△ 无监督学习

基本问题
- 聚类
  - 硬聚类：一个样本只属一个类（类间无交集） $z=g_\theta(x)$
  - 软聚类：一个样本可属多个类（类间有交集） $P_\theta(z|x)$ → 给定类别数。 $P(z|x)=P(x|z)P(z)$. 可看作概率模型估计。
- 降维
  - 线性
  - 非线性
  - $z=g_\theta(x)$. 给定低维空间维数。
- 概率（模型）估计：
  - 混合模型
  - 概率图模型 → $z$：图的结构。
  - $P_\theta(x|z)$：观测数据（离散/连续）；隐式结构（离散）

给定模型的结构类型或模型集合，找到最可能生成训练数据的模型结构和参数。

△ 聚类

层次聚类
- 聚合（自下而上）聚类：开始每个样本各属一类，逐渐合并最近两类
- 分裂（自上而下）聚类：开始所有样本同属一类，逐渐将相距最远的样本分到两个新类。

k-means 聚类：基于中心。

样本间距离
- 闵可夫斯基距离
- 马氏距离
- 相关系数
- 夹角余弦 ..

类间距离
- 最短距离
- 最长距离
- 中心距离（中心：均值）
- 平均距离（任两个样本间距离求平均）。

△ 层次聚类 - 聚合聚类

① 计算 $n$ 个样本两两间距离

② 构造 $n$ 个类，每类只有一个样本

③ 合并类间距离最小的两类为一新类

④ 重复③至类个数为 1.

△ k-means 聚类

划分方法为 $C$，$C(x_i)$ 表示 $x_i$ 所在类。

用欧氏距离平方作样本间距离。

target：$C^* = \arg\min_C \sum_{l=1}^{k} \sum_{C(x_i)=l} \|x_i - \bar{x}_l\|^2 = W(C)$

（NP难）

能量。表示相同类中样本相似程度。（越相似，簇类/能量越小）

迭代求解：① 给定K个中心 $(m_1, m_2, \cdots, m_K)$，求划分 $\hat{C}$：

$$\hat{C} = \arg\min_{C} \sum_{l=1}^{K} \sum_{C(i)=l} \|x_i - m_l\|^2$$

得：$\hat{C}$：将每个样本分配到与其最近的中心所在类。

② 给定划分 C，再求各个类中心 $(\hat{m}_1, \cdots, \hat{m}_K)$：

$$\hat{m}_1, \cdots, \hat{m}_K = \arg\min_{m_1, \cdots, m_K} \sum_{l=1}^{K} \sum_{C(i)=l} \|x_i - m_l\|^2$$

得：$\hat{m}_l = \frac{1}{n_l} \sum_{C(i)=l} x_i$ 类中心为该类样本均值。

步骤：① 随机选K个样本点作为初始中心。

② 分配样本至最近中心

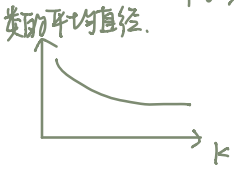
③ 计算新中心（均值）

④ 收敛/符合停止条件，结束。否则返回②。
（划分不再改变）

不保证收敛到全局最优。

初始中心的选择影响结果：可用层次聚类先对样本进行聚类，得K个类时停止，在每个类选择一个与其中心（均值）最接近的样本点。

最优K：实验。




平均直径越小，聚类结果质量越高。
二分查找 找最优K值。