余弦距离，欧式距离，马氏距离之间的关系

blog.csdn.net/u014453898/article/details/98657357

欧式距离：

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+...+(x_n-y_n)^2}=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

马氏距离：

S为协方差矩阵，当样本集合的协方差矩阵是单位矩阵时，即样本的各个维度上的方差均为１．马氏距离就等于欧式距离相等。

$$d(\vec{x},\vec{y})=\sqrt{(\vec{x}-\vec{y})^T S^{-1}(\vec{x}-\vec{y})}.$$

余弦距离：

$$cos\theta=\frac{\vec{x}\cdot\vec{y}}{||x||\cdot||y||}$$

一，余弦距离和欧式距离：

两个向量间的余弦值可以通过使用欧几里得点积公式求出：

$$\mathbf{a}\cdot\mathbf{b}=\|\mathbf{a}\|\|\mathbf{b}\|\cos\theta.$$

从三维图可以看出：

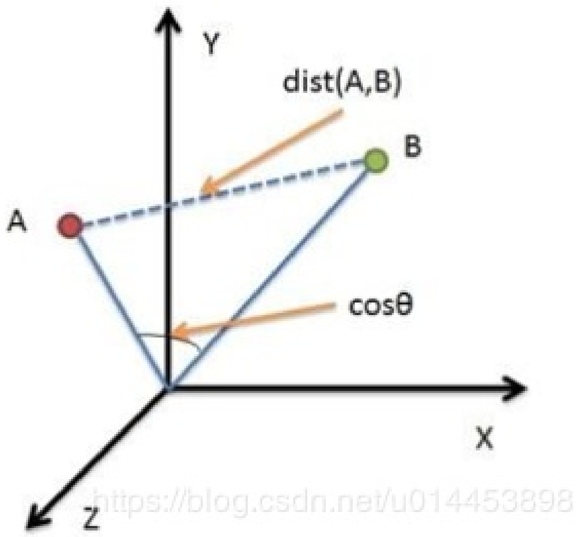
虚线为欧式距离：欧氏距离衡量的是空间各点的绝对距离，跟各个点所在的位置坐标直接相关。

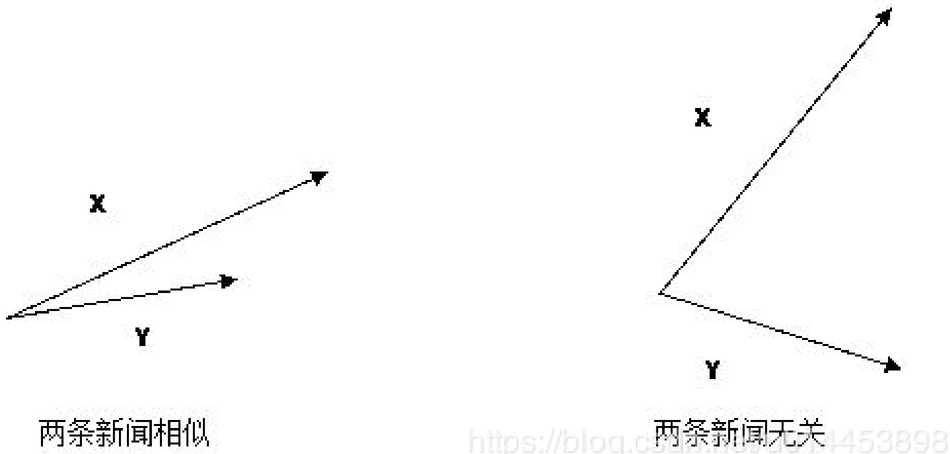
夹角为余弦距离：衡量的是空间向量的夹角，更加体现在方向上的差异，而不是位置。

欧氏距离更倾向于：体现个体数值特征的绝对差异，所以更多的用于需要从维度的数值大小中体现差异的分析

余弦距离：更多的是从方向上区分差异，而对绝对的数值不敏感，更多的用于使用用户对内容评分来区分兴趣的相似度和差异，同时修正了用户间可能存在的度量标准不统一的问题（因为余弦距离对绝对数值不敏感）。

以一个例子说明余弦距离：





x，y向量为两条新闻，x，y的各个维度就相当于该新闻所包含的信息。两向量的夹角越小，两条新闻越相似。这与x，y各个维度的数值无关，至于夹角有关。

假设二维空间两个点，$A(x_1,y_1),B(x_2,y_2)$

然后归一化为**单位向量**，$A\left(\frac{x_1}{\sqrt{x_1^2+y_1^2}},\frac{y_1}{\sqrt{x_1^2+y_1^2}}\right),B\left(\frac{x_2}{\sqrt{x_2^2+y_2^2}},\frac{y_2}{\sqrt{x_2^2+y_2^2}}\right)$

那么余弦相似度就是：

$$cos=\frac{x_1}{\sqrt{x_1^2+y_1^2}}\times\frac{x_2}{\sqrt{x_2^2+y_2^2}}+\frac{y_1}{\sqrt{x_1^2+y_1^2}}\times\frac{y_2}{\sqrt{x_2^2+y_2^2}}$$

（分母是1，省略了）

欧式距离就是：

$$euc=\sqrt{\left(\frac{x_1}{\sqrt{x_1^2+y_1^2}}-\frac{x_2}{\sqrt{x_2^2+y_2^2}}\right)^2+\left(\frac{y_1}{\sqrt{x_1^2+y_1^2}}-\frac{y_2}{\sqrt{x_2^2+y_2^2}}\right)^2}$$

化简后就是：$euc=\sqrt{2-2\times cos}$

因为归一化后，A,B的模为1.

二，马氏距离与欧式距离：

马氏距离是不受量纲影响的欧式距离。

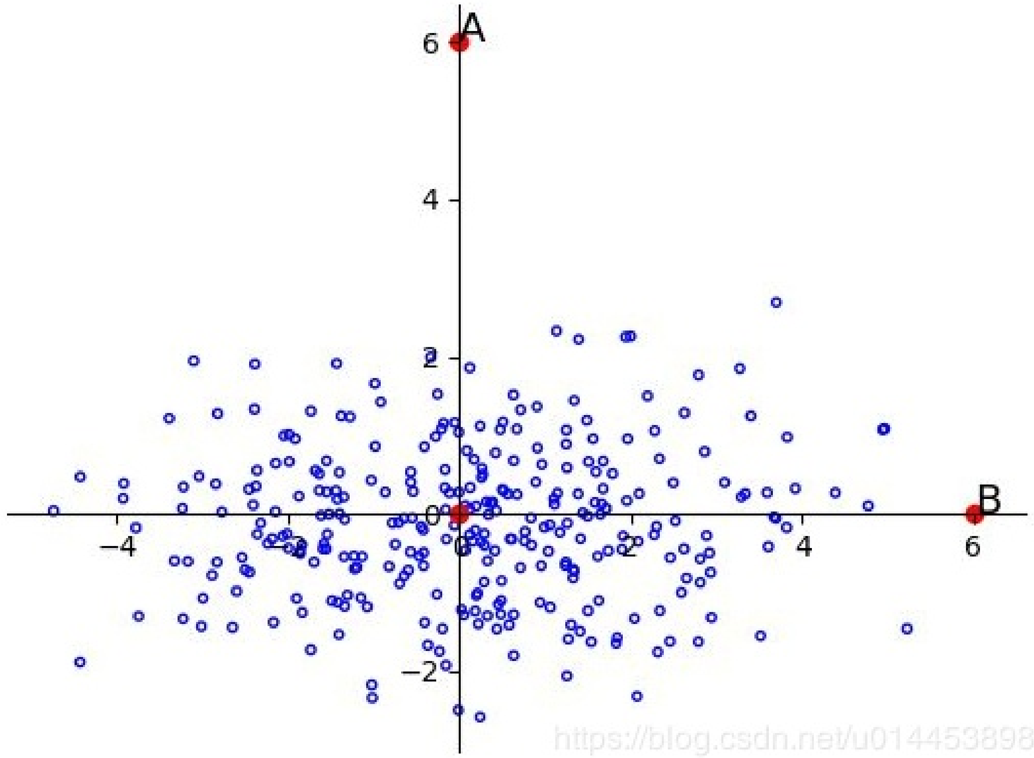
欧式距离的缺点：

若判断A,B两个人哪个适合打篮球：A身高1.8米，弹跳50厘米。B身高2.6米，弹跳20厘米。按照欧式距离，虽然A身高比B少了0.8米，但是A弹跳比B多30厘米啊，所以欧式距离会认为A应该比B适合。但是我们直觉会认为，B两米多高啊，A才1.8米，A的弹跳多那30厘米也没什么 用啊。所以结论就是 欧式距离 容易受量纲影响。

但是归一化后的欧式距离越近，就越相似？

当然我们可以先做归一化来消除这种维度间量纲不同的问题，但是样本分布也会影响分类

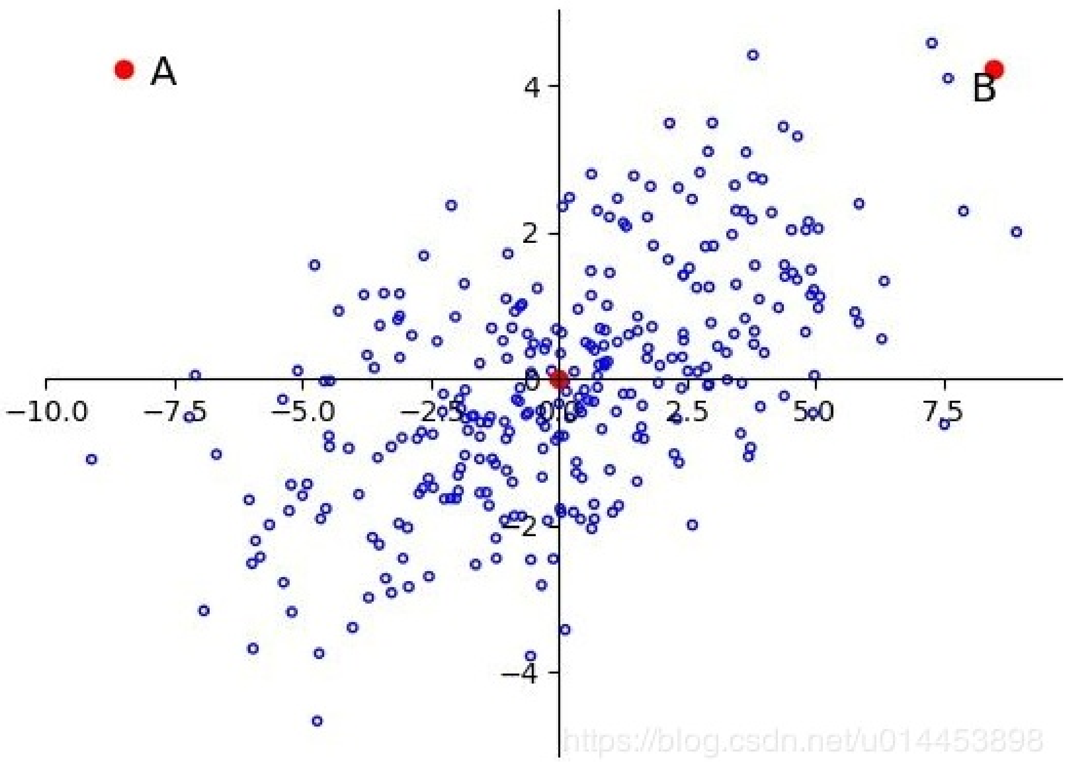
如下图一样，设现在有x，y两个维度，且x，y的均值都为0。A与B相对于原点的欧式距离是相同的。但是由于样本总体沿着横轴分布，所以B点更有可能是这个样本中的点，而A则更有可能是离群点。



因此，也要考虑上方差(数据偏离均值的程度)的影响，上图明显y轴的偏离程度少与x轴的偏离程度。在一个方差较小的维度下很小的差别就有可能成为离群点。

那么当各个维度的方差相同就足够了？

如下图所示：



其中x，y两个维度的方差(离均值0的偏离程度)相同，但是为什么看上去还是A比较像离群点？因为可以看出样本集是服从f(x)=x分布的。所以要马氏距离等于欧式距离，就必须要求样本集是独立分布的。

由于常用的标准化方法为：

标准化后数据 = (原数据 - 均值)/标准差

所以标准化后，均值为0，标准差等于方差等于1.

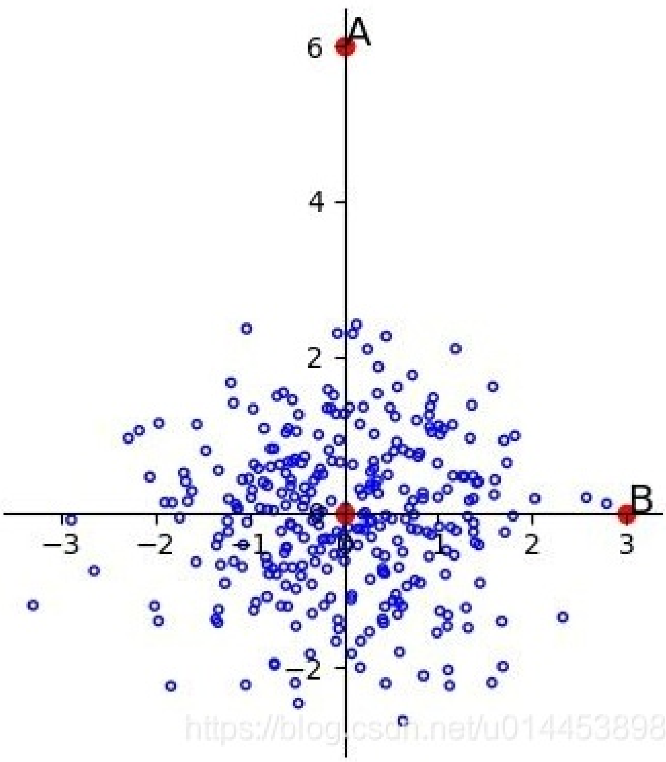
所以要马氏距离等于欧式距离：

就要求数据：

1.各维度独立分布

2.方差相等为1，均值为0，且进行标准化过。

这时候马氏距离就等于欧式距离。
此时图上，B距离比A近，因此可以看作B为样本点，A为离群点。



（Euclidean distance）是一个通常采用的*距离*定义，它是在m维空间中两个点*之间*的真实*距离距离之间*的

，如聚类、KNN，K-means等，使用的*距离*为欧式*距离*。其实，除了欧氏*距离距离*计算标准，本文主要介绍欧氏*距离*和*马氏距离之间*或多点*之间*的*距离*表示法，又称之为欧几里得度量，它定义于欧几里得空间中，如点 x=(x1,...,xn)x = (x_1,...,x_n)x=(x1 ,...,xn) 和 y=(y1,...,yn)y......