CS 480/680

3/19/26

Lecture 19

Q3 : 3/26/26 : 11:30

L VIII : PCA

- Mixtures

Lecture VII: Clustering: K-means and Mixtures of Gaussians

Marina Meilă
mmp@uwaterloo.ca

With Thanks to Pascal Poupart & Gautam Kamath
Cheriton School of Computer Science
University of Waterloo

March, 2026

Paradigms for clustering ✓

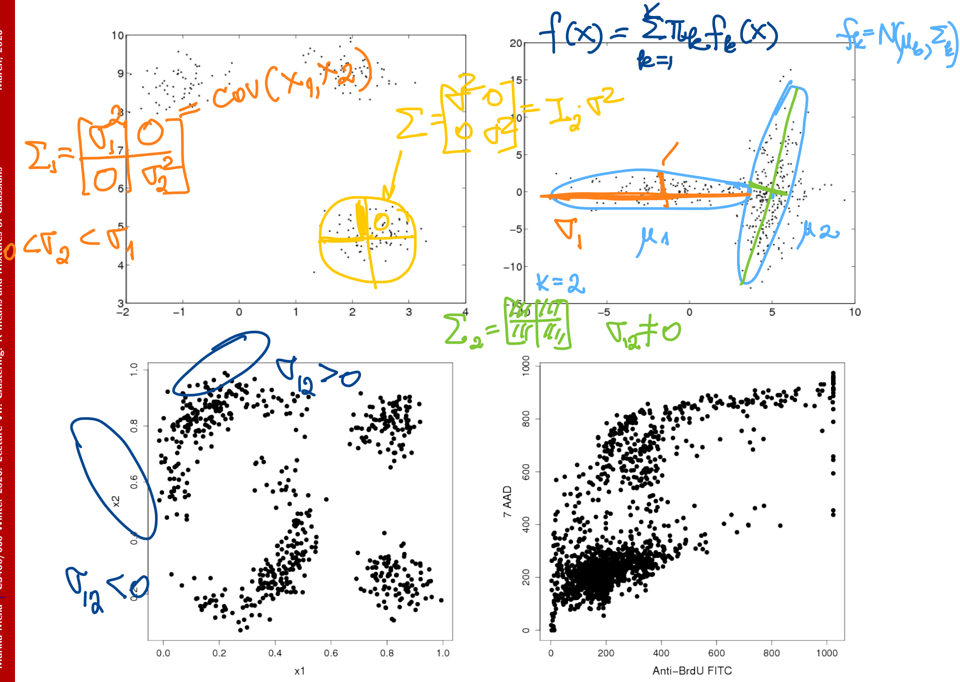
K-means clustering ✓

Mixtures of Gaussians and the EM algorithm ✓ ←

Initialization ←

Special topics in clustering

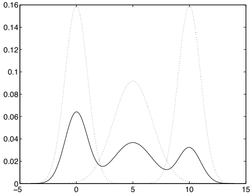
Reading HTF Ch.: 14.3, Murphy Ch.: Ch 11.[1], 11.2.1-3, 11.3, Ch 25, Bach Ch.:



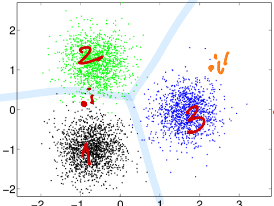
(from)

Model based clustering: Mixture models

Mixture in 1D



Mixture in 2D



- The **mixture density**

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$$

- $f_k(x)$ = the **components** of the mixture
  - each is a density
  - f called **mixture of Gaussians** if $f_k = Normal_{\mu_k, \Sigma_k}$
- π_k = the **mixing proportions**, $\sum_k = 1^K \pi_k = 1, \quad \pi_k \geq 0$.
- **model parameters** $\theta = (\pi_{1:K}, \mu_{1:K}, \Sigma_{1:K})$
- The **degree of membership** of point i to cluster k

$$\gamma_{ki} \stackrel{\text{def}}{=} P[x_i \in C_k] = \frac{\pi_k f_k(x)}{f(x)} \text{ for } i = 1:n, k = 1:K \quad (8)$$

- depends on x_i and on the model parameters

$\gamma_{1i} = \gamma_{2i} = 0.44 \quad \gamma_{3i} = .12$

$\gamma_{i'3} = 0.8 \quad \gamma_{i'2} = .12 \quad \gamma_{i'1} = .08$

Guess which Gaussian?

Parameters $(\pi_{1:K}, \mu_{1:K}, \Sigma_{1:K}) = \theta$ ⇙ known

Bayes' rule

given x, $k = ?$

$$\Pr[k \mid x^i] = \frac{p(x^i \mid k) \cdot P[k]}{f(x^i)} = \gamma_{ki} \geq 0 \qquad k = 1:K$$

($p(x^i \mid k) = f_k(x^i)$, $P[k] = \pi_k$)

Ex → $\sum_{k=1}^{K} \gamma_{ki} = 1$

$x \in \mathbb{R}^D$

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & \\ \sigma_{12} & \sigma_2^2 & \\ \sigma_{13} & & \ddots \\ \vdots & & & \sigma_D^2 \\ \sigma_{1D} & & \end{bmatrix}$$

$\sigma_{ij} = Cov(x_i, x_j)$

coordinates i, j

Criterion for clustering: Max likelihood

- denote $\theta = (\pi_{1:K}, \mu_{1:K}, \Sigma_{1:K})$ (the parameters of the mixture model)
- Define **likelihood** $P[\mathcal{D}|\theta] = \prod_{i=1}^{n} f(x_i)$
- Typically, we use the **log likelihood** ← has local maxima :(

$$l(\theta) = \ln \prod_{i=1}^{n} f(x_i) = \sum_{i=1}^{n} \ln \sum_{k} \pi_k f_k(x_i) \quad (9)$$

f(x^i)

- denote $\theta^{ML} = \underset{\theta}{\operatorname{argmax}}\, l(\theta)$
- θ^{ML} determines a soft clustering γ by (8)
- a soft clustering γ determines a θ (see later)
- Therefore we can write

Loss $\mathcal{L}(\gamma) = -l(\theta(\gamma))$

Maximize l → GAscent
Maximize l → EM

Algorithms for model-based clustering

Maximize the (log-)likelihood w.r.t θ

- directly - (e.g by gradient ascent in θ)
- by the EM algorithm (very popular!)
- indirectly, w.h.p. by "computer science" algorithms

w.h.p = with high probability (over data sets)

The Expectation-Maximization (EM) Algorithm

Algorithm Expectation-Maximization (EM)

Input Data $\mathcal{D} = \{x_i\}_{i=1:n}$, number clusters K

tialize parameters $\pi_{1:K} \in \mathbb{R}$, $\mu_{1:K} \in \mathbb{R}^d$, $\Sigma_{1:K} \in \mathbb{R}^{d\times d}$ at random[1]

terate until convergence

[soft] "assigns" x^i to $C_{1:K}$

E step (Optimize clustering) for $i = 1:n,\ k = 1:K$

$$\gamma_{ki} = \frac{\pi_k f_k(x)}{f(x)}$$

M step (Optimize parameters) set $\Gamma_k = \sum_{i=1}^n \gamma_{ki}$, $k = 1:K$ (number of points in cluster k)

"n_k" =

Recalculate cluster parameters

$$\pi_k = \frac{\Gamma_k}{n}, \quad k = 1:K$$

$$\mu_k = \sum_{i=1}^n \frac{\gamma_{ki}}{\Gamma_k} x_i$$

$$\Sigma_k = \frac{\sum_{i=1}^n \gamma_{ki}(x_i - \mu_k)(x_i - \mu_k)^T}{\Gamma_k}$$

- $\pi_{1:K}, \mu_{1:K}, \Sigma_{1:K}$ are the maximizers of $l_c(\theta)$ in (13)
- $\sum_k \Gamma_k = n$

[1] Σ_k need to be symmetric, positive definite matrices

The EM Algorithm – Motivation

- Define the **indicator variables**

$$z_{ik} = \begin{cases} 1 & \text{if } i \in C_k \\ 0 & \text{if } i \notin C_k \end{cases} \quad (10)$$

  denote $\bar{z} = \{z_{ki}\}_{k=1:K}^{i=1:n}$
- Define the **complete log-likelihood**

$$l_c(\theta, \bar{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ki} \ln \pi_k f_k(x_i) \quad (11)$$

- $E[z_{ki}] = \gamma_{ki}$
- Then

$$E[l_c(\theta, \bar{z})] = \sum_{i=1}^{n} \sum_{k=1}^{K} E[z_{ki}][\ln \pi_k + \ln f_k(x_i)] \quad (12)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ki} \ln \pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ki} \ln f_k(x_i)] \quad (13)$$

- If θ known, γ_{ki} can be obtained by (8) **(Expectation)**
- If γ_{ki} known, π_k, μ_k, Σ_k can be obtained by separately maximizing the terms of $E[l_c]$ **(Maximization)**

Brief analysis of EM

$$Q(\theta,\gamma) \;=\; \sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{ki}\ln \underbrace{\pi_k f_k(x_i)}_{\theta}$$

- each step of EM increases $Q(\theta,\gamma)$
- Q converges to a local maximum
- at every local maxi of Q, $\theta \leftrightarrow \gamma$ are fixed point
- $Q(\theta^*,\gamma^*)$ local max for $Q \Rightarrow\; l(\theta^*)$ local max for $l(\theta)$
- under certain regularity conditions $\theta \longrightarrow \theta^{ML}$
- the E and M steps can be seen as projections

- Exact maximization in **M step** is not essential.
  Sufficient to increase Q.
  This is called **Generalized EM**

EM w.r.t Q

$\theta^0 \xrightarrow{E} \gamma^0 \xrightarrow{M} \theta^1 \rightarrow \gamma^1 \rightarrow \theta^2 \dots$

converge? YES

At convergence

E step, M step: $\gamma^*(\theta^*) = \theta^*(\gamma^*) \Leftarrow$ max of $Q(\gamma,\theta)$ $\Downarrow$ local max of $l(\theta)$

fixed point

Probablistic alternate projection view of EM

- let $z_i =$ which gaussian generated i? (random variable), $X = (x_{1:n})$, $Z = (z_{1:n})$
- Redefine Q

$$Q(\tilde{P}, \theta) \; = \ell(\theta) - KL(\tilde{P}||P(Z|X,\theta)$$ ⇐ Thm

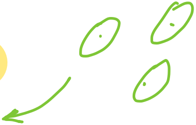
  where $P(X, Z|\theta) = \prod_i \prod_k P[z_i = k]P[x_i|\theta_k]$
  $\tilde{P}(Z)$ is any distribution over Z,
  $KL(P(w)||Q(w)) = \sum_w P(w) \ln \frac{P(w)}{Q(w)}$ the **Kullbach-Leibler divergence**

  Then,
  - **E step** $\max_{\tilde{P}} Q \Leftrightarrow \; KL(\tilde{P}||P(Z|X,\theta)$
  - **M step** $\max_\theta Q \Leftrightarrow \; KL(P(X|Z,\theta^{old})||P(X|\theta))$
- Interpretation: KL is “distance”, “shortest distance” = projection

The M step in special cases

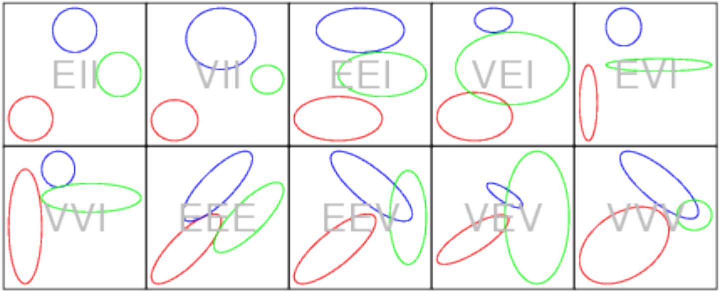
- Note that the expressions for μ_k, Σ_k = expressions for μ, Σ in the normal distribution, with data points x_i weighted by $\frac{\gamma_{ki}}{\Gamma_k}$

	M step
general case	$\Sigma_k = \sum_{i=1}^n \frac{\gamma_{ki}}{\Gamma_k}(x_i - \mu_k)(x_i - \mu_k)^T$
$\Sigma_k = \Sigma$ "same shape & size" clusters	$\Sigma \leftarrow \frac{\sum_{i=1}^n \sum_{k=1}^K \gamma_{ki}(x_i - \mu_k)(x_i - \mu_k)^T}{n}$
$\Sigma_k = \sigma_k^2 I_d$ "round" clusters	$\sigma_k^2 \leftarrow \frac{\sum_{i=1}^n \gamma_{ki} \|\|x_i - \mu_k\|\|^2}{d\Gamma_k}$
$\Sigma_k = \sigma^2 I_d$ "round, same size" clusters	$\sigma^2 \leftarrow \frac{\sum_{i=1}^n \sum_{k=1}^K \gamma_{ki} \|\|x_i - \mu_k\|\|^2}{nd}$



Exercise Prove the formulas above

- Note also that **K-means** is **EM** with $\Sigma_k = \sigma^2 I_d, \; \sigma^2 \to 0$ Exercise Prove it



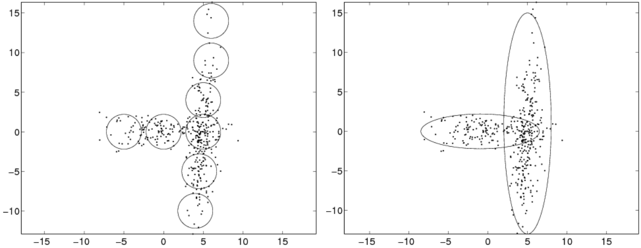
More special cases introduce the following description for a covariance matrice in terms of *volume, shape, alignment with axes* (=determinant, trace, e-vectors). The letters below mean: I=unitary (shape, axes), E=equal (for all k), V=unequal

- EII: equal volume, round shape (spherical covariance)
- VII: varying volume, round shape (spherical covariance)
- EEI: equal volume, equal shape, axis parallel orientation (diagonal covariance)
- VEI: varying volume, equal shape, axis parallel orientation (diagonal covariance)
- EVI: equal volume, varying shape, axis parallel orientation (diagonal covariance)
- VVI: varying volume, varying shape, equal orientation (diagonal covariance)
- EEE: equal volume, equal shape, equal orientation (ellipsoidal covariance)
- EEV: equal volume, equal shape, varying orientation (ellipsoidal covariance)
- VEV: varying volume, equal shape, varying orientation (ellipsoidal covariance)
- VVV: varying volume, varying shape, varying orientation (ellipsoidal covariance)

(from)

EM versus K-means

- Alternates between cluster assignments and parameter estimation
- Cluster assignments γ_{ki} are probabilistic
- Cluster parametrization more flexible



- Converges to local optimum of **log-likelihood**
  Initialization recommended by K-logK method

- **Modern algorithms with guarantees** (for e.g. mixtures of Gaussians)
  - Random projections
  - Projection on principal subspace
  - Two step EM (=K-logK initialization + one more EM iteration)

A fundamental result

The Johnson-Lindenstrauss Lemma For any $\varepsilon \in (0, 1]$ and any integer n, let d' be a positive integer such that $d' \geq 4(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \ln n$. Then for any set $\mathcal{D}$ of n points in $\mathbb{R}^d$, there is a map $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ such that for all $u, v \in V$,

$$(1-\varepsilon)||u-v||^2 \leq ||f(u) - f(v)||^2 \leq (1+\varepsilon)||u-v||^2 \tag{14}$$

Furthermore, this map can be found in randomized polynomial time.

- note that the **embedding dimension** d' does not depend on the original dimension d, but depends on n, ε
- show that: the mapping f is linear and that w.p. $1 - \frac{1}{n}$ a random projection (rescaled) has this property
- their proof is elementary Projecting a fixed vector v on a a random subspace is the same as projecting a random vector v on a fixed subspace. Assume $v = [v_1, \ldots v_d]$ with $v \sim$ i.i.d. and let $\tilde{v}$ = projection of v on axes $1 : d'$. Then $E[||\tilde{v}||^2 = d' E[v_j^2] = \frac{d'}{d} E[||v||^2]$. The next step is to show that the variance of $||\tilde{v}||^2$ is very small when d' is sufficiently large.

A two-step EM algorithm

Assumes K spherical gaussians, separation $||\mu_k^{true} - \mu_{k'}^{true} \geq C\sqrt{d}\sigma_k$

1. Pick $K' = \mathcal{O}(K \ln K)$ centers μ_k^0 at random from the data
2. Set $\sigma_k^0 = \frac{d}{2} \min_{k \neq k'} ||\mu_k^0 - \mu_{k'}^0||^2$, $\pi_k^0 = 1/K'$
3. Run one E step and one M step $\Longrightarrow \{\pi_k^1, \mu_k^1, \sigma_k^1\}_{k=1:K'}$
4. Compute "distances" $d(\mu_k^1, \mu_{k'}^1) = \frac{||\mu_k^1 - \mu_{k'}^1||}{\sigma_k^1 - \sigma_{k'}^1}$
5. Prune all clusters with $\pi_k^1 \leq 1/4K'$
6. Run Fastest First Traversal with distances $d(\mu_k^1, \mu_{k'}^1)$ to select K of the remaining centers. Set $\pi_k^1 = 1/K$.
7. Run one E step and one M step $\Longrightarrow \{\pi_k^2, \mu_k^2, \sigma_k^2\}_{k=1:K}$

eorem For any $\delta, \varepsilon > 0$ if d large, n large enough, separation $C \geq d^{1/4}$ the **Two step EM** algorithm obtains centers μ_k so that

$$||\mu_k - \mu_k^{true}|| \leq ||\mathrm{mean}(C_k^{true}) - \mu_k^{true}|| + \varepsilon\sigma_k\sqrt{d}$$

In practice

$\pi_k^0 = 1/K$

$\Sigma_k^0 = \sigma_0^2 I$ ← σ_0 not too small

μ_k^0 ← K-log K

Clustering for large D

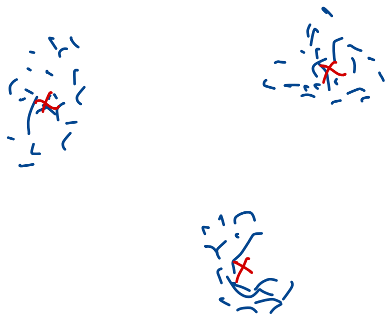
- Easier
- by projecting to lower $D' < D$

1) Random projection
   - choose $D' \sim K \ln K$
   - project $x^{1:n}$ on random V $\quad V \in \mathbb{R}^{D \times D'}$ orthogonal

   ↑ D' subspace basis

   $$V V^T x = \mathrm{Proj}_V\, x$$

2. PCA $(K-1)$



Selecting K for mixture models

The **BIC (Bayesian Information) Criterion**

HEURISTIC

- let θ_K = parameters for γ_K
- let $\#\theta_K$=number independent parameters in θ_K
  - e.g for mixture of Gaussians with full Σ_k's in d dimensions

$$\#\theta_K = \underbrace{K-1}_{\pi_{1:K}} + \underbrace{Kd}_{\mu_{1:K}} + \underbrace{Kd(d-1)/2}_{\Sigma_{1:K}}.$$

- define

$$BIC(\theta_K) = l(\theta_K) - \frac{\#\theta_K}{2}\ln n$$

↗ with K

- **Select** K that maximizes $BIC(\theta_K)$
- selects true K for $n \to \infty$ and other technical conditions (e.g parameters in compact set)
- but theoretically not justified (and overpenalizing) for finite n

1. Run clustering algo for K=2,3,... Kmax
⇒ Δ2, Δ3, ... ΔKmax
Select K* based on ΔK
K=2:Kmax

#θ = nr params ↗ with K