Lecture 10 - scribbles

Tuesday, October 8, 2019 14:35

today: * finish MLE for Gaussian
* K-means, GMM, EM

continuation: MLE for multivariate Gaussian

example 2: derivative of $f(A) \triangleq \log\det(A)$ where $\begin{pmatrix} A \text{ symmetric} \\ A \succ 0 \end{pmatrix}$

can represent derivative of a fct. from matrix to scalar, as a matrix

$$f(A+\Delta) - f(A) = \mathrm{tr}(f'(A)^T \Delta) + o(\|\Delta\|)$$
$$= \langle f'(A), \Delta \rangle + \quad //$$

$\log\det(A+\Delta) - \log\det(A)$ — $A \succ 0 \Rightarrow$ invertible & has unique positive square root

$$= \log\det\left(A^{1/2}\left(I + A^{-1/2}\Delta A^{-1/2}\right)A^{1/2}\right) - \log\det(A)$$
$$= \log|A|^{1/2}\,|I + A^{-1/2}\Delta A^{-1/2}|\,|A|^{1/2} - \log|A|$$
$$= \log\det|I + A^{-1/2}\Delta A^{-1/2}|$$

use $\det(A) = \prod_i \lambda_i(A)$ (e-values)

$$= \sum_i \log \lambda_i\left(I + A^{-1/2}\Delta A^{-1/2}\right)$$
$$= \sum_i \log\left(1 + \lambda_i(A^{-1/2}\Delta A^{-1/2})\right)$$

$\log(1+x) = x + O(x^2)$ for $|x|<1$

$$= \sum_i \lambda_i(A^{-1/2}\Delta A^{-1/2}) + O(\lambda_i(\cdot)^2)$$

$o(\|\Delta\|)$ — $\mathrm{tr}(A) = \sum_i \lambda_i(A)$

$$= \mathrm{tr}(A^{-1/2}\Delta A^{-1/2}) + o(\|\Delta\|)$$
$$= \mathrm{tr}(A^{-1}\Delta) + o(\|\Delta\|)$$

(recall A is symmetric) $\mathrm{tr}(A^{-1}\Delta) = \langle A^{-1}, \Delta \rangle$

$$\Rightarrow \boxed{\frac{\partial}{\partial A}\log\det(A) = A^{-1}} \quad //$$

back to log-likelihood for Gaussian:

concave as function of $\Lambda = \Sigma^{-1}$?

$$+\frac{n}{2}\log|\Sigma^{-1}| - \frac{n}{2}\langle \Sigma^{-1}, \tilde{\Sigma}(\mu) \rangle$$

derivative

want

$\hat{\Sigma}$

derivative w.r.t. Σ^{-1} (where $\Sigma = \Lambda$)

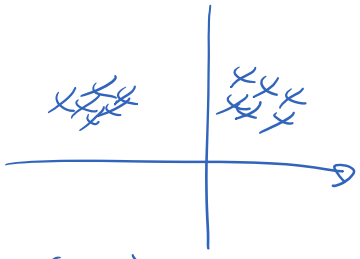
$$\frac{n}{2}\underbrace{(\Sigma^{-1})^{-1}}_{\Sigma} - \frac{n}{2}\hat{\Sigma}(\mu) = 0 \quad \text{want}$$

$$\Rightarrow \hat{\Sigma}_{MLE} = \hat{\Sigma}(\mu_{MLE}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{MLE})(x_i - \mu_{MLE})^T$$

15h

Unsupervised learning

here X without any labels Y



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

$$Y \sim \text{Mult}(\pi) \quad \pi \in \Delta_K$$

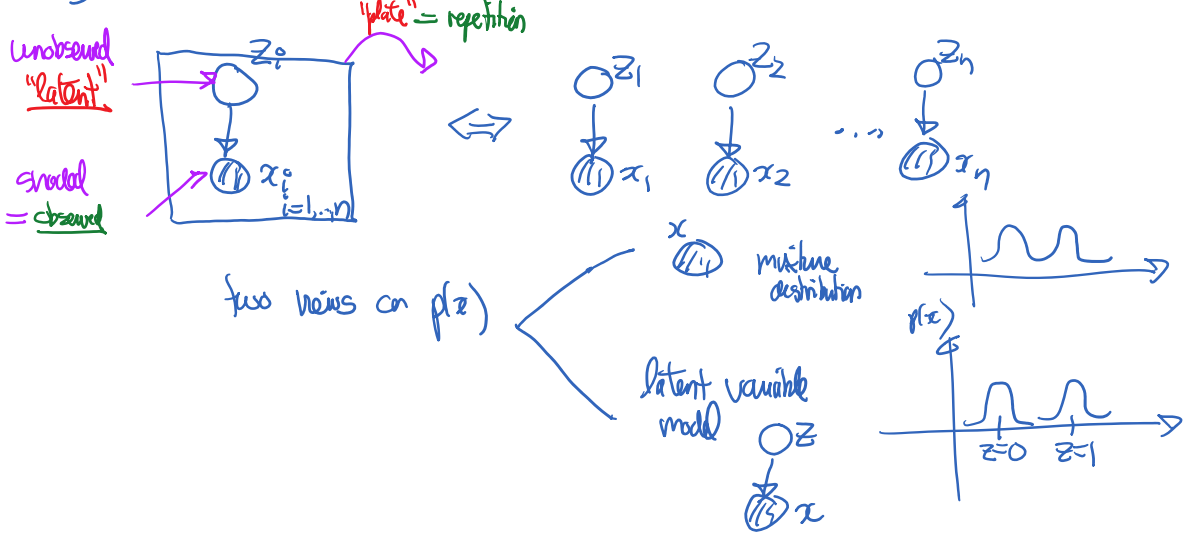
$$X \mid Y = j \sim N(\mu_j, \Sigma)$$

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

"GMM model"

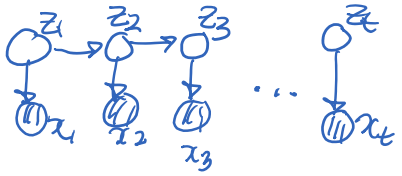
more generally, can have different Σ_j per class

graphical model for this "latent variable model"



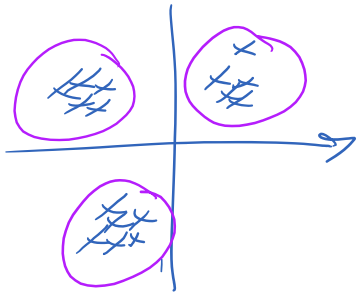
(later in class, we will add time structure: HMM)

(xaxis in cross, we can add time structure: HMM)



K-means → to do clustering ie. group data

(can be seen as a limit of GMM $\sigma^2 \to 0$)



we want to get a cluster assignment for every data pt. x_i

represent $z_{ij}=1$ to mean that x_i in cluster j

$j=1,\dots,K$

# of clusters (specified in advance for K-means)

applications: • vector quantization

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

[sh31]

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

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

Z: cluster assign. $z_1,\dots,z_n \in$ corners of Δ_K

$\mu_1,\dots,\mu_K \in \mathbb{R}^d$: cluster means

"distortion measure"

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

2) iterate until convergence

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

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

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

$\Rightarrow \boxed{\mu_j^{(t+1)} = \frac{1}{\sum_i z_{ij}^{(t+1)}} \sum_i z_{ij}^{(t+1)} x_i}$ empirical

μ_j is ...

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

demo: http://web.stanford.edu/class/ee103/visualizations/kmeans/kmeans.html

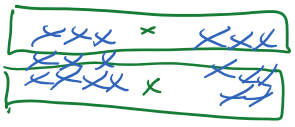
properties of K-means:

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

2) NP hard in general to find best z

kmeans++: clever initialization scheme which guarantees that obj. is within $\log k$ of global opt. (w.h.p.)

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

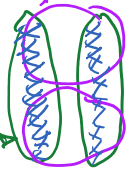


3) choice of k?

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

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

e.g. Dirichlet process mixture model



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

↳ GMM fixes that

16h12

EM - maximum likelihood in latent variable model

Setup:

x observed variable

z latent "

log-likelihood

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

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

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

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

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

options for ML in latent variable model

1) do gradient ascent on 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"

ie. E step → fill z with "soft-values"

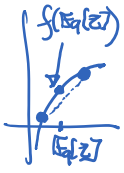
M step → max w.r.t. θ for fully observed model

trick overview:

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

Jensen's inequality

$\mathbb{E}_q[f(z)] \leq f(\mathbb{E}_q[z])$ when f is concave



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

Jensen's ineq. trick?

$$\triangleq \mathcal{L}(q, \theta) \triangleq \mathbb{E}_{q(z)} [\log p(x,z;\theta)] + H(q)$$

"entropy of q"

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

in Jensen ineq, we get equality only if $f(z) =$ constant.

ie. $\frac{p(x,z)}{q(z)} =$ constant $\forall z \Rightarrow q^*(z) \propto p(x,z)$

ie. $q^*(z) = p(z|x;\theta)$

this means argmax $\mathcal{L}(q, \theta_t) = p(z|x;\theta_t)$ $q \in$ all dist. over z

inference

EM algorithm: E step: $q_{t+1} \triangleq \arg\max_q \mathcal{L}(q, \theta_t) \Rightarrow q_{t+1}(z) = p(z|x;\theta_t)$

M step: $\theta_{t+1} \triangleq \arg\max \mathcal{L}(q_{t+1}, \theta)$

$= \arg\max_\theta \mathbb{E}_{q_{t+1}(z)} \log[p(x,z;\theta_{t+1})]$

"expected complete log-likelihood"

this is another ML problem, but for complete information

(usually, replace z with $\mathbb{E}_q[z]$)