Lecture 9 - scribbles

Tuesday, October 2, 2018 14:31

today: • Fisher model for classification + math tricks
• unsupervised learning: k-mean

generative model for classification: (Fisher) linear discriminant analysis

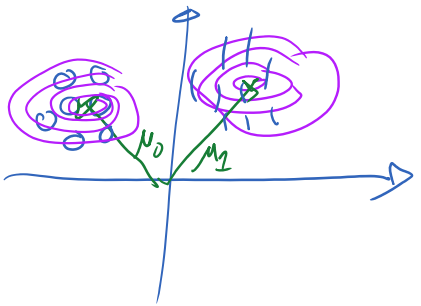
for classification $Y \in \{0,1\}$
$X \in \mathbb{R}^d$

generative approach: $p(x,y;\theta) = \overbrace{p(x|y;\theta)}^{\text{class conditionals}} p(y;\theta)$

vs.

cond. approach: $p(y|x;\theta)$

for Fisher model, we assume $p(x|y;\theta) = N(x|\mu_y, \Sigma)$ (same for both classes)



$\theta = (\mu_0, \mu_1, \underset{d \times d}{\Sigma}, \pi)$

- μ_0, μ_1: mean for class 0
- Σ: shared for 2 classes
- π: prior $p(y=1)$

can then show that $p(y|x;\theta) = \sigma(w^T x)$

where w is a function of $(\mu_0, \mu_1, \Sigma, \pi)$

[if you use $\Sigma_0 \neq \Sigma_1$, get "quadratic discriminant analysis" (QDA)]

i.e. $\sigma(w^T \varphi(x))$ where $\varphi(x)$ is quadratic function of x

~ see hwk 2

(*) gen. approach: do joint MLE to estimate $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_i \log p(x_i, y_i; \theta)$

Sidenote: MLE for multivariate Gaussian

$X_i \sim N(\mu, \Sigma)$ $\quad \mu \in \mathbb{R}^d$

$\Sigma \in \mathbb{R}^{d \times d}$, Σ is symmetric

$\Sigma \succ 0$

$$p(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

$\operatorname{tr}(AB) = \operatorname{tr}(BA)$

$$\operatorname{tr}\left((x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

$$\operatorname{tr}\left(\Sigma^{-1} (x-\mu)(x-\mu)^T\right) = \langle \Sigma^{-1}, (x-\mu)(x-\mu)^T \rangle$$

(inner product)

$\theta = (\mu, \Sigma)$

$\langle A, B \rangle \triangleq \sum_{ij} A_{ij} B_{ij} = \operatorname{tr}(A^T B)$

$$\text{log-likelihood}: \sum_{i=1}^n \log p(x_i; \theta) = \text{cst.} - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \cdot n \left\langle \Sigma^{-1}, \underbrace{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T}_{\tilde{\Sigma}} \right\rangle$$

vector derivative review:

Suppose $f: \mathbb{R}^m \to \mathbb{R}^n$

f is differentiable at x_0 iff $\exists$ a linear operator $df_{x_0}: \mathbb{R}^m \to \mathbb{R}^n$

s.t. $\forall \Delta \in \mathbb{R}^m \quad f(x_0+\Delta) - f(x_0) = df_{x_0}(\Delta) + o(\|\Delta\|)$

"little oh" → $o(\|\Delta\|)$ means is some fct. $h(\|\Delta\|)$ s.t. $\lim_{\|\Delta\|\to 0} \frac{h(\|\Delta\|)}{\|\Delta\|} \to 0$

"differential"

$f(x) \in \mathbb{R}^n$

$\Delta \in \mathbb{R}^m$

df_{x_0} is linear

means $df_{x_0}(\Delta_1 + b\Delta_2) = df_{x_0}(\Delta_1) + b\, df_{x_0}(\Delta_2)$

can represent as a $n \times m$ matrix

called the Jacobian matrix

standard representation $(df_{x_0})_{ij} = \frac{\partial f_i}{\partial x_j}$ (i^{th} component of f, j^{th} component of x)

1) this gives you a way to get df_{x_0} for anything (matrix, tensor, ...)

2) be careful with dimension: differential of $f: \mathbb{R}^m \to \mathbb{R}$

is a row vector ($1 \times m$) i.e. $df_{x_0} = (\nabla f(x_0))^T$

chain rule: suppose $f: \mathbb{R}^m \to \mathbb{R}^n$

$g: \mathbb{R}^n \to \mathbb{R}^q$

$$d(g \circ f)_{x_0} = dg_{f(x_0)} \circ df_{x_0}$$

"product of Jacobians"

"product of Jacobians"

e.g. $f(x) = x - \mu$ $\quad df_{x_0} = I$

$g(x) = x^T A x$ $\quad dg_{x_0} = x^T (A + A^T)$ $\quad [\nabla g^T]$

$$d(g \circ f)_{x_0} = dg_{f(x_0)} \circ df_{x_0}$$
$$= (x-\mu)^T (A + A^T) \cdot I$$

$$g(f(x)) = (x-\mu)^T A (x-\mu)$$

---

for Gaussian : $-\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$

∇_μ $\quad -\frac{1}{2} \sum_{i=1}^{n} 2\Sigma^{-1}(x_i - \mu) \overset{\text{want}}{=} 0$

$$\Rightarrow \boxed{\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

example 2 : derivative of $f(A) \triangleq \log \det(A)$

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

$$f(A+\Delta) - f(A) = \mathrm{tr}\left(f'(A)^T \Delta\right) + O(\|\Delta\|)$$
$$= \langle f'(A), \Delta \rangle + \|$$

$\log\det(A+\Delta) - \log\det(A)$ $\quad$ [assume A is symmetric and invertible]

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

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

$$= \log|A^{1/2}|\,|I + A^{-1/2}\Delta A^{-1/2}|\,|A^{1/2}| - \log|A|$$

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

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

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

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

$$= \underbrace{\mathrm{tr}(A^{-1}\Delta)}_{\langle A^{-1}, \Delta\rangle} + O(\|\Delta\|)$$

e-values

use $\det(A) = \prod_i \lambda_i(A)$

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

use $\mathrm{tr}(A) = \sum_i \lambda_i(A)$

$$\Rightarrow \boxed{\frac{d}{dA}\log\det(A) = A^{-1}}$$

back to our log-likelihood:

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

derivative w.r. to Σ^{-1}: $\frac{n}{2}\underbrace{(\Sigma^{-1})^{-1}}_{\Sigma} - \frac{n}{2}\tilde{\Sigma} \overset{\text{want}}{=} 0$

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

some references:

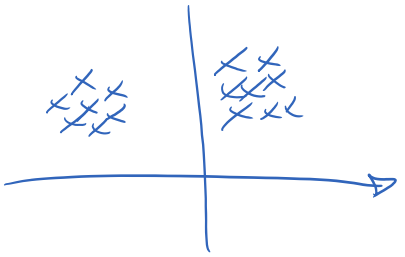
- for convex optimization: Boyd & Vandenberghe's book
- DL book -- chapter 4.3 on gradient-based optimization
- for matrix calculus:
  Matrix Differential Calculus with Applications in Statistics and Econometrics, Heinz Neudecker and Jan R. Magnus -- free

online

15h55

Unsupervised learning

here X without labels Y



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

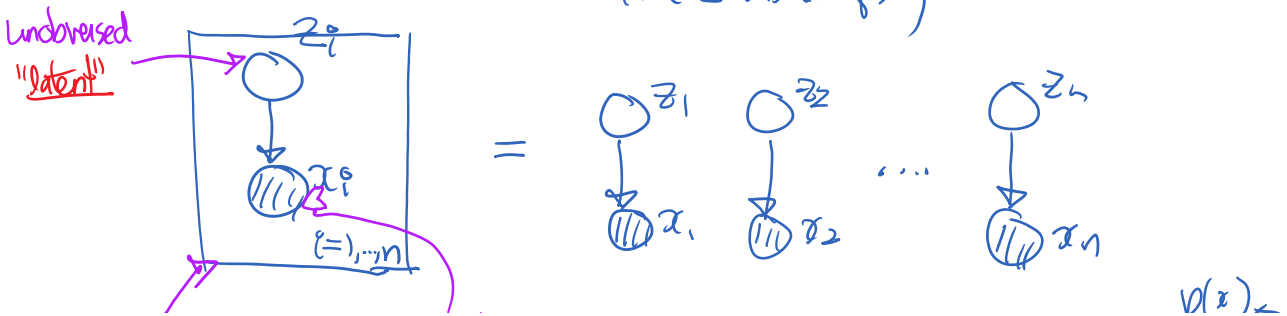
$$Y \sim \text{Mult}(\pi) \quad \pi \in \Delta_K$$

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

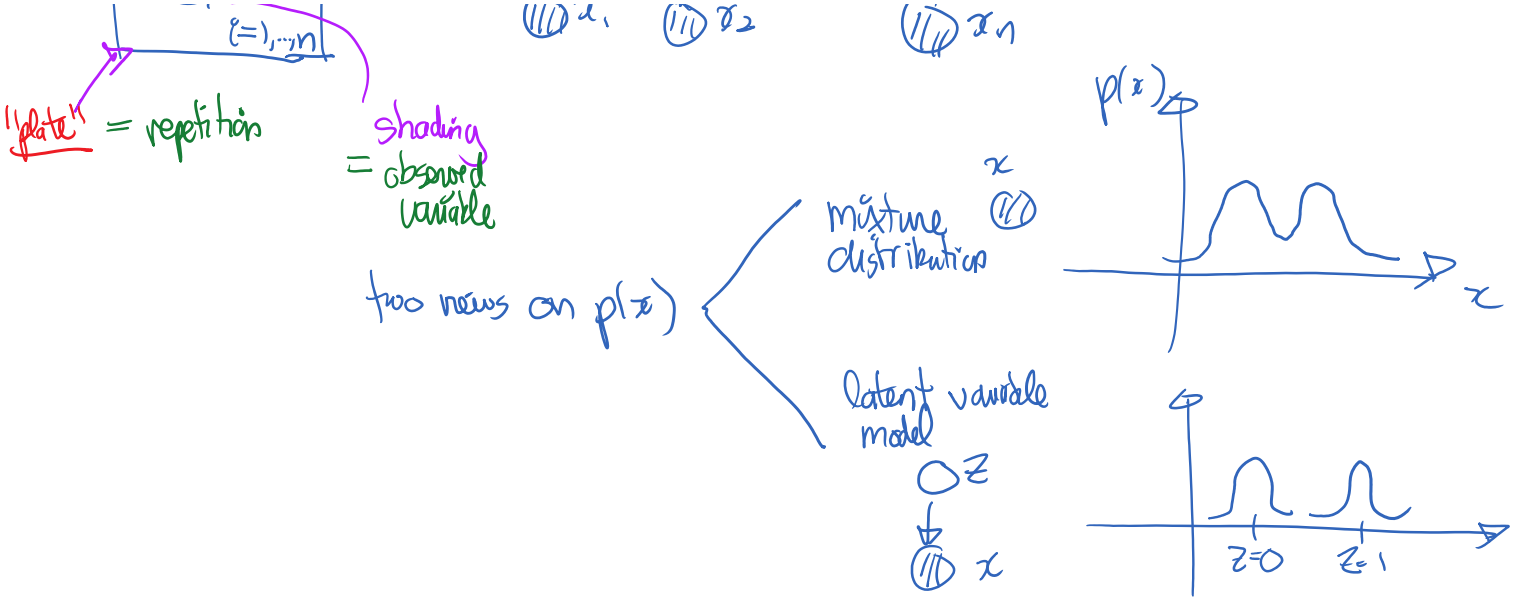
$$p(x) = \sum_y p(x,y) = \sum_{j=1}^{K} \pi_j N(x | \mu_j, \Sigma)$$

GMM model, more generally could have different covariance per class Σ_j

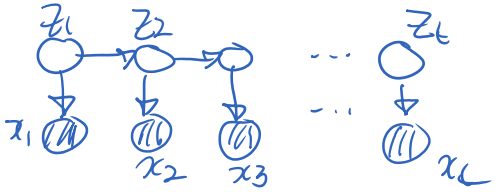
graphical model for this "latent variable model"
(use z instead of y)



$p(x)$



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



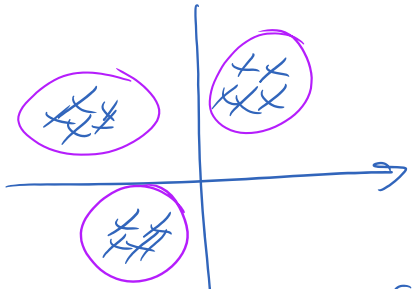
K-means → to do clustering, i.e. group data

can be seen a limit of GMM

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

represent $z_{ij}=1$ to mean 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

k-mean algorithm → can be seen as block-coordinate minimization of objective fct :

$$J(z, \mu) \triangleq \sum_{i=1}^{n} \left(\sum_{j=1}^{k} z_{i,j} \|x_i - \mu_j\|^2 \right)$$

"distorsion measure"

z: cluster assign. $z_1, \dots, z_n \in$ corners of Δ_k

$\mu_1, \dots \mu_k \in \mathbb{R}^d$ cluster means $\quad \|x_i - \mu_{z_i}\|^2$

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

2) iterate until convergence

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

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

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

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

empirical mean with cluster

16h24

mean with cluster

16h24

properties (next class)