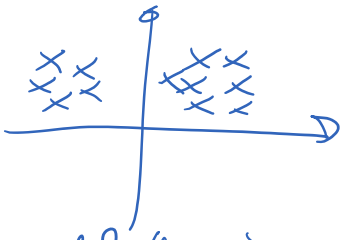
Lecture 11 - K-means & EM

Tuesday, October 6, 2020 14:30

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)

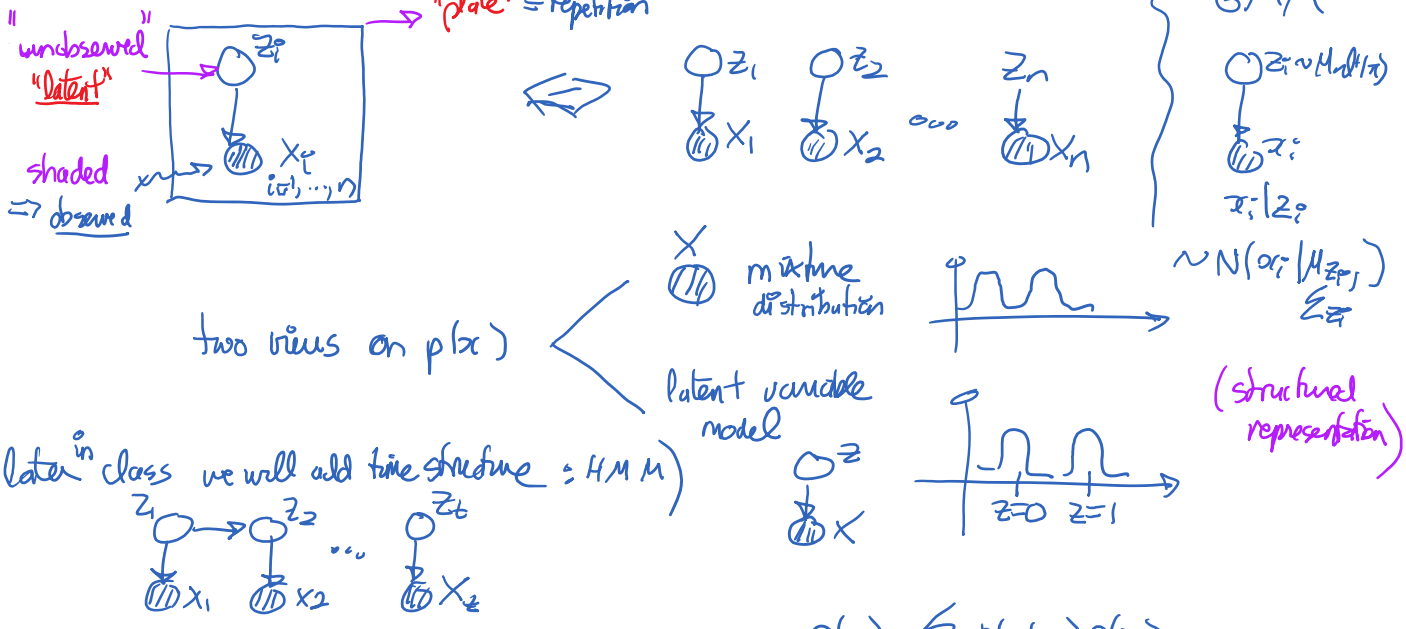
$Y \sim Mult(\pi) \quad \pi \in \Delta_K$ [extension of FLD to multiple classes]

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

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

"GMM model" more generally can have Σ_j per class

graphical model for this "latent variable model"

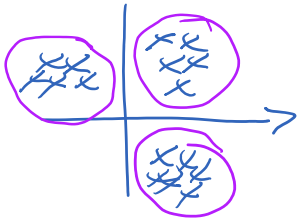


$$p(x) = \sum_z p(x|z)p(z)$$

$p(x|z)$ $p(z)$

$p(\theta \mid \text{data})$

K-means → to do clustering i.e. 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

$j = 1, \dots, K$ ← # 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 image patches
- many others?

15h36

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

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

- z: cluster assignment $z_1, \dots, z_n \in$ corners of Δ_K ("one-hot encoding")
- μ: $\mu_1, \dots, \mu_K \in \mathbb{R}^d$ cluster centroids
- μ_{z_i}: cluster index represented by z_i

alg.:

1) initialize $\mu^{(1)}$

2) iterate until convergence:

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

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

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

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

Visualization:
https://stanford.edu/class/engr108/visualizations/kmeans/kmeans.html

properties of K-means:

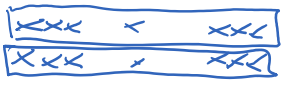
1) converge in finite # of iterations to a local min

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

2) NP hard in general to compute the global minimum in $\mathbb{Z}$

K-means++ : clever initialization scheme which 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 : $J(\mu, z, K) = \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} \|x_i - \mu_j\|^2 + K$

K : hyperparameter

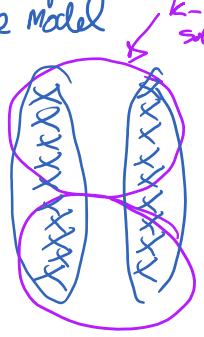
→ we'll see later in class "non-parametric" models where "K" is basically infinite and can get $p(K|data)$

e.g. Dirichlet process mixture model

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, I was wrong on the author! ;)]

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

↳ GMM fixes that problem →



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