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

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