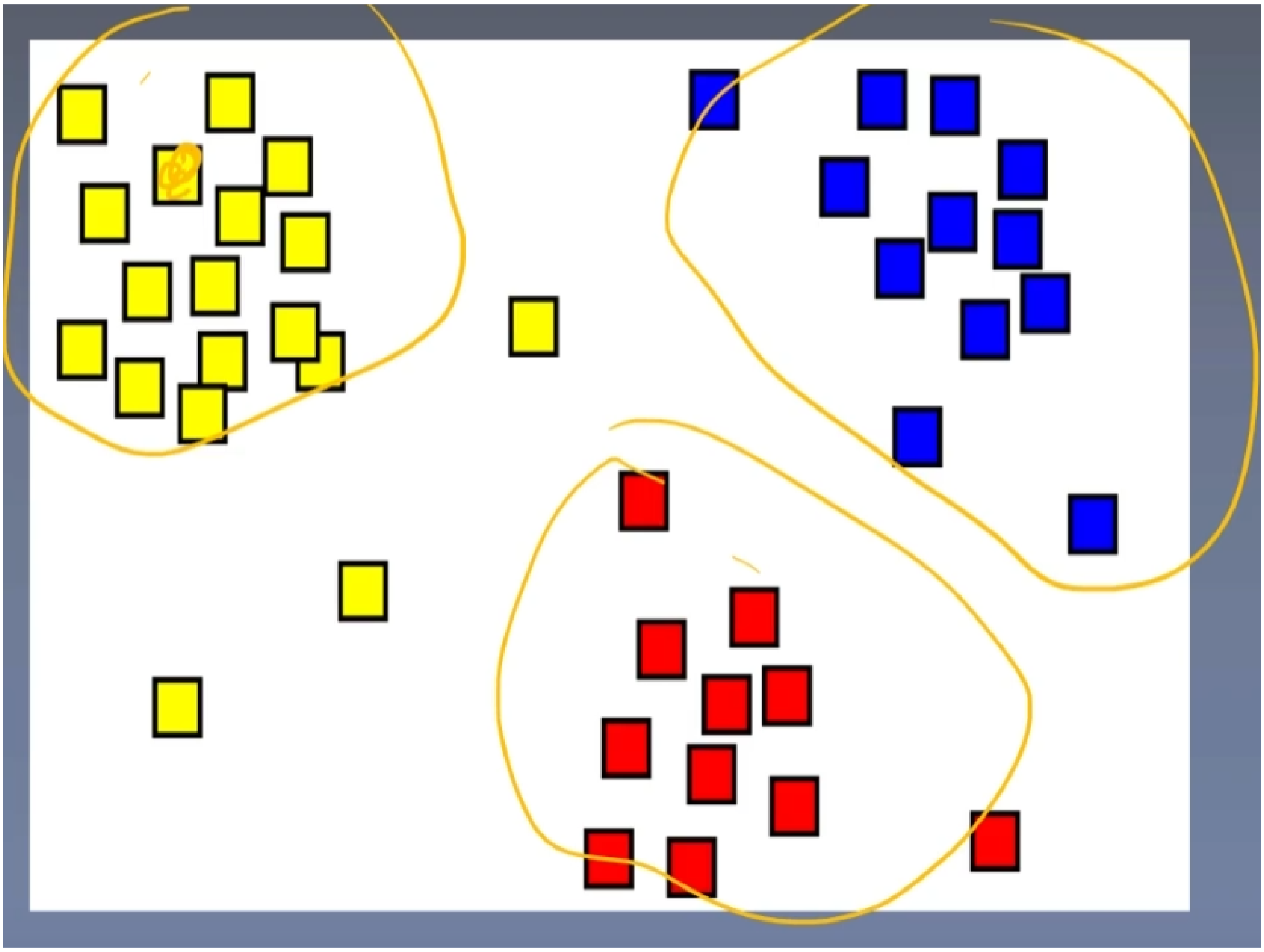
统计学习方法——K近邻法

create：2021/11/2 广州

引言

同一标签的样本通常有很多相似的特征，这就是**物以类聚**的现象，在下图中有三个不同标签的数据，如果一个新的数据进来，我们可以**通过查看新样本周围的样本来确定它的标签。**



算法详解

给定一个数据集，对于新输入的实例，在训练数据中找到与该实例最近邻的k个实例，这k个实例的多数属于某个类，就把该输入实例分到这个类中。

距离度量

上面说到，要找到与输入实例距离最近的k个邻居，那么如何确定距离就成了一个很重要的事情，在书中给出了三个方法：

设$x_i=(x_i^{(1)},x_i^{(2)},x_i^{(3)},\cdots,x_i^{(n)})^T$，$x_j=(x_j^{(1)},x_j^{(2)},x_j^{(3)},\cdots,x_j^{(n)})^T$。

x_i,x_j的L_p距离定义为P范数，公式为：

$$L_p(x_i,x_j)=(\sum_{l=1}^{n}|x_i^{(l)}-x_j^{(l)}|^p)^{\frac{1}{p}}$$

当p=2时，称为欧氏距离，2范数

$$L_2(x_i,x_j)=(\sum_{l=1}^{n}|x_i^{(l)}-x_j^{(l)}|^2)^{\frac{1}{2}}$$

当p=1时，称为曼哈顿距离，1范数

$$L_1(x_i,x_j)=\sum_{l=1}^{n}|x_i^{(l)}-x_j^{(l)}|$$

当$p=\infty$时，是各个坐标距离的最大值，即

$$L_\infty(x_i,x_j)=\max_{l=1}|x_i^{(l)}-x_j^{(l)}|$$

下面这张图展示了二维空间中p取不同值时，与原点的L_p距离为1($L_p=1$)的点的图形。

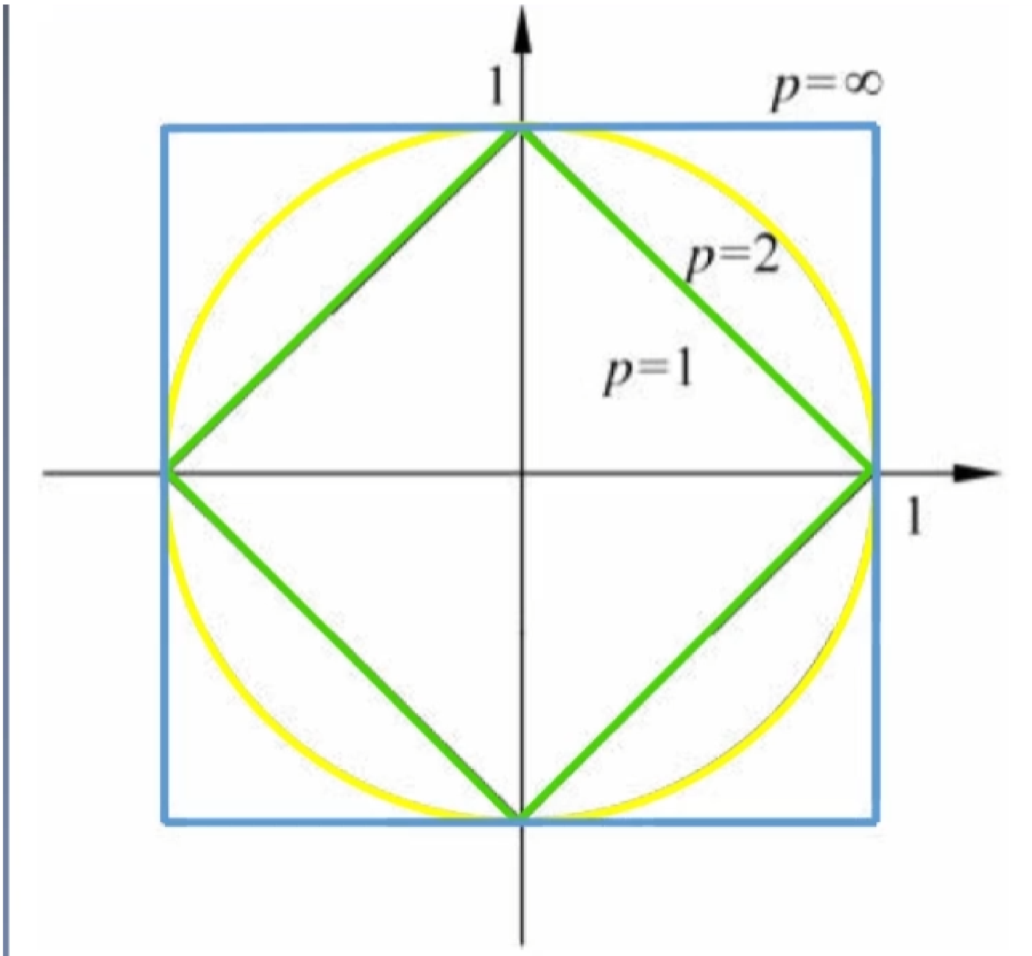
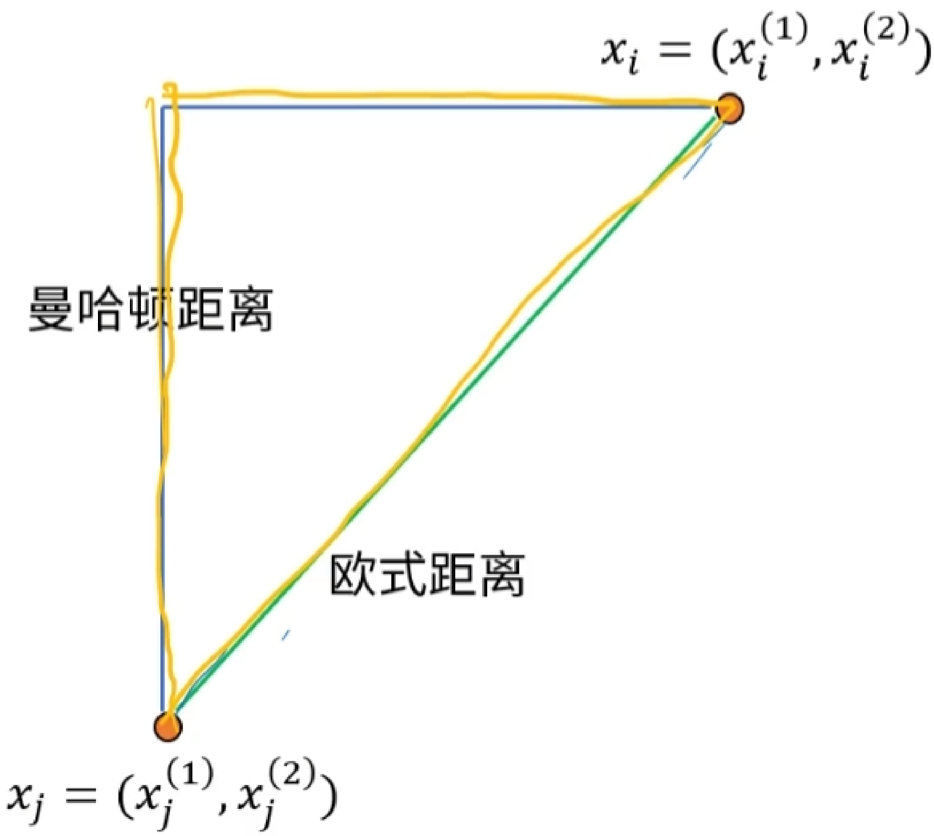


图3.2 L_p距离间的关系

下面的图更为直观的展示了曼哈顿距离与欧式距离的效果



在k近邻算法中，**一般采用欧式距离**

K值的选择

k值过大或者过小都会出现错误的结果

所以在实例应用中，会先选取较小的k值，然后通过交叉验证来得到合适的k值。

分类决策规则

kNN中的分类决策规则通常是多数表决，即由测试样本的k个临近样本的多数类决定测试样本的类别。

假设新输入实例的最近邻k个数据构成集合$N_k(x)$, 分类损失函数为0-1函数，即分错了得分为0，分对了得分为1，$N_k(x)$区域中的类别为c_j，那么误分类率为：

$$\frac{1}{k}\sum_{x_i \in N_k(x)} I(y_i \neq c_j) = 1 - \frac{1}{k}\sum_{x_i \in N_k(x)} I(y_i = c_j)$$

根据上式很容易得知，我们的目标就是要使$\sum_{x_i \in N_k(x)} I(y_i = c_j)$最大，公式中的I就是0-1损失函数。

K近邻算法的实现：kd树

具体可以参考，统计学习方法p41

https://zhuanlan.zhihu.com/p/23966698 这个知乎回答写的很直观

Reference

b站视频—K近邻算法