

最近邻规则分类

KNN (K-Nearest Neighbor)

KNN例子



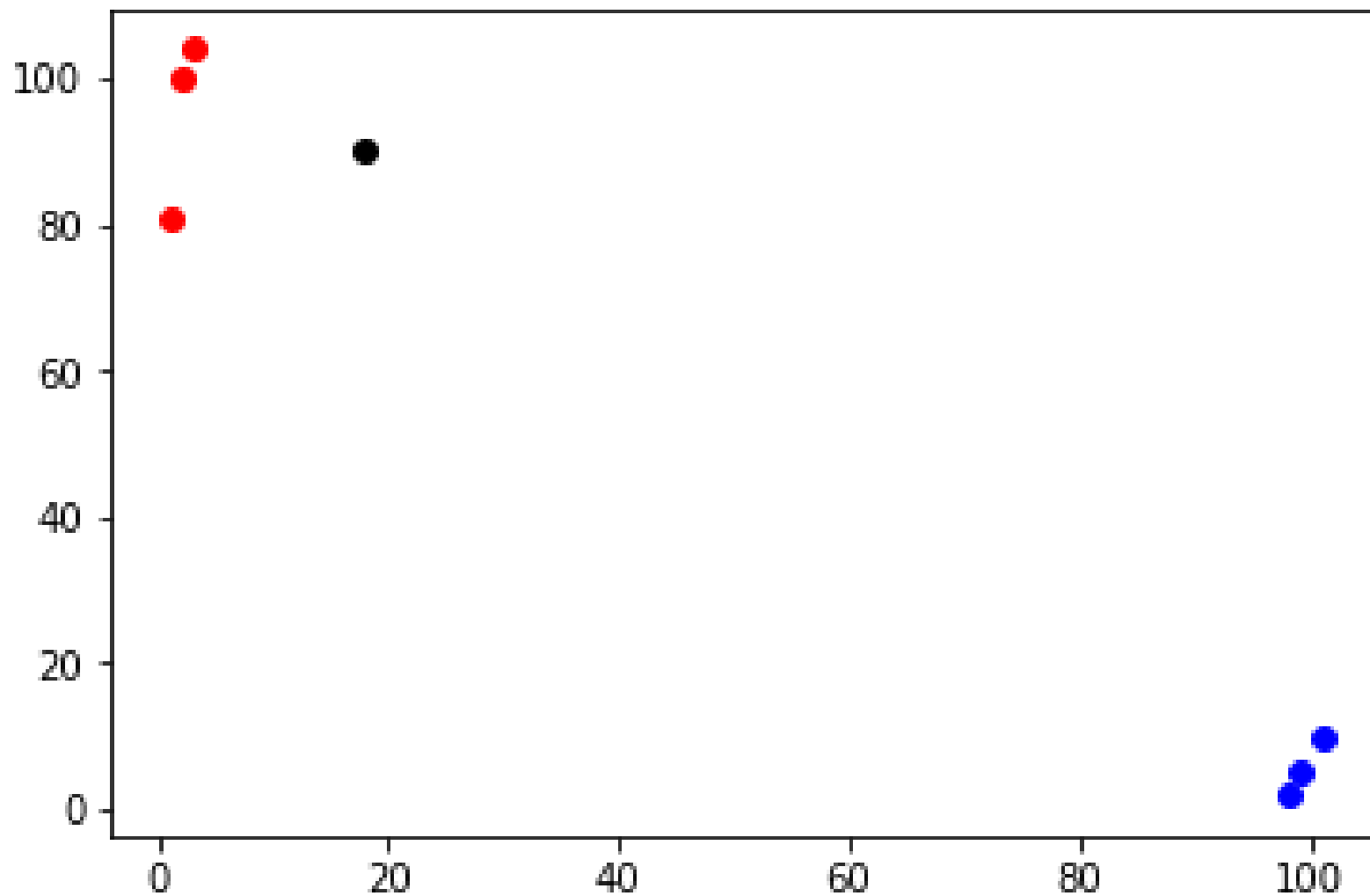
电影名称	打斗次数	接吻次数	电影类型
California Man	3	104	Romance
He's Not Really into Dudes	2	100	Romance
Beautiful Woman	1	81	Romance
Kevin Longblade	101	10	Action
Robo Slayer 3000	99	5	Action
Amped II	98	2	Action
未知	18	90	Unknown

KNN例子



点	X坐标	Y坐标	点类型
A点	3	104	Romance
B点	2	100	Romance
C点	1	81	Romance
D点	101	10	Action
E点	99	5	Action
F点	98	2	Action
G点	18	90	Unknown

KNN例子

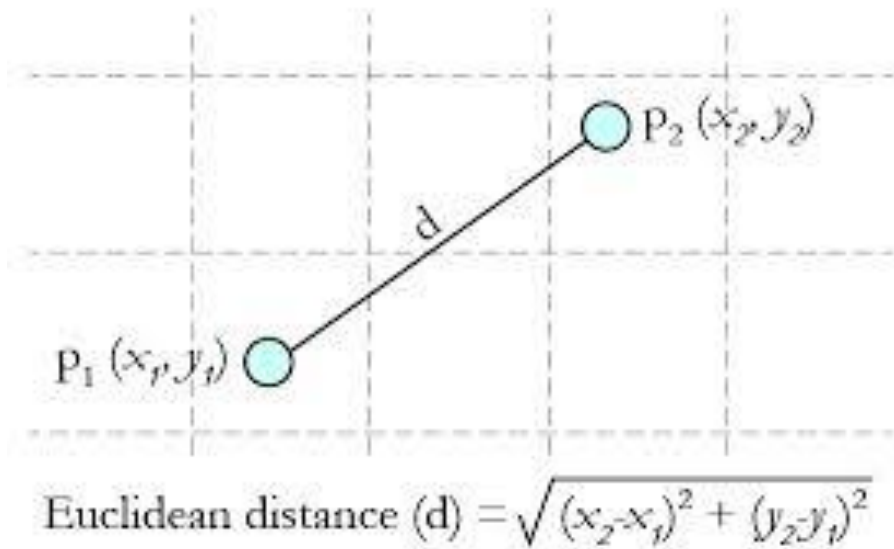




- 为了判断未知实例的类别，以所有已知类别的实例作为参照选择参数K
- 计算未知实例与所有已知实例的距离
- 选择最近K个已知实例
- 根据少数服从多数的投票法则(majority-voting)，让未知实例归类为K个最邻近样本中最多数的类别



欧式距离也称为欧几里得距离



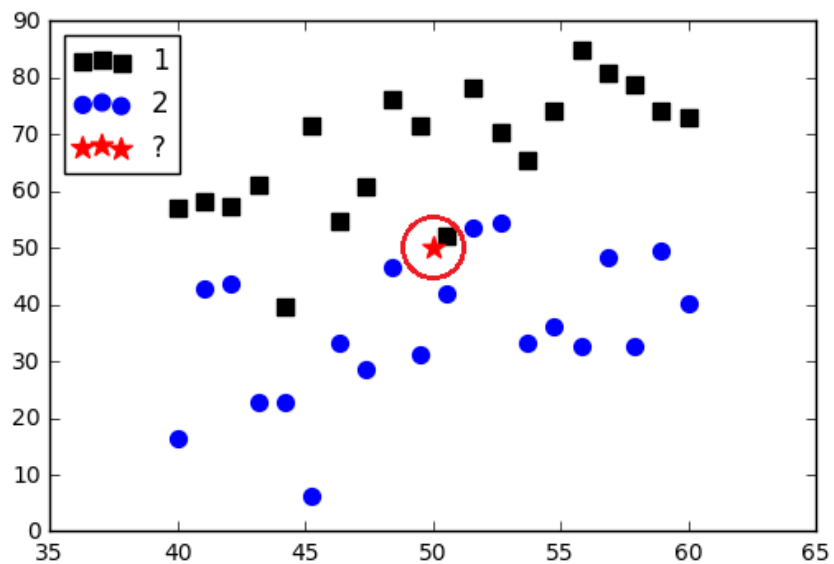
$$E(x, y) = \sqrt{\sum_{i=0}^n (x_i - y_i)^2}$$

其他距离衡量：余弦值距离 (cos), 相关度
(correlation), 曼哈顿距离 (Manhattan distance)
<http://www.cnblogs.com/belfuture/p/5871452.html>

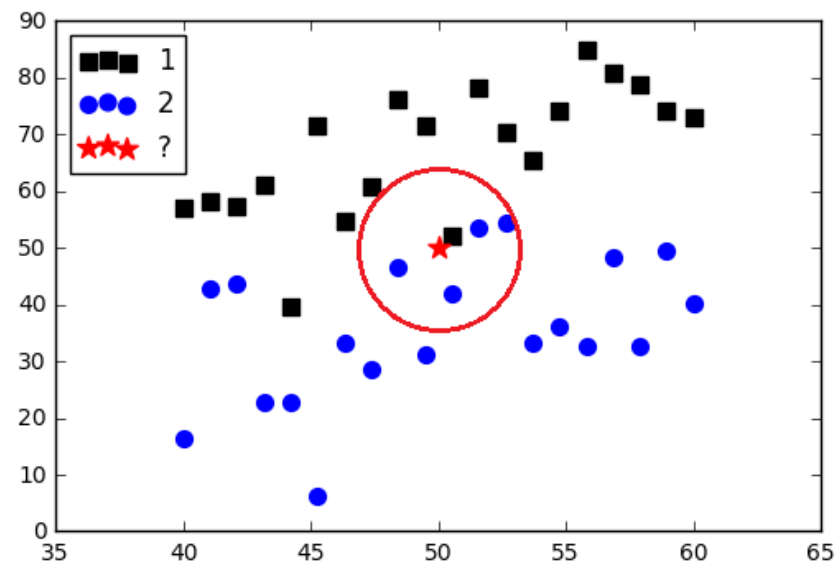
K值选取

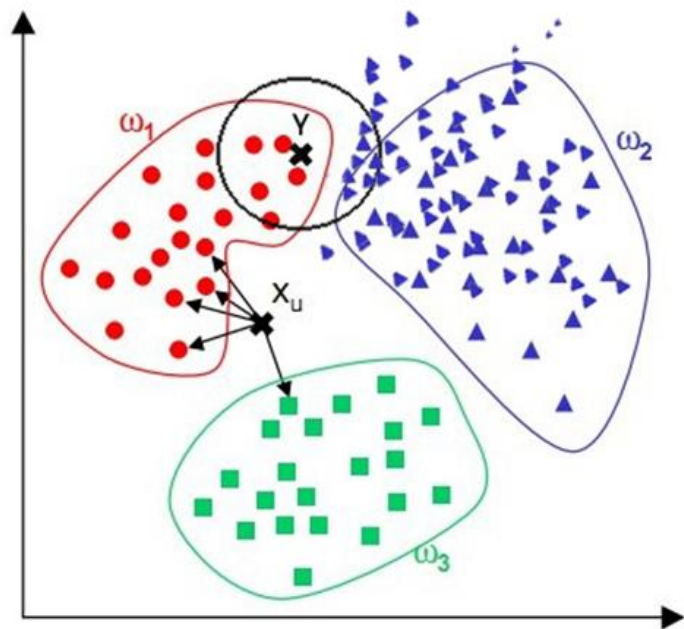


K=1



K=5





- 算法复杂度较高（需要比较所有已知实例与要分类的实例）
- 当其样本分布不平衡时，比如其中一类样本过大（实例数量过多）占主导的时候，新的未知实例容易被归类为这个主导样本，因为这类样本实例的数量过大，但这个新的未知实例实际并没有接近目标样本