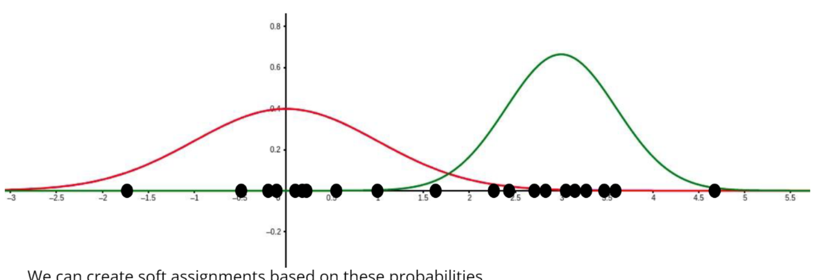
10.5 soft clustering

Tuesday, October 5, 2021 4:38 AM

- Hard clustering
  - 1 point -> 1 cluster
- Soft clustering
  - **Soft Clustering - Example**

    Generate data using $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$

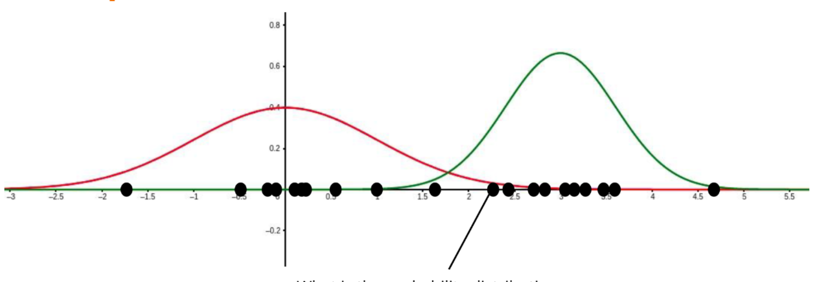
    

    We can create soft assignments based on these probabilities.
  -
- Mixture model
  - $$P(X = x) = \sum_{j=1}^{k} P(C_j) P(X = x \mid C_j)$$

    Mixture proportion Represents the probability of belonging to C_j

    Probability of seeing x when sampling from C_j

    Example

  - 

    $$P(X = x) = P(C_1)P(X = x|C_1) + P(C_2)P(X = x|C_2)$$

    $$P(X = x) = P(C_1)\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} + P(C_2)\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}$$
- Gaussian mixture model
  - $$P(X = x|C_i) \sim N(\mu, \sigma)$$
  - Maximum the overall probability
  - **GMM Clustering**

    Goal:

    $$\theta^* = \arg\max_{\theta} \prod_{i=1}^{n} \sum_{j=1}^{k} P(C_j) P(X_i \mid C_j)$$

    Where $\theta = \{\mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k, P(C_1), \ldots, P(C_k)\}$

    Joint probability distribution of our data

    Assuming our data are independent
  - Theta is everything we keep track of
  - ?????if we care about the order????
  - Apply log to both side of the equation don't change the curve of the function
  - $$l(\theta) = \log(L(\theta)) = \sum_{i=1}^{n} \log\Big(\sum_{j=1}^{k} P(C_j) P(X_i \mid C_j)\Big)$$
  - We can solve partial derivative (multi variance distribution)
  - $$\frac{d}{d\Sigma} l(\theta) = 0 \qquad \frac{d}{d\mu} l(\theta) = 0$$
  - Finally we get the equation:
  - $$\hat{\mu}_j = \frac{\sum_{i=1}^{n} P(C_j|X_i) X_i}{\sum_{i=1}^{n} P(C_j|X_i)}$$

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

    $$\hat{P}(C_j) = \frac{1}{n}\sum_{i=1}^{n} P(C_j|X_i)$$
  - Then we use bayes' rule to compute
  - $$P(C_j|X_i) = \frac{P(X_i|C_j)P(C_j)}{P(X_i)} = \frac{P(X_i|C_j)P(C_j)}{\sum_{j=1}^{k} P(C_j)P(X_i|C_j)}$$

    missing
  - So we implement expectation maximization alg
    - 1. Start with random $\boldsymbol{\theta}$
      2. Compute $\mathbf{P(C_j \mid X_i)}$ for all $\mathbf{X_i}$ by using $\boldsymbol{\theta}$
      3. Compute / Update $\boldsymbol{\theta}$ from $\mathbf{P(C_j \mid X_i)}$
      4. Repeat 2 & 3 until convergence

- Clustering aggregation
  - Goals:
    - Compare clusterings
    - Combine the information from multiple clustering to create a new clustering
  - Disagreement distance
    - Given 2 clusterings P and C

      $$D(P, C) = \sum_{x,y} \mathbb{I}_{P,C}(x, y)$$

      where

      $$\mathbb{I}_{P,C}(x, y) = \begin{cases} 1 & \text{if P \& C disagree on which clusters x \& y belong to} \\ 0 & \end{cases}$$
    - How many nodes have different assignments between p and c.