt中，每个样本数据占据一行，总共有1000行

- 海伦收集的样本数据主要包含以下3种特征：
  - ✓ 每年获得的飞行常客里程数
  - ✓ 玩视频游戏所消耗时间百分比
  - ✓ 每周消费的冰淇淋公升数

k-近邻算法实战：数据归一化

样本	玩游戏所耗时间百分比	每年获得的飞行常用里程数	每周消费的冰淇淋公升数	样本分类
1	0.8	400	0.5	1
2	12	134000	0.9	3
3	0	20000	1.1	2
4	67	32000	0.1	2

表2.1 约会网站样本数据

$$\sqrt{(0-67)^2+(20000-32000)^2+(1.1-0.1)^2}$$

图2.4 计算公式

k-近邻算法实战：数据归一化

- ✓ 这三种特征是同等重要的，因此作为三个等权重的特征之一，飞行常客里程数并不应该如此严重地影响到计算结果
- ✓ 将任意取值范围的特征值转化为0到1区间内的值：

  newValue = (oldValue - min) / (max - min)

k-近邻算法实战：约会网站配对效果判定

- ✓ 测试算法：评估算法的正确率
- ✓ 一般提供已有数据的90%作为训练样本来训练分类器，而使用其余的10%数据去测试分类器，检测分类器的正确率

- k-近邻算法实战：约会网站配对效果判定
  - ✓ 构建完整可用系统
  - ✓ 构建小程序，约会网站上找到某个人并输入他的信息。程序会给出海伦对男方喜欢程度的预测值

KNN

- sklearn中的KNN
  - ✓ 分类
  - ✓ 回归

Talk is cheap

Show me the

CODE