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： $\mathbf{d} = \sqrt{\sum_{i=1}^{n}(x_i - x_i')^2}$

曼哈顿距离： $\mathbf{d} = \sum_{i=1}^{n}(|x_i - x_i'|)$

切比雪夫距离：$\mathbf{d}=\mathbf{Max}(|x_1 - x_1'|,\ |x_2 - x_2'|,,,\ |x_n - x_n'|)$

Jaccard 相似系数：$\mathbf{J}(\mathbf{A},\mathbf{B}) = \frac{|A\cap B|}{|A\cup B|}$

相关系数：$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$

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：

$$L(c_i) = \sum_{i=1}^{K}\sum_{x\in C_i}(c_i - x)^2$$

K 表示簇数目，c_i表示每个簇的质心，优化目标是使得$L(c_i)$最小，求解c_i的位置。$L(c_i)$对c_i求偏导，令导数等于 0，求解c_i的位置，m_k代表每个簇的元素数目

$$\frac{\partial L(c_i)}{\partial c_i} = \frac{\partial}{\partial c_i}\sum_{i=1}^{K}\sum_{x\in C_i}(c_i - x)^2$$

$$= \sum_{i=1}^{K}\sum_{x\in C_i}\frac{\partial}{\partial c_i}(c_i - x)^2$$

$$= \sum_{i=1}^{K}2(c_i - x) = 0$$

$$\xRightarrow{变换} m_k c_i = \sum_{i=1} x$$

$$c_i = \frac{1}{m_k}\sum_{i=1} x$$