Lecture 9 - Fisher LDA + Gaussian MLE

Tuesday, October 3, 2023 2:32 PM

today: • Fisher LDA
• math tricks & MLE for Gaussian

generative model for classification: (Fisher) linear discriminant analysis

FLD (instead of LDA)

for classification $Y \in \{0,1\}$

$X \in \mathbb{R}^d$

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

Vs.

conditional approach $p(y|x;\theta)$

⊛ for Fisher model: we assume $p(x|y;\theta) = N(x|\mu_y, \Sigma)$ ← shared across classes

$\theta = (\mu_0, \mu_1, \Sigma, \pi)$

μ_0: mean for class 0; Σ: shared covariance; π: $p(y=1)$

as before (see exponential family argument)

can show that $p(y|x;\theta) = \sigma(w^T x)$ where w is a fct. of $(\mu_0, \mu_1, \Sigma, \pi)$

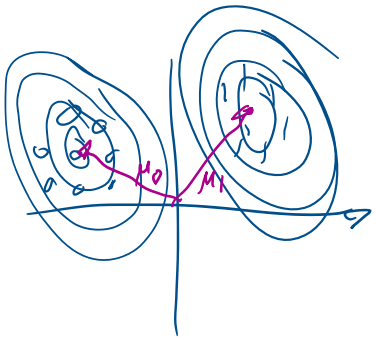
[note: if use $\Sigma_0 \neq \Sigma_1$, get "quadratic discriminant analysis" (QDA)]

ie. $\sigma(w^T \varphi(x))$ where $\varphi(x)$ is a quadratic fct. of x [see hwk. 2]

⊛ generative approach: do joint MLE to estimate

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_i \log p(x_i, y_i; \theta)$$

[Vs. $\arg\max_w \sum_i \log p(y_i|x_i;w)$ for logistic regression]



sidenote: MLE for multivariate Gaussian

$X_i \overset{iid}{\sim} N(\mu, \Sigma)$

$\mu \in \mathbb{R}^d$

$\Sigma \triangleq \mathbb{E}[(X-\mu)(X-\mu)^T]$

$\Sigma = \mathbb{E}[(x-\mu)(x-\mu)^T]$

$\mu \in \mathbb{R}^d$

$\Sigma \in \mathbb{R}^{d \times d}$

Σ is symmetric

$\Sigma \succeq 0$

$\Sigma^T = \quad = \Sigma$

$v^T \Sigma v = \mathbb{E}[v^T (x-\mu)(x-\mu)^T v]$

$\mathbb{E}[((x-\mu)^T v)^2] \geq 0$ $\Rightarrow \Sigma \succeq 0$

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

$tr(AB) = tr(BA)$

$(x-\mu)^T \Sigma^{-1}(x-\mu)$ is a scalar, $= tr((x-\mu)^T \Sigma^{-1}(x-\mu))$

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

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

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

$|\Sigma^{-1}| = |\Sigma|^{-1} = \frac{1}{|\Sigma|}$

Vector derivative review:

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

$h \in o(\|\Delta\|)$ "little oh" $\Leftrightarrow \lim_{\|\Delta\| \to 0} \frac{h(\|\Delta\|)}{\|\Delta\|} = 0$ e.g. $\|\Delta\|^2 \in o(\|\Delta\|)$

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

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

"differential"

"derivative"

$$\lim_{\|\Delta\| \to 0} \frac{f(x_0+\Delta) - f(x_0)}{\|\Delta\|} = \lim_{\|\Delta\| \to 0} \left(\underbrace{df_{x_0}\left(\frac{\Delta}{\|\Delta\|}\right)}_{d} + \frac{o(\|\Delta\|)}{\|\Delta\|} \right)$$

$\frac{o(\|\Delta\|)}{\|\Delta\|} \to 0$

$df_{x_0}(\frac{\Delta}{\|\Delta\|}) \to d$

→ directional derivative at x_0 in direction d

if $n=1$: $df_{x_0}(d) = \langle \nabla f(x_0), d \rangle$

$\nabla f(x_0) = (df_{x_0})^T$

df_{x_0} is linear

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

can represent as a $n \times m$ matrix called the Jacobian matrix.

standard representation $(df_{x_0})_{ij} = \frac{\partial f_i}{\partial x_j}$

then $df_{x_0}(\Delta) = df_{x_0} \cdot \Delta$

1) this gives a way to get df_{x_0} for "anything" (matrix, tensor, no-dim fct., etc.)

2) [illegible]

1) this gives a way to get df_{x_0} for "anything" (matrix, tensor, ∞-dim fct., etc.)

2) be careful with dimensions

$f: \mathbb{R}^m \to \mathbb{R}$ $\quad df_{x_0}$ is a row vector $(1 \times m)$

$df_{x_0} = (\nabla f(x_0))^T$

Chain rule: $f: \mathbb{R}^m \to \mathbb{R}^n$, $g: \mathbb{R}^n \to \mathbb{R}^q$

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

$g(f(x_0))$ $\quad = (\ ) \cdot (\ )$ matrix product of Jacobians

e.g. $f(\mu) = x - \mu$

$g(v) = v^T A v$

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

$df_{\mu_0} = -I$

$dg_{v_0} = v^T (A + A^T)$

$d(g \circ f)_{\mu_0} = dg_{f(\mu_0)} \circ df_{\mu_0}$

$= (x - \mu_0)^T (A + A^T)(-I)$

---

for Gaussian: $-\frac{1}{2} \sum_i (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$

∇_μ $\quad +\frac{1}{2} \sum_i 2\Sigma^{-1}(x_i - \mu) \overset{want}{=} 0 \quad \Rightarrow \quad \boxed{\mu_{MLE} = \frac{1}{n} \sum_i x_i}$

$(d(g \circ f)_\mu)^T$

15h29

example 2: derivative of $f(A) \triangleq \log \det(A)$ where assume A is symmetric $A \succ 0$

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

$$f(A+\Delta) - f(A) = \mathrm{tr}(f'(A)^T \Delta) + O(\|\Delta\|)$$

$$= \langle f'(A), \Delta \rangle + O(\|\Delta\|)$$

$f'(A) = df_A$

$A \succ 0 \Rightarrow$ invertible; has unique square root $A^{1/2}$

$\log\det(A+\Delta) - \log\det(A)$

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

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

$= \log |I + A^{-1/2}\Delta A^{-1/2}|$ $\quad$ use $\det(B) = \prod_i \lambda_i(B)$ (e-values of B) $\quad Bv = \lambda v$

$= \sum \log \lambda_i(I + A^{-1/2}\Delta A^{-1/2})$ $\quad$ use $\lambda(I+B) = 1 + \lambda(B)$ $\quad (I+B)v = (1+\lambda)v$

$= \sum_i \log \lambda_i(I + A^{-1/2}\Delta A^{-1/2})$ use $\lambda(I+B) = 1 + \lambda(B)$ $(I+B)v = (1+\lambda)v$

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

$= \sum_i \lambda_i(A^{-1/2}\Delta A^{-1/2}) + O\left(\left[\sum_i \lambda_i(A^{-1/2}\Delta A^{-1/2})\right]^2\right)$

because λ is homogenous fct. i.e. $Bv = \lambda v$, $\left(\frac{B}{b}\right)v = \left(\frac{\lambda}{b}\right)v$

$\|\Delta\|^2 \lambda_i\left(A^{-1/2}\frac{\Delta}{\|\Delta\|}A^{-1/2}\right)^2$

sup over $\|\Delta\| \le 1$; cst. with respect $\|\Delta\|$

$O(\|\Delta\|^2) = o(\|\Delta\|)$

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

$= \mathrm{tr}(A^{-1}\Delta) + o(\|\Delta\|)$ $\Rightarrow$ $\boxed{\frac{d}{dA}\log\det(A) = A^{-1}}$ //

$= \langle A^{-1}, \Delta \rangle$

(recall A is symmetric)

see Boyd's book A.4.1 for the above proof

back to log-likelihood of Gaussian

$+\frac{n}{2}\log|\Sigma^{-1}| - \frac{n}{2}\langle \Sigma^{-1}, \tilde{\Sigma}(\mu)\rangle$ (concave fct. of $\Lambda = \Sigma^{-1}$)

take derivative w.r.t. to $\Sigma^{-1} = \Lambda$

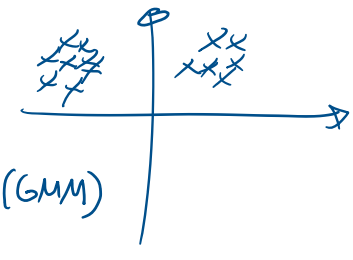
$\frac{n}{2}\underbrace{(\Sigma^{-1})^{-1}}_{\Sigma} - \frac{n}{2}\tilde{\Sigma}(\mu) \overset{\text{want}}{=} 0$

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

(the empirical covariance matrix)

unsupervised learning

here X without any label Y



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

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

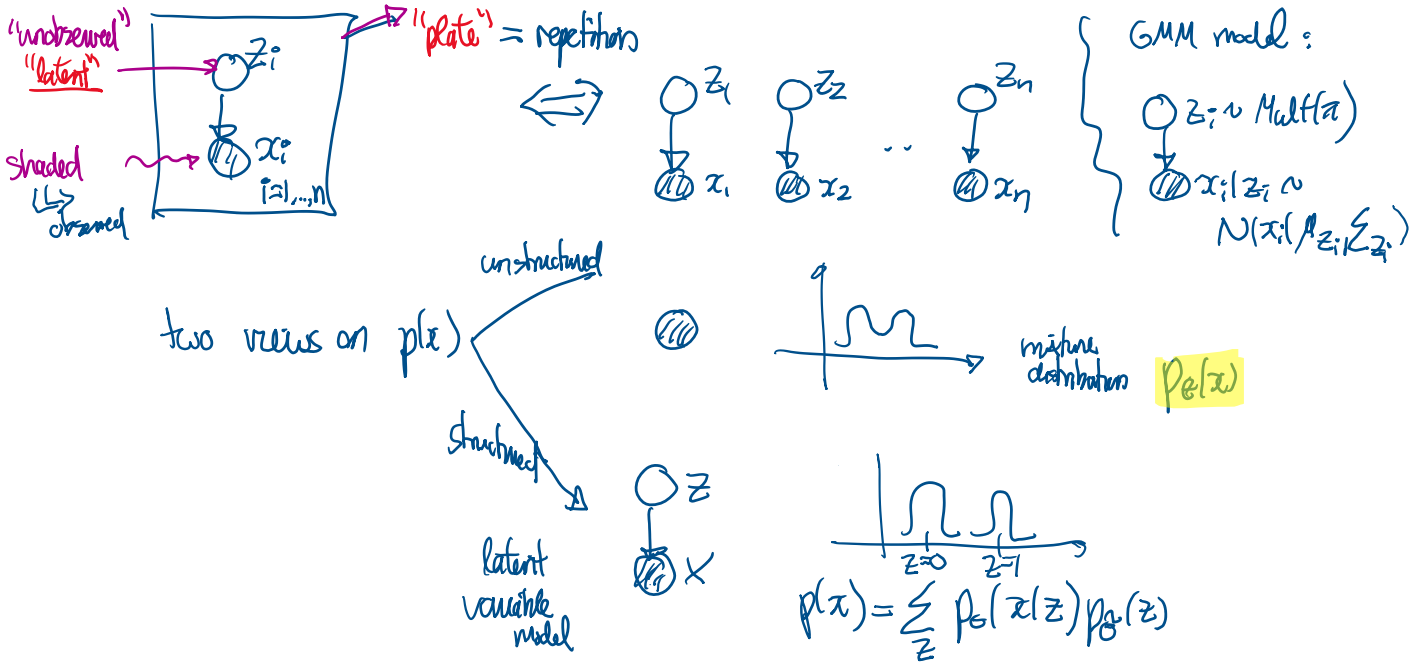
$Y \sim \text{Mult}(\pi)$ $\pi \in \Delta_K$ [extension of FLD to multiple classes]

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

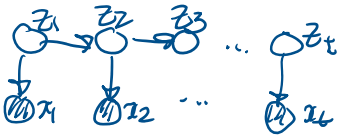
$$p(x) = \sum_y p(x,y) = \sum_y p(x|y)p(y) = \sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma)$$

"GMM" (more generally, can have Σ_j per class)

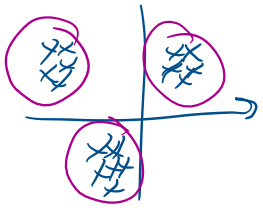
graphical model for this "latent variable model"



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



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



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

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

$j = 1, \ldots, K$ # 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 many others!