2022.02.16

Lecture 7

Soft Clustering

对于每个点，算出它属于各个 cluster 的概率，而非将其 assign 到某组

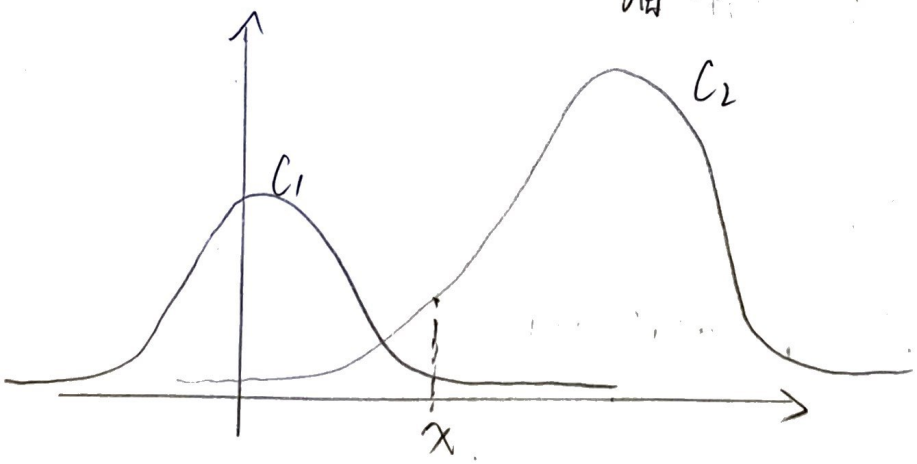
例：2种动物，根据其 weight 判断分类

Things to consider:

1. Prior probability of being one species
2. 每个 species 内部有其体重的 distribution

例：该地区 80% 猫，20% 狗　狗的体重 distribution 为 $N(\mu_1, \sigma_1)$

猫…　为 $N(\mu_2, \sigma_2)$



$X=x$ 时的 pdf：$P(X=x) = P(C_1) \cdot P(X=x|C_1) + P(C_2) \cdot P(X=x|C_2)$

$P(C_1 \text{ and } X=x) + P(C_2 \text{ and } X=x)$

Mixture Model

上述为 mixture model.

若有 k 个种类

$$P(X=x) = \sum_{j=1}^{k} P(C_j) \cdot P(X=x|C_j)$$

Gaussian Mixture Model

↓

即1个 Mixture Model中，所有 species 内部 distribution 都为 N

所需信息：① prior probability

② distribution 的 parameters

GMM Clustering

用 MLE 的想法算 GMM.

The probability of seeing the data we saw is the product of the probabilities of observing each data point.

将所有 data point 代入其 distribution，全部乘起来，用 MLE 的方法找 parameters.

$$\theta^* = \arg\max_{\theta} \prod_{i=1}^{n} \prod_{j=1}^{K} P(C_j) P(x_i | C_j)$$

↓

包含所有 C… 和 (μ, σ).

$$l(\theta) = \log(L(\theta))$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{K} \log(P(C_j) \cdot P(x_i | C_j))$$

求 partial derivative

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} P(C_j | x_i) x_i}{\sum_{i=1}^{n} P(C_j | x_i)}$$

$$\hat{\sigma}_j = \frac{\sum_{i=1}^{n} P(C_j | x_i)(x_i - \hat{\mu}_j)^T (x_i - \hat{\mu}_j)}{\sum_{i=1}^{n} P(C_j | x_i)}$$

$$\hat{P}(C_j) = \frac{1}{n} \sum_{i=1}^{n} P(C_j | x_i)$$

Clustering Aggregation

Compare clustering.

同组 data，两种 clustering，比较

↓

方法：对于每对 data point，看它们在 A clustering 和 B clustering 中是否一致。（x, y 在 A 中为一组，在 B 中是否为一组？）

Disagreement Distance

Given 2 clustering P, C

$$D(P,C) = \sum_{x,y} \Pi_{P,C}(x,y)$$

$$(x,y) = \begin{cases} 1 & \text{if } x \text{和} y \text{在} 1 \text{个 clustering 中在同组另} 1 \text{个 clustering 中不同组} \\ 0 & \text{o.w.} \end{cases}$$

Properties:

1. $D(C,P)=0$ 时 $C=P$
2. $D(C,P)=D(P,C)$
3. Triangle Inequality

Aggregate Clustering

从 $C_1,\cdots,C_m$ 中 generate a C^* that minimizes:

$$\sum_{i=1}^{m} D(C^*,C_i)$$

例：10个人，计录其所在地，职业，国籍.

↓ ↓ ↓

此3个得10个人3种不同方法 Cluster 走来

Benefits:

1. 可找到最佳 cluster 数
2. 不需分享 data，只需 assignments.