聚类算法

机器学习笔记 create by siwanghu v1.0

**K均值算法**，无监督学习算法，用于将相似的样本自动归到一个类别中

欧式距离：

曼哈顿距离：

切比雪夫距离：**d=Max(||，||，，，||)**

Jaccard相似系数：

相关系数：

选择k个点作为初始质心

repeat

将每个点指派到最近的质心，形成k个簇

重新计算每个簇的质心

until 簇不发生变化或达到最大迭代次数

假设使用欧式距离计算数据之间的离散程度，则优化损失函数为：

K表示簇数目，表示每个簇的质心，优化目标是使得最小，求解的位置。对求偏导，令导数等于0，求解的位置,代表每个簇的元素数目

**高斯分布：**

若随机变量服从一个位置参数为，尺度参数为的概率分布，且其概率密度函数为：

简称。

**高斯混合聚类**，聚类算法之一，假设样本中的每个聚类类别服从各自的高斯分布

假设样本为X，高斯混合分布定义如下：

K表示聚类的类别数目（人为指定），代表样本有K个高斯分布。代表每个高斯分布中样本占总样本的比重，而且：

其中：

代表样本中第k个高斯分布的位置参数，代表样本中第k个高斯分布的尺度参数，代表样本中属于第k类高斯分布的样本点占总样本的比重。

算法流程：

1. 首先初始化聚类大小K，然后在初始化每个聚类高斯分布的，，，必须满足所有的相加和为1
2. 依次取出每个样本点，比较样本点x在每个高斯函数中的概率大小，计算公式为：

将样本点分配到概率最大的高斯分布中。

3．用极大似然估计法，重新估计每个聚类中高斯分布的参数，，并且重新计算，也就是重新计算每个聚类数目大小占总样本的比例，计算公式为：

4.反复进行第2步和第3步，知道每个聚类中高斯分布参数更新不明显或者迭代次数结束为止。

**k近邻算法**，如果样本在特征空间中的k个最邻近样本中的大多数样本属于某一个类别，则该样本也属于这个类别，

明可夫斯基距离：

当p=1时，明氏距离即为曼哈顿距离

当p=2时，明氏距离即为欧氏距离

当，明氏距离即为切比雪夫距离

算法步骤：

计算已知类别数据集中的点与当前点之间的距离

按照距离递增次序排序

选取与当前点距离最小的k个点

确定k个点所在类别的出现频率

选择出现频率最大的类别作为当前样本的类别

**常常构造kd树来查找最近邻**