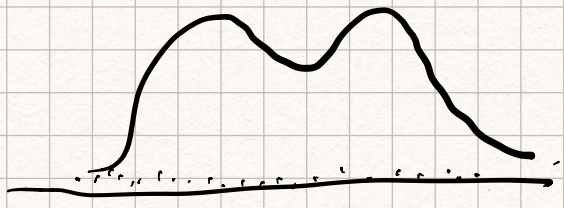
ECE-GY 6143 Intro to ML

- Recap: PCA
- Clustering
  - Expectation - Maximization
    - K-Means

Probability distribution:

Mixture model:



$$P(X) = \sum_{k=1}^{K} P(X \mid C=k)\, P(C=k)$$

Conditional prob. of k^{th} distribution

Probability of k^{th} distribution

Gaussian mixture model (GMM):

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

$$:= \sum_{k=1}^{K} N(\mu_k, \sigma_k^2) \; P(C=k).$$

Goal: Given samples from $P(x)$, estimate the parameters μ_k, σ_k. { Assume $P(C=k) = 1/K$ }

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.

(Likelihood) $$P = \prod_{i=1}^{n} \sum_{k=1}^{K} N(x_i; \mu_k, \sigma_k^2) \cdot \underbrace{\left(\frac{1}{K}\right)}_{\text{ignore}}$$

(Log-likelihood)

$$L = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} N(x_i; \mu_k, \sigma_k^2) \right).$$

Simpler case: $K=1$.

$$\frac{(x-\mu)^2}{}$$

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

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

Maximize w.r.t. μ, σ.

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

$$\frac{\partial L}{\partial \sigma^2} = 0 \qquad \boxed{\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2}$$

General case:

$$\frac{\partial L}{\partial \mu_k} = 0$$

$$\frac{\partial}{\partial \mu_k} e^{-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}}$$

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

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

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

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

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

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

- If oracle 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 $\gamma_i(k)$.

[Maximization, M-step]

Alternatively iterate → **EM**

Once μ_u, σ_u^2 are computed, then

$r_i(k)$ gives $P(C=k \mid X=x_i)$.

K-means

EM: soft-clustering

↳ Hard clustering

Input: $\{x_1, x_2, \dots, x_n\} = X$

Output: $S_1 \cup S_2 \cup \dots \cup S_K = X$

partition of X.

$\{\mu_1, \mu_2, \dots, \mu_K\}$ → "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 w.r.t μ $\quad \sum_{x_i \in S} \|x_i - \mu\|_2^2$

Minimizer: Sample mean

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

General case

- Assume that cluster identities of each data point are known.

Optimize over μ_k $\quad \sum_{x_i \in S_k} \|x_i - \mu_k\|^2$

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

$O(nd)$

- Assume cluster centers are known.

- Assign each data point to the nearest μ_k.

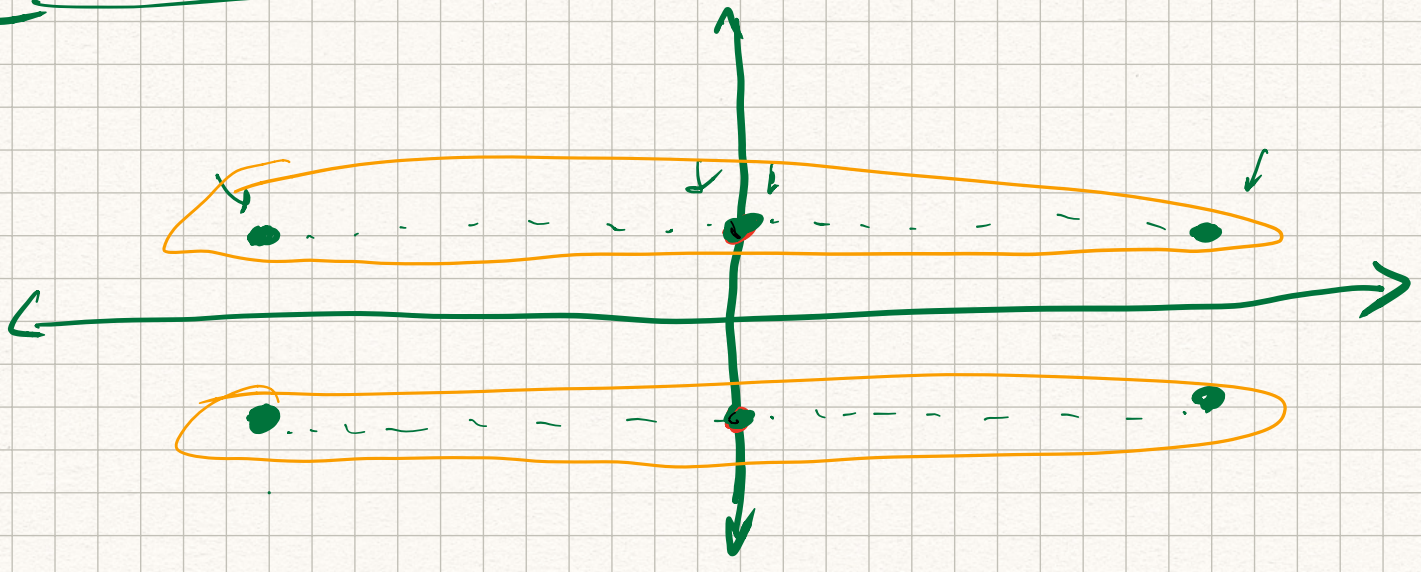
- For each i, assign $\hat{k}_i = \arg\min_k \|x_i - \mu_k\|_2$

- Alternating between these steps → **K-Means**

$O(ndk)$

Worst case: k^n possible cluster choices.

Other issue



Caveats: 1) Sensitive to initialization

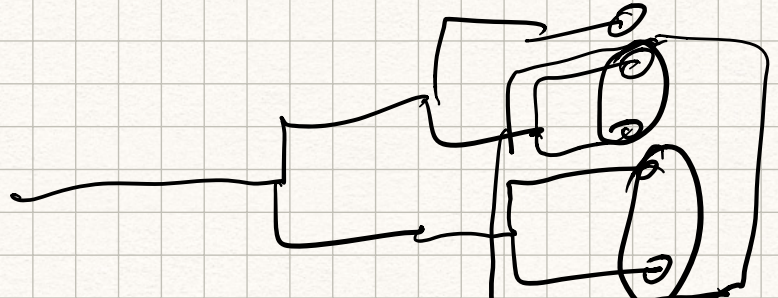
2) Exponential running time

3) Generally circular/ clusters.
Spherical.

K-Means ++
↳ Smart initialization

Approaches that do not require knowledge of k

- Hierarchical clustering.
  - ↳ Bioinformatics
  - ↳ Genomics
  - ↳ Linguistics.



- Bayesian methods