Python
机器学习实战

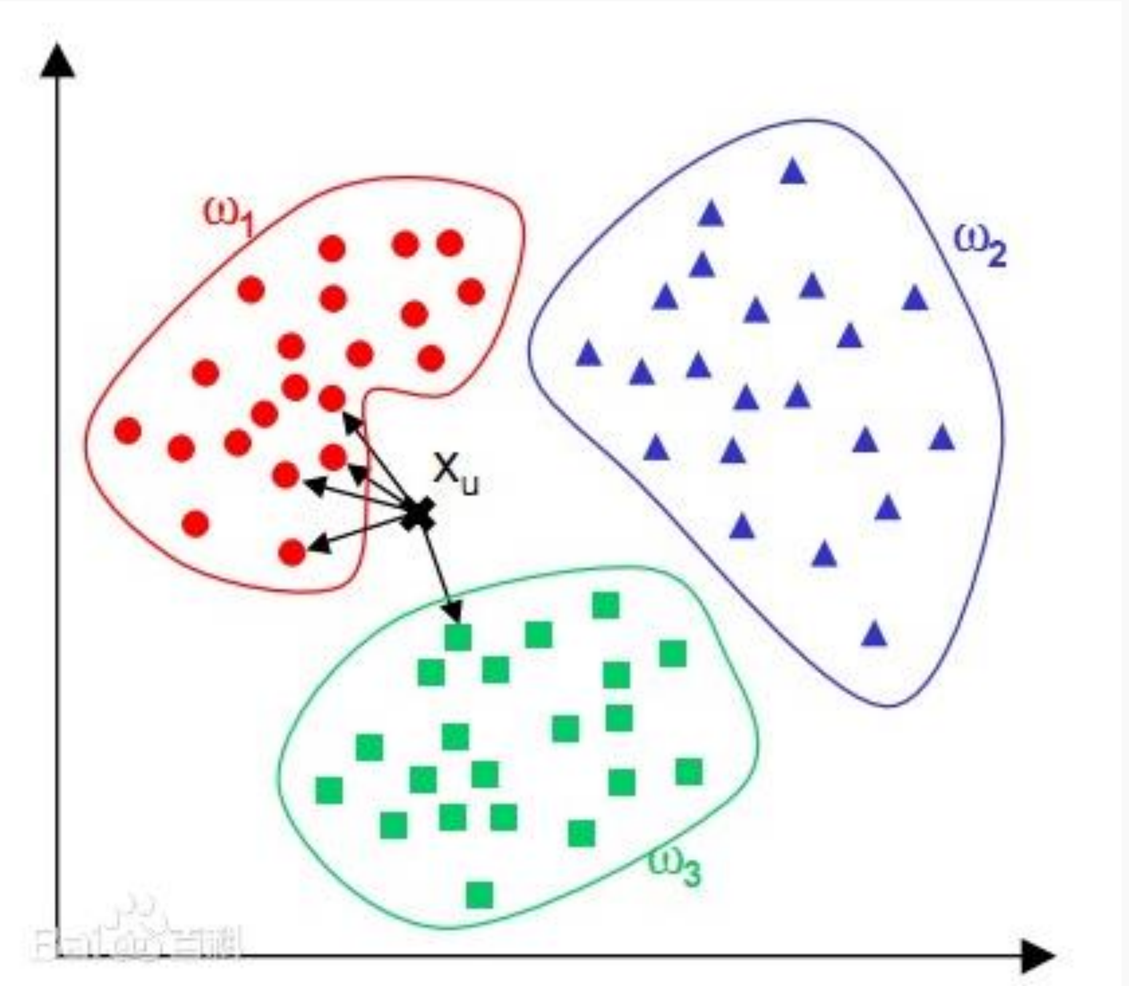
K最近邻算法

K最近邻算法

KNN（K-Nearest Neighbor）

原理

- 数据映射到高维空间中的点
- 找出k个最近的样本
- 投票结果



K最近邻算法

如何衡量距离

数学中距离满足三个要求

- 必须为正数
- 必须对称
- 满足三角不等式

K最近邻算法

闵可夫斯基距离（Minkowski）

- 曼哈顿距离
- 欧式距离
- 切比雪夫距离

$$d_{ij}(q) = \left[\sum_{k=1}^{p}(x_{ik} - x_{jk})^q\right]^{1/q}, \quad q > 0.$$

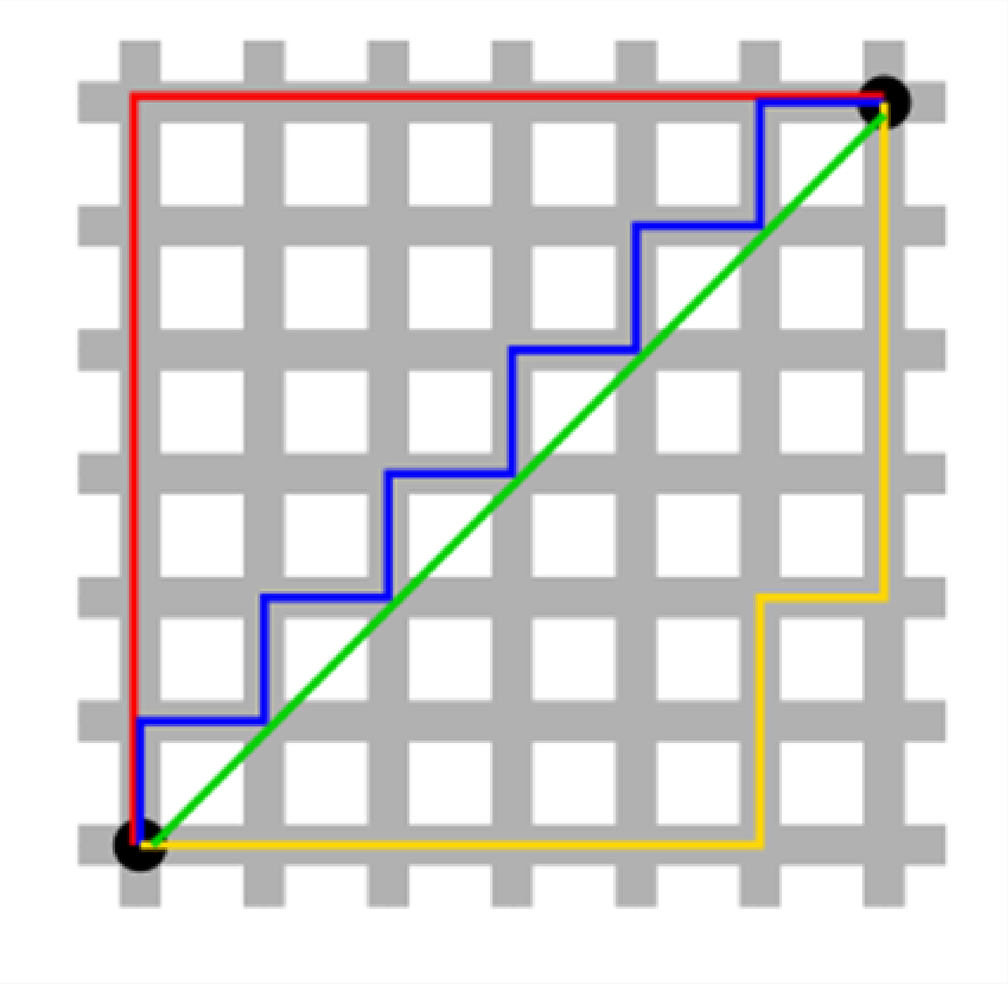
q越大，差异越大的维度对最终距离影响越大

马氏距离

K最近邻算法

曼哈顿距离

$$d(i,j)=|X1-X2|+|Y1-Y2|$$

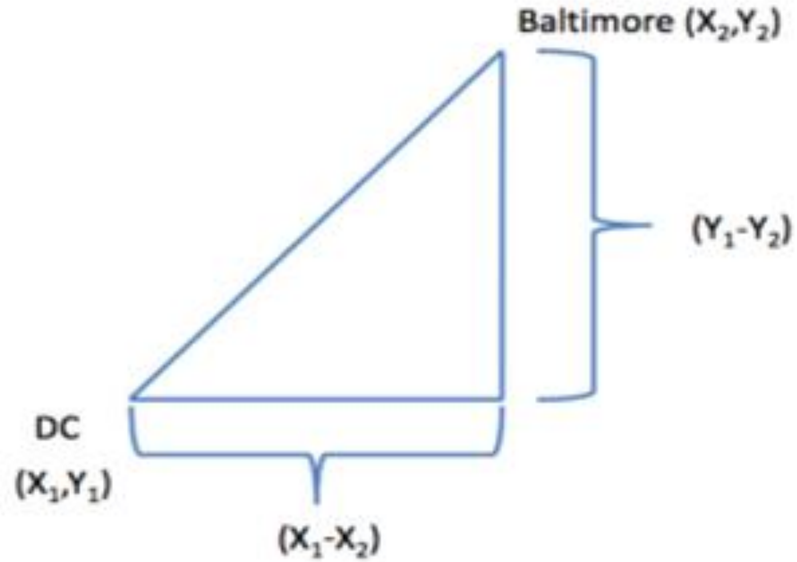


K最近邻算法

欧式距离

$$\sqrt{(A_1 - A_2)^2 + (B_1 - B_2)^2 + \ldots + (Z_1 - Z_2)^2}$$

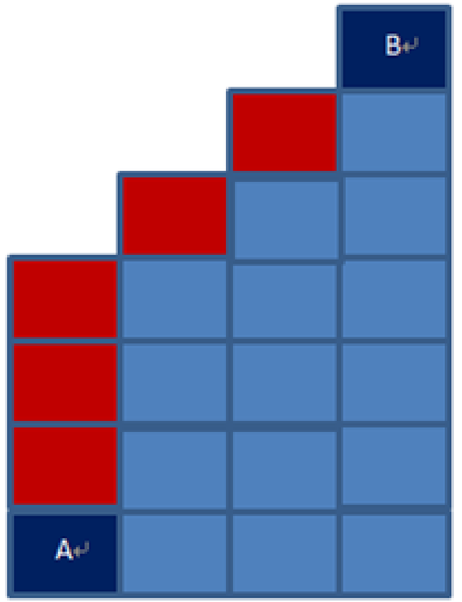
$$\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$$



K最近邻算法

切比雪夫距离

$$dist(X,Y) = \lim_{p\to\infty}(\sum_{i=1}^{n}|xi-yi|^{p})^{1/p} = \max|xi-yi|$$

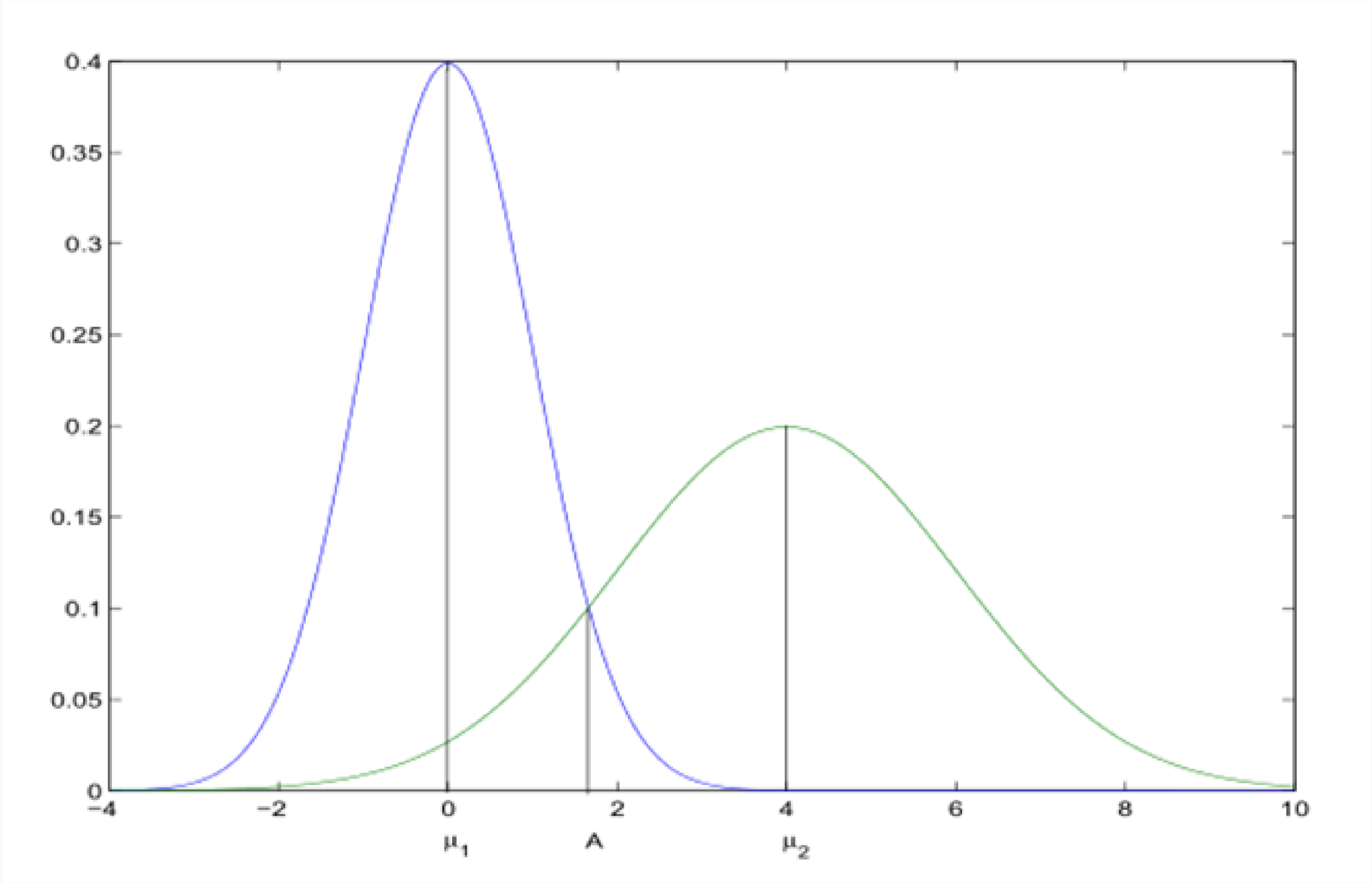


K最近邻算法

马氏距离

马氏距离

- 考虑数据分布



实战

K最近邻算法

鸢尾花分类

训练数据集

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.5,2.8,4.6,1.5,Iris-versicolor
5.7,2.8,4.5,1.3,Iris-versicolor
6.3,3.3,4.7,1.6,Iris-versicolor
4.9,2.4,3.3,1.0,Iris-versicolor
6.6,2.9,4.6,1.3,Iris-versicolor
5.2,2.7,3.9,1.4,Iris-versicolor
```