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：**d=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

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

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

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

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

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

高斯分布：

若随机变量服从一个位置参数为μ，尺度参数为σ的概率分布，且其概率密度函数为：

$$\mathrm{f(x)} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\mathrm{x}-\mu)^2}{2\sigma^2}}$$ 简称$\mathrm{X} \sim \mathrm{N}(\mu, \sigma^2)$。

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：

$$p(x) = \sum_{k=1}^{K} \pi_k \mathrm{N}(\mathrm{x}|\mu_k, \sigma_k^2)$$

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。π_k代表每个高斯分布中样本占总样本的比重，而且：

$$\sum_{k=1}^{K} \pi_k = 1$$

其中：

$$\mu_k = \frac{1}{\mathrm{N_k}} \sum_{\mathrm{k=1}} x_k$$

$$\sigma_k^2 = \frac{1}{\mathrm{N_k}} \sum_{\mathrm{k=1}} (x_k - \mu_k)(x_k - \mu_k)^T$$

$$\pi_k = \frac{\mathrm{N_k}}{N}$$

μ_k代表样本中第 k 个高斯分布的位置参数，σ_k^2代表样本中第 k 个高斯分布的尺度参数，π_k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π_k，μ_k，σ_k^2，必须满足所有的π_k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：

$$\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} e^{\frac{(x-\mu_k)(x-\mu_k)^T}{2\sigma_k^2}}$$

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μ_k，σ_k^2，并且重新计算π_k，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：

$$\mu_k = \frac{1}{\mathrm{N_k}}\sum_{\mathrm{k=1}} x_k$$

$$\sigma_k^2 = \frac{1}{\mathrm{N_k}}\sum_{\mathrm{k=1}}(x_k - \mu_k)(x_k - \mu_k)^T$$

$$\pi_k = \frac{\mathrm{N_k}}{N}$$

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：

$$d = \sqrt[p]{\sum_{i=1}^{n}|x_i - y_i|^p}$$

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 p → ∞，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