目录

第九章　聚类

常见的无监督学习任务有：聚类 (clustering)，密度估计 (density estimation)，异常检测 (anomaly detection) 等。

0.1　聚类性能度量

聚类的期望结果：簇内相似度 (intra-cluster similarity) 高且簇间相似度 (inter-cluster similarity) 低。

聚类性能度量分为两类：

- 将聚类结果与某个“参考模型”(reference model) 进行比较，称为外部指标 (external index)，主要考虑两个模型分类结果的相似性和相异性；
- 直接考察聚类结果而不利用任何参考模型，称为内部指标 (internal index)，主要考虑类间的“距离”。

0.2　距离计算

用函数 $\mathrm{dist}(\cdot,\cdot)$ 计算两个样本间的距离。定义“距离度量”(distance measure) 是满足下面性质的函数：

- 非负性：$\mathrm{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j)\geq 0$
- 同一性：$\mathrm{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j)=0$ 当且仅当 $\boldsymbol{x}_i=\boldsymbol{x}_j$
- 对称性：$\mathrm{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\mathrm{dist}(\boldsymbol{x}_j,\boldsymbol{x}_i)$
- 直递性 (三角不等式)：$\mathrm{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j)\leq \mathrm{dist}(\boldsymbol{x}_i,\boldsymbol{x}_k)+\mathrm{dist}(\boldsymbol{x}_k,\boldsymbol{x}_j)$

给定样本 $\boldsymbol{x}_i=(x_{i1};x_{i2};\cdots;x_{in})$ 与 $\boldsymbol{x}_j=(x_{j1};x_{j2};\cdots;x_{jn})$，定义闵可夫斯基距离 (Minkowski distance)：

$$\mathrm{dist}_{mk}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\left(\sum_{u=1}^{n}|x_{iu}-x_{ju}|^p\right)^{\frac{1}{p}}$$

对于 $\forall p\geq 1$，闵可夫斯基距离均满足距离度量的性质。$p=1$ 即为曼哈顿距离 (Manhattan distance)，$p=2$ 即为欧氏距离 (Euclidean distance).

连续属性可在属性值上直接计算距离，这样的属性称为“有序属性”，离散属性无法在属性值上计算距离，称为“无序属性”。闵可夫斯基距离可用于有序属性。

对无序属性采用 VDM(Value Difference Metric)。令 $m_{u,a}$ 表示在属性 u 上取值为 a 的样本数，$m_{u,a,i}$ 表示在第 i 个样本簇中在属性 u 上取值为 a 的样本数，k 为样本簇数，则属性 u 上两个离散值 a 与 b 之间的 VDM 距离为：

$$\mathrm{VDM}_p(a,b)=\sum_{i=1}^{k}\left|\frac{m_{u,a,i}}{m_{u,a}}-\frac{m_{u,b,i}}{m_{u,b}}\right|^p$$

绝对值里即是每个样本簇上属性 u 取值为 a 的比例减去每个样本簇上属性 u 取值为 b 的比例。

0.3 原型聚类

假设聚类结构能通过一组原型刻画，算法先对原型进行初始化，然后对原型进行迭代更新求解。

0.3.1 k 均值算法

给定样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m\}$,$k$ 均值 (k-means) 算法针对聚类所得簇划分 $\mathcal{C}=\{C_1,C_2,\cdots,C_k\}$ 最小化平方误差:

$$E=\sum_{i=1}^{k}\sum_{\boldsymbol{x}\in C_i}\|\boldsymbol{x}-\boldsymbol{\mu}_i\|_2^2 \tag{1}$$

其中 $\boldsymbol{\mu}_i=\frac{1}{|C_i|}\sum_{\boldsymbol{x}\in C_i}\boldsymbol{x}$ 是簇 C_i 的均值向量。E 的值越小，簇内样本相似度越高。

最小化式子 (1) 是 NP 难问题，k 均值算法采用贪心策略，通过迭代优化来近似求解。

---

k 均值算法

输入: 样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m\}$，聚类簇数 k。

从 D 中随机选择 k 个样本作为初始均值向量 $\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_k\}$
repeat
  令 $C_i=\varnothing(1\le i\le k)$
  for $j=1,2,\cdots,m$ **do**
    计算样本 $\boldsymbol{x}_j$ 与各均值向量 $\boldsymbol{\mu}_i(1\le i\le k)$ 的距离: $d_{ji}=\|\boldsymbol{x}_j-\boldsymbol{\mu}_i\|_2$;
    根据距离最近的均值向量确定 $\boldsymbol{x}_j$ 的簇标记: $\lambda_j=\arg\min_{i\in\{1,2,\cdots,k\}} d_{ji}$;
    将样本 $\boldsymbol{x}_j$ 划入相应的簇: $C_{\lambda_j}=C_{\lambda_j}\cup\{\boldsymbol{x}_j\}$;
  end for
  for $i=1,2,\cdots,k$ **do**
    计算新均值向量: $\boldsymbol{\mu}_i'=\frac{1}{|C_i|}\sum_{\boldsymbol{x}\in C_i}\boldsymbol{x}$
    更新均值向量: $\boldsymbol{\mu}_i=\boldsymbol{\mu}_i'$
  end for
until 当前均值向量均未更新

输出: 簇划分 $\mathcal{C}=\{C_1,C_2,\cdots,C_k\}$

0.3.2 学习向量量化

与一般聚类算法不同的是，学习向量量化 (Learning Vector Quantization，简称 LVQ) **假设数据样本带有标记信息**。

0.4 密度聚类

密度聚类算法假设聚类结构能通过样本分布的紧密程度确定，从样本密度的角度考察样本之间的连续性，并基于可连接样本不断扩展聚类簇以获得最终的聚类结果。

DBSCAN 算法 (Density-Based Spatial Clustering of Applications with Noise) 基于“邻域”刻画样本分布的紧密程度。

- $\boldsymbol{x}_j$ 的 ϵ-邻域是指样本集 D 中与 $\boldsymbol{x}_j$ 的距离不大于 ϵ 的样本构成的集合。
- 如果 $\boldsymbol{x}_j$ 的 ϵ-邻域至少包含 $MinPts$ 个样本，则称 $\boldsymbol{x}_j$ 是一个核心对象。
- $\boldsymbol{x}_j$ 在 $\boldsymbol{x}_i$ 的 ϵ-邻域中，且 $\boldsymbol{x}_i$ 是核心对象，则称 $\boldsymbol{x}_j$ 可由 $\boldsymbol{x}_i$ 密度直达。**通常密度直达关系不满足对称性**。

LVQ 算法

输入: 样本集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$;
原型向量个数 q，各原型向量预设的类别标记 $\{t_1, t_2, \cdots, t_q\}$;
学习率 $\eta \in (0, 1)$。

初始化一组原型向量 $\{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_q\}$，可以随机选择 q 原型样本
repeat
  从样本集 D 中随机选择样本 $(\boldsymbol{x}_i, y_j)$;
  计算样本 $\boldsymbol{x}_j$ 与 $\boldsymbol{p}_i (1 \leq i \leq q)$ 的距离: $d_{ji} = \|\boldsymbol{x}_j - \boldsymbol{p}_i\|^2$
  找出与 $\boldsymbol{x}_j$ 距离最近的原型向量 $\boldsymbol{p}_{i*}$
  if $y_j = t_{i*}$ **then**
    $\boldsymbol{p}' = \boldsymbol{p}_{i*} + \eta \cdot (\boldsymbol{x}_j - \boldsymbol{p}_{i*})$ $\boldsymbol{x}_j$ 与 $\boldsymbol{p}_{i*}$ 的类别相同
  else
    $\boldsymbol{p}' = \boldsymbol{p}_{i*} - \eta \cdot (\boldsymbol{x}_j - \boldsymbol{p}_{i*})$ $\boldsymbol{x}_j$ 与 $\boldsymbol{p}_{i*}$ 的类别不同
  end if
  将原型向量 $\boldsymbol{p}_{i*}$ 更新为 $\boldsymbol{p}'$
until 满足停止条件，如迭代的次数等。

输出: 簇划分 $\mathcal{C} = \{C_1, C_2, \cdots, C_k\}$

- $\boldsymbol{x}_j$ 可由 $\boldsymbol{x}_i$ 密度可达是指存在从 $\boldsymbol{x}_i$ 到 $\boldsymbol{x}_j$ 的密度直达序列。
- $\boldsymbol{x}_j$ 和 $\boldsymbol{x}_i$ 密度相连是指存在 $\boldsymbol{x}_k$ 使得 $\boldsymbol{x}_j$ 和 $\boldsymbol{x}_i$ 均可由 $\boldsymbol{x}_k$ 密度可达。

DBSCAN 的"簇"定义为由密度可达关系导出的最大的密度相连样本集合: **把一个点所有可达的点加入这个集合，并把可达点的所有可达点也加入**。初始点是核心对象集中的随机一个。

0.5 层次聚类

数据集的划分策略: 自底向上的聚合策略和自顶向下的分拆策略。

AGNES(AGglomerative NESting 的缩写) 是一种自底向上聚合策略的层次聚类算法: 先将数据集中的每个样本看做一个初始聚类簇，然后在算法运行的每一步中找出**距离最近**两个聚类簇进行合并，不断重复，直至达到预设的聚类簇个数。

两个聚类簇的距离有: 最小距离 (两个簇的最近样本距离确定)，最大距离 (两个簇的最远样本距离确定)，平均距离 (两个簇的所有样本共同确定)。相应的 AGNES 算法依次称为单链接，全链接和均链接算法。

0.6 常用的距离

0.6.1 切比雪夫距离

切比雪夫距离 (Chebyshev distance) 是空间中的一种度量，二个点之间的距离定义为其各坐标数值差的最大值。

$$D(\boldsymbol{p}, \boldsymbol{q}) = \max_i |\boldsymbol{p}_i - \boldsymbol{q}_i| = \lim_{k \to \infty} \left(\sum_{i=1}^{n} |\boldsymbol{p}_i - \boldsymbol{q}_i|^k \right)^{\frac{1}{k}}$$

0.6.2 马氏距离

马氏距离 (Mahalanobis distance) 表示数据的协方差距离。它是一种有效的计算两个未知样本集的相似度的方法。与欧氏距离不同的是它考虑到各种特性之间的联系（例如: 一条关于身高的信息会带来一条关于体重的信息，因为两者是有关联的）并且是尺度无关的 (scale-invariant)，即独立于测量尺度。

有 m 个样本向量 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m$，协方差矩阵记为 S，均值记为向量 $\boldsymbol{\mu}$，则其中一个样本向量 $\boldsymbol{x}_i$ 到 $\boldsymbol{\mu}$ 的马氏距离表示为：

$$D(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu})^T S^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})}$$

如果协方差矩阵为单位矩阵，马氏距离就简化为欧式距离，如果协方差矩阵为对角阵，其也可称为正规化的马氏距离。马氏距离的优缺点：**量纲无关，排除变量之间的相关性的干扰**。

0.6.3 余弦相似度

余弦相似度 (Cosine Similarity)，又称为余弦相似性，是通过计算两个向量的夹角余弦值来评估他们的相似度。余弦相似度将向量根据坐标值，绘制到向量空间中，如最常见的二维空间。

$$\cos(\theta) = \frac{\sum_{i=1}^{n} \boldsymbol{x}_{1i} \times \boldsymbol{x}_{2i}}{\sqrt{\sum_{i=1}^{n} \boldsymbol{x}_{1i}^2} \times \sqrt{\sum_{i=1}^{n} \boldsymbol{x}_{2i}^2}}$$

两个向量有相同的指向时，余弦相似度的值为 1；两个向量夹角为 90° 时，余弦相似度的值为 0；两个向量指向完全相反的方向时，余弦相似度的值为-1。这结果是**与向量的长度无关的，仅仅与向量的指向方向相关**。

0.6.4 汉明距离

汉明距离 (Hamming distance)：两个等长字符串 s1 与 s2 之间的汉明距离定义为将其中一个变为另外一个所需要作的最小替换次数。例如字符串“1111”与“1001”之间的汉明距离为 2。可以利用动态规划求解。

应用：信息编码 (为了增强容错性，应使得编码间的最小汉明距离尽可能大)。

0.6.5 杰卡德相似系数

杰卡德相似系数 (Jaccard similarity coefficient) 表示两个集合 A 和 B 的交集元素在 A 与 B 的并集中所占的比例，用符号 $J(A, B)$ 表示。

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

杰卡德相似系数是衡量两个集合的相似度一种指标。

与杰卡德相似系数相反的概念是杰卡德距离 (Jaccard distance)。杰卡德距离可用如下公式表示：

$$J_\delta(A, B) = 1 - J_{A,B}$$

0.6.6 皮尔森相关系数

皮尔森相关系数 (Pearson correlation coefficient)：也称皮尔森积矩相关系数 (Pearson product-moment correlation coefficient)，是一种线性相关系数。皮尔森相关系数是用来反映两个变量线性相关程度的统计量，用 r 表示。

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n}(Y_i - \bar{y})^2}}$$

其中 n 为样本量。r 描述的是两个变量间线性相关强弱的程度。r 的绝对值越大表明相关性越强。

0.6.7 K-L 散度

K-L 散度 (Kullback-Leibler Divergence)：即相对熵；是衡量两个分布 (P、Q) 之间的距离；越小越相似。

$$D(P \| Q) = \sum_{i=1}^{n} P(i) \log \frac{P(i)}{Q(i)}$$