K-Nearest Neighbor

2019年6月28日 16:02

最邻近分类.

如电影分类问题.

电影	打斗次数	接吻次数	电影类型.
①	3	104	Romance
②	2	100	Romance
③	100	2	Action
④	1	81	Romance
⑤	99	5	Action
⑥	97	3	Action.

打斗次数 → x坐轴

接吻次数 → y坐轴.

① (3,104) ④ (1,81)

② (2,100) ⑤ (99,5)

③ (100,2) ⑥ (97,3).

现在给出一个点, M(x0, y0)

通过计算M与①~⑥之间的距离, 再定义一个K值(K一般为奇数)

选择K个与M最近的点, 以少数服从多数的方式, 来判断M的类型.

计算距离的方法.

1. Euclidean Distance 即两点间距公式.
2. 余弦值.
3. 相关度.
4. 曼哈顿距离(街区数).

算法优点:

算法优点：

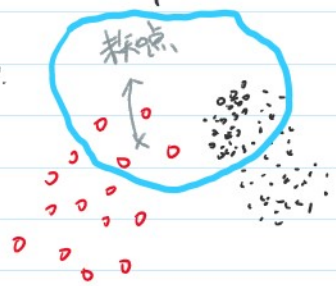
简单、易理解、易实现

通过对K的选择可具备丢噪声数据的健壮性.

算法缺点：

1. 要储备大量的实例.
2. 算法复杂度高.(要比较所有已知实例与要分类的实例)
3. 当样本分布不均时，如其中一个样本过大，实例过多，占优势时，新的未知样本易被归为此类样本，但事实并非如此.

如：



当K的取值过大时
X会被归为黑
但事实上为红.

算法的改进

依距离加上权重，如距离为d，则权重为 1/d.