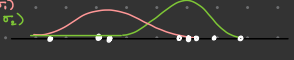
Oct 4th

Hard Clustering - 1pt → 1 cluster

Soft Clustering - helps w/ overlapping clusters, to give each pt a probability of being in a given cluster.



For each pt, we can compute probability of it being generated from any cluster.

MIXTURE MODEL

$$P(X = x) = \sum_{j=1}^{K} P(C_j)\, P(X = x \mid C_j)$$

- $P(C_j)$: MIXTURE PROPORTION
- $P(X = x \mid C_j)$: PROBABILITY OF SEEING x WHEN SAMPLING FROM C_j
- K clusters, so probability of $X=x$ is weighted by the prob of each cluster and $X=x$ given the cluster
- * ASSUMING discrete

$$\left\{ \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2} \right\}$$ NORMAL DISTRIBUTION μ_i σ_i

GMM

$$P(X = x \mid C_j) \sim N(\mu, \sigma)$$

GMM Clustering

GOAL: find the GMM that maximizes the probability of seeing the data we have.

Probability of seeing data we saw (assume iid) = the product of the probabilities of observing each data point.

Parameters we have: $P(C_i)$, μ_i, σ_i $\forall$ K clusters. → $\theta = \{\mu_1 \cdots \mu_K, \sigma_1 \cdots \sigma_K, P(C_1) \cdots P(C_K)\}$

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

(argument that maximizes θ)

Note: taking log transform. doesn't change critical pts

$$\ell(\theta^*) = \sum_{i=1}^{n} \log \sum_{j=1}^{K} P(C_j) P(X_i \mid C_j)$$

solve for

$$\frac{d}{d\mu} \ell(\theta) = 0 \qquad \frac{d}{d\Sigma} \ell(\theta) = 0$$

mle estimates: →

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

$$P(C_j \mid X_i) = \frac{P(X_i \mid C_j)\, P(C_j)}{\sum_{i=1}^{K} P(C_j)\, P(X_i \mid C_j)}$$ * BAYES RULE

* CIRCULAR PROBLEM → NEED estimators to estimate.

EXPECTATION-MAXIMIZATION ALGO

1. start w/ random θ
2. Compute $P(C_j \mid X_i)$ for all X_i using θ
3. Compute/update θ from $P(C_j \mid X_i)$
4. Repeat 2-3 until convergence.

- ESTIMATE WILL NOT GET WORSE.
- MUST CONVERGE
- NOT NECESSARILY THE OPTIMAL SOLN UPON CONVERGENCE

Clustering - a group of clusters output by a clustering algo

Cluster - a group of points.

Clustering Aggregation

GOALS: a. Compare clusterings b. Combine info from mult clusterings to create a new clustering.

To compare clusters, for each pair of points are they clustered to each other in each k clusters.

↳ DISAGREEMENT DISTANCE

$$D(P,C) = \sum_{x,y} \mathbb{I}_{P,C}(x,y) \quad \text{where} \quad \mathbb{I}_{P,C}(x,y) = \begin{cases} 1 \\ 0 \end{cases} \text{ if } P \& C \text{ disagree on which clusters } x,y \text{ belong to.}$$

distance function on 2 clusters P & C for all pairs of points. [IT'S A METRIC ♥]

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

generate C^* from set of clusterings and look to minimize

NOTE: PROBLEM IS EQUIVALENT TO CLUSTERING.

Benefits

- can identify best num of clusters
- can handle/detect outliers
- improve robustness of clustering algo
- privacy preserving clustering

BUT IT IS NP-HARD

MAJORITY RULE → ISSUE IS DOES NOT ALWAYS PRODUCE A CLUSTERING.