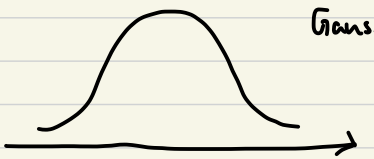
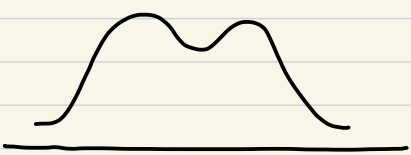
L12

- [x] Recap : PCA
- [ ] Clustering
  - [ ] Expectation Maximization
  - [ ] k-means

Probability distribution



Mixture model



$$p(x) = \sum_{k=1}^{k} p(X \mid C=k)\, p(C=k)$$

C=k 的情况下 X 的概率

Conditional prob of k-th distribution

probability of k-th distribution

Gaussian mixture model (GMM)

$$p(x) = \sum_{k=1}^{k} \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} p(C=k)$$

$$= \sum_{k=1}^{k} N(\mu_k, \sigma_k^2)\, p(C=k)$$

(we know k in advance)

Goal: Given samples from $P(X)$, estimate the parameters μ_k, σ_k

assume $P(C=k) = \frac{1}{2}$

Approach: Maximum likelihood

Input: I.I.D. samples $\{x_1, x_2, \dots, x_n\}$

Output: $\{(\mu_k, \sigma_k^2)\}_{k=1}^{K}$

Optimize: log-likelihood

$$p = \prod_{i=1}^{n} \sum_{k=1}^{k} N(x_i, \mu_k, \sigma_k^2) \frac{1}{k}$$

~ ignore

Log-likelihood $$L = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{k} N(x_i, \mu_k, \sigma_k^2) \right)$$

Simpler case: $k=1$

$$L = \sum_{i=1}^{n} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \right]$$

$$L = \sum_{i=1}^{n} -\frac{1}{2} \log(2\pi\sigma^2) - \left[\frac{x_i-\mu)^2}{2\sigma^2} \right]$$

Maximize w.r.t μ, σ

$$\frac{\partial L}{\partial \mu} = 0 \qquad \mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\frac{\partial L}{\partial \sigma^2} = 0 \qquad \sigma^2 = \frac{1}{n} \sum (x_i - \bar{\mu})^2$$

General Case :

$$\frac{\partial L}{\partial \mu_k} = 0 \qquad \text{apply } \frac{\partial}{\partial \mu_k}:$$

$$\sum_{i=1}^{n} \frac{1}{\sum N(x_i; \mu_k, \sigma_k^2)} \cdot N(x_i; \mu_k, \sigma_k^2) \cdot \left(\frac{x_i - \mu_k}{\sigma_k^2}\right) = 0$$

$\gamma_i(k)$: itself a probability distribution

Posterior distribution : $= P(C = k \mid X = x_i)$

$$\sum_{i=1}^{n} \gamma_i(k) \cdot \left(\frac{x_i - \mu_k}{\sigma_k^2}\right) = 0$$

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} \gamma_i(k) x_i}{\sum_{i=1}^{n} \gamma_i(k)}$$

$$\sigma_k^2 = \frac{\sum_{i=1}^{n} \gamma_i(k) (x_i - \hat{\mu}_k)^2}{\sum_{i=1}^{n} \gamma_i(k)}$$

if Oracles tells us posterior distribution ($\gamma_i(k)$)

then we can compute μ_k, σ_k^2

[Expectation E-step]

if oracle tells us means μ_k, variances σ_k^2,

then we can compute $\gamma(k)$

[Maximization, M-step]

Alternatively iterate → EM

Once μ_k, σ_k^2 are computed, then $\gamma_i(k)$ gives

$P(C=k \mid X=x_i)$

☆ K-means

↳ Hard clustering

EM soft-clustering

知道几类，回发label.

初始化K个点为随机的

input : $\{x_1, x_2, \ldots, x_n\}$

Output : $S_1 \cup S_2 \cup \ldots \cup S_k = X$ subsets

partition of data

$\{\mu_1, \mu_2, \ldots, \mu_k\}$ → "Cluster centers"

Loss function : "K-means objective function"

$$L = (\{S_1, S_2, \dots, S_k\}, \{\mu_1, \dots, \mu_k\})$$

$$= \sum_{k=1}^{K} \sum_{x_i \in S_k} \|x_i - \mu_k\|_2^2$$

Simpler Case : $k=1$

Optimize : $\sum_{x_i \in S} \|x_i - \mu\|_2^2$

w.r.t μ

Minimize Sample mean

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

General Case ?

Assume that cluster identities of each data points are known

optimize over $\sum_{x_i \in S_k} \|x_i - \mu_k\|_2^2$ ← O(nd)

$$\hat{\mu}_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} x_i$$

— Assume clusters centers are known. Assign each data points to the nearest μ_k.

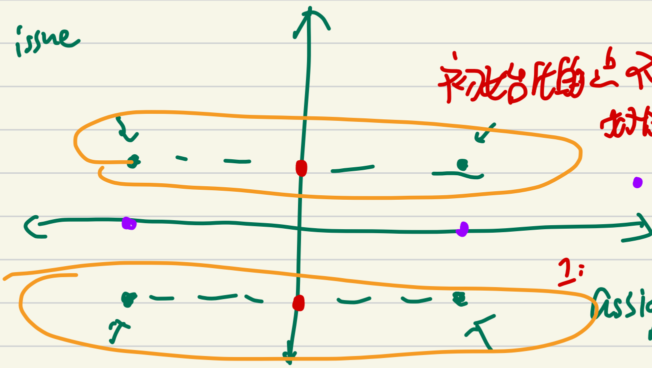
For each i, assign $\hat{c}_i = \arg\min_k \|x_i - \mu_k\|_2$ ← O(ndk)

— alternativity between these steps

→ k-Means

Worst case: k^n possible possible cluster choices

Other issue



初始化的点不好，得不到好的结果

• 才是好的结果

1: assign to •

2: compute mean of •

Caveats: 1) sensitive to initialization

2) exponential running time

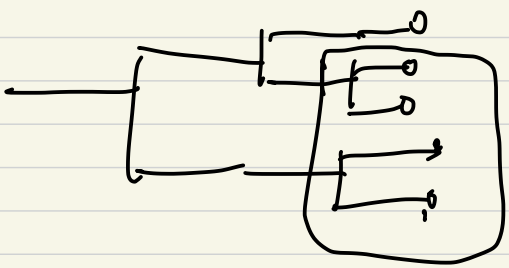
3) generally circular / clusters
spherize

k-Means ++

↳ smart initialization

Approaches requires no knowledge of k.

- Hierachical clustering
  - → Bioinformatics
  - → Genomics
  - → Linguistics



模型训练

- Bayesian methods