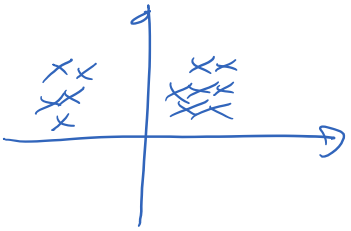
Lecture 10 - K-means

October 5, 2021 4:31 PM

today : • unsupervised learning
• K-means & EM

Unsupervised learning

here X without any labels Y



consider the Gaussian mixture model (GMM)
(can be obtained from FLD)

$Y \sim \text{Mult}(\pi) \quad \pi \in \Delta_K$ [extension of FLD to multiple classes]

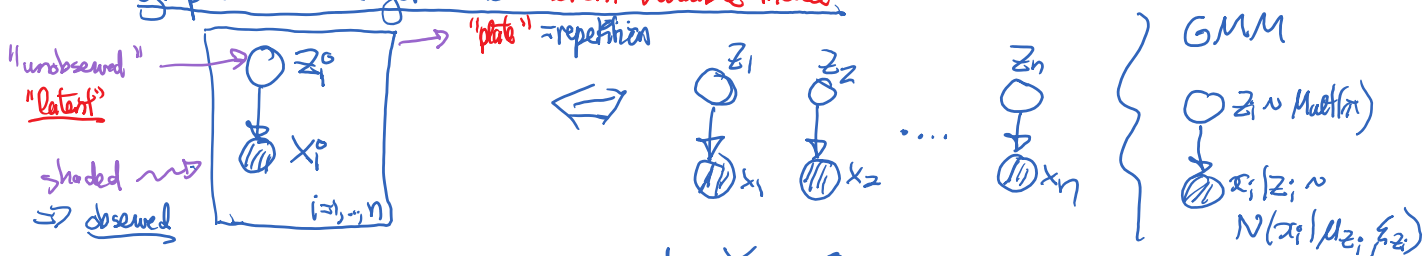
$X | Y=j \sim N(\mu_j, \Sigma)$

$$p(x) = \sum_y p(x,y) = \sum_y p(x|y)p(y) = \sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma)$$

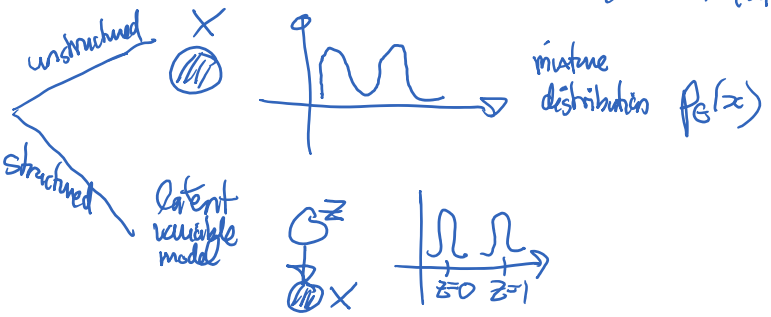
"GMM model"

more generally can have Σ_j per class

graphical model for this "latent variable model"

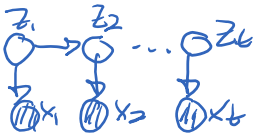


two views on $p(x)$

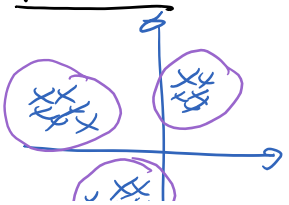


$$p(x) = \sum_z p_\theta(x|z)\, p_\theta(z)$$

(later in class, we will add time structure : HMM)

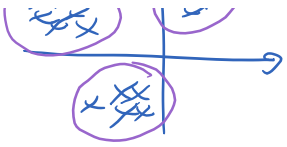


K-means → to do clustering ie. group data



we want to get a cluster assignment for every data point x_i

represent $z_{i,j}=1$ to mean that x_i belongs to cluster j



represent $z_{i,j}=1$ to mean that x_i belongs to cluster j

$j=1,\dots,k$ ← # of clusters (specified in advance for k-means)

Applications: • vector quantization (compression)

• in computer vision: use k-mean to get "bag of visual words" representation of images patches

• many many others!?

k-mean alg.: → can be derived as a block-coordinate minimization alg. of objective fct.:

"distorsion measure" → $J(z,\mu) \triangleq \sum_{i=1}^{n} \|x_i - \mu_{z_i}\|_2^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{k} z_{i,j} \|x_i - \mu_j\|^2 \right)$

cluster index represented by z_i

Cluster assignment $z_1,\dots,z_n \in$ corners of Δ_k ("one hot encoding")

$\mu_1,\dots,\mu_k \in \mathbb{R}^d$ Cluster centroids

$\underbrace{x_1, x_2}_{\text{block 1}}, \underbrace{x_3, x_4}_{\text{block 2}}$

alg.: 1) initialize $\mu^{(1)}$

2) iterate until convergence:

"E" step: $z^{(t+1)} = \underset{z \in \text{valid ass}}{\operatorname{argmin}} J(z, \mu^{(t)})$

$\Rightarrow z_{i,j^*}^{(t+1)} = 1$ for $j^* = \underset{j}{\operatorname{argmin}} \|x_i - \mu_j^{(t)}\|$

"M" step: $\mu^{(t+1)} = \underset{\mu \in \mathbb{R}^{d \times k}}{\operatorname{argmin}} J(z^{(t+1)}, \mu)$

$\Rightarrow \boxed{\mu_j^{(t+1)} = \frac{\sum_i z_{i,j} x_i}{\left(\sum_i z_{i,j}\right)}}$ empirical mean of cluster

Visualization:
https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

17h32

Properties of k-means:

1) converge in finite # of iterations to a local min

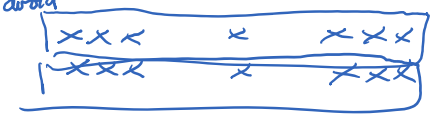
2) NP hard in general to compute the global minimum in z

K-means++: clever initialization scheme which guarantees that alg. is within $\log k$ of global opt. (w.h.p.)

min ... to avoid

guarantees that alg. is within $\log k$ of global opt. (w.h.p.)

→ idea: spread out as much as possible the initial means to avoid

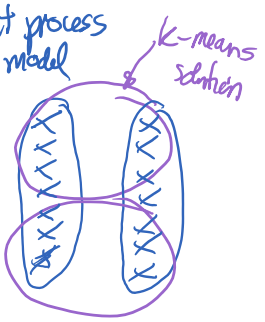


3) choice of k?

- one heuristic is: $J(\mu, z, k) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\|x_i - \mu_j\|^2 + \lambda k$ (λ: hyperparameter)

→ we'll see later in class "non-parametric" models where "k" is basically infinite and can get $p(k \mid \text{data})$ e.g. Dirichlet process mixture model

4) K-mean is very sensitive on distance measure: it assumes spherical clusters



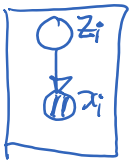
↳ GMM fixes that problem

Mahalanobis distance $d_\Sigma(x,x') \triangleq \sqrt{(x-x')^T \Sigma^{-1}(x-x')}$

Side reference I mentioned:
see https://icml.cc/2012/papers/291.pdf
for interpreting regularized K-means as approximate inference in a Dirichlet process mixture model... [by Kullis & Jordan]

EM ~ maximum likelihood in latent variable model

setup



z latent variable

x observed variable

log-likelihood $\log p(x_{1:n}; \theta) = \log\left(\prod_{i=1}^{n} p(x_i; \theta)\right)$

$= \sum_{i=1}^{n} \log p(x_i; \theta)$

$= \sum_{i=1}^{n} \log \left[\sum_{z_i} p(x_i, z_i; \theta)\right]$

problem → gives multi-modal opt. problem (non-convex)

options for MLE in latent variable model

f is convex $\Leftrightarrow$ $-f$ is concave

1) do gradient ascent on a non-concave obj.

2) EM alg. → block-coordinate ascent on auxiliary fct. which lower bounds $\log p(x_{1:n}; \theta)$

nice interpretation in terms of filling "missing data")

i.e. E step → fill z with "soft-values"

M step → max w.r.to θ for fully observed model

trick overview:

Jensen's inequality

trick overview:

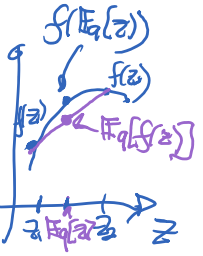
$$\log \sum_z p(x,z) = \log \sum_z q(z) \frac{p(x,z)}{q(z)}$$

$$= \log \mathbb{E}_q\left[\frac{p(x,z)}{q(z)}\right]$$

Jensen's inequality

$$\mathbb{E}_q[f(g(z))] \leq f(\mathbb{E}_q(g(z)))$$

when f is concave



$$\geq \mathbb{E}_q\left[\log \frac{p(x,z)}{q(z)}\right] = \sum_z q(z) \log p_\theta(x,z) - \sum_z q(z) \log q(z)$$

Jensen's inequality trick!

$$\triangleq \mathcal{L}(q,\theta) \triangleq \underbrace{\mathbb{E}_q[\log p(x,z;\theta)]}_{\text{"expected complete log-likelihood"}} + \underbrace{H(q)}_{\text{"entropy of } q\text{"}}$$

we have $\log p(x;\theta) \geq \mathcal{L}(q,\theta) \quad \forall q, \theta$

EM algorithm: E step: $q_{t+1} \triangleq \underset{q \text{ dist over } z}{\operatorname{argmax}} \mathcal{L}(q,\theta_t) \Rightarrow q_{t+1}(z) = p(z|x;\theta_t)$

M step: $\theta_{t+1} \triangleq \underset{\theta}{\operatorname{argmax}} \mathcal{L}(q_{t+1},\theta)$

$= \underset{\theta}{\operatorname{argmax}} \mathbb{E}_{q_{t+1}(z)}[\log p(x,z;\theta)]$

this is another MLE problem, but for complete information

(often, replace z with $\mathbb{E}_q[z]$ in this expression)

* we had $\log p(x;\theta) \geq \mathcal{L}(q,\theta)$ from Jensen's inequality

$\log(\mathbb{E}_q[g(z)]) \geq \mathbb{E}_q[\log(g(z))]$

in Jensen's ineq., you get a strict ineq unless the dist. for $g(z)$ is degenerate (i.e. takes only one value)

(when f is strictly concave)

[log is " "]

i.e. when $g(z) =$ constant

$g(z) =$ constant i.e. $\frac{p(x,z)}{q(z)} = \text{const.} \ \forall z \Rightarrow q(z) \propto p(x,z)$

i.e. $\boxed{q(z) = p(z|x,\theta)}$

makes equality $\mathcal{L}(q,\theta) = \log p(x;\theta)$

$\mathcal{L}(q_{t+1},\theta_t) = \log p(x;\theta_t) \geq \mathcal{L}(q,\theta_t) \ \forall q$

$\Rightarrow q_{t+1}$ maximizes $\mathcal{L}(q,\theta_t)$ w.r. to q

i.e. $\underset{q}{\operatorname{argmax}} \mathcal{L}(q,\theta_t) = p(z|x;\theta_t) = q_{t+1}$

and $\mathcal{L}(q_{t+1},\theta_t) = \log p(x;\theta_t)$