# Lecture 9 - Fisher LDA + Gaussian MLE

Thursday, October 3, 2024 2:39 PM

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

generative model for classification : (Fisher) linear discriminant analysis

FLD (instead of LDA)

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

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 class)

$\theta = (\mu_0, \mu_1, \Sigma, \pi)$ ($\pi$ = $p(y=1)$)

$\begin{pmatrix} \mu_0 \\ \mu_1 \\ \text{vec}(\Sigma) \\ \pi \end{pmatrix}$

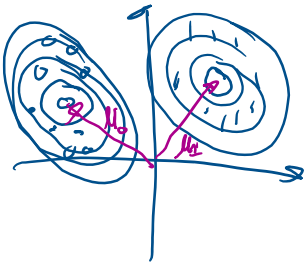
as before (see exponential family argument)

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 $\Sigma_0$ & $\Sigma_1$, get "quadratic discriminant analysis" QDA]

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

⊛ generative approach: do joint MLE to estimate $\theta$:




$$\hat{\theta} = \underset{\theta \in \Theta}{\text{argmax}} \sum_i \log p(x_i, y_i; \theta)$$

[vs. $\underset{w}{\text{argmax}} \sum_i \log p(y_i|x_i; w)$ for logistic regression]

scribble: MLE for multivariate Gaussian

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

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

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

$\Sigma$ is symmetric

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

$\Sigma^T = \mathbb{E}[\ \ ] = \Sigma$

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

$\Sigma \in \mathbb{R}^{d\times d}$
$\Sigma$ is symmetric
$\Sigma \succ 0$

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

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

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

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

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

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

## Vector derivative review:

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

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

"differential"

"derivative"

$$\lim_{\delta\to 0} \frac{f(x_0+\delta\vec{d}) - f(x_0)}{\delta} = \lim_{\delta\to 0} \frac{df_{x_0}(\delta\vec{d})}{\delta} + \frac{o(\delta\|\vec{d}\|)}{\delta\|\vec{d}\|} \to 0$$

$$\frac{\delta[df_{x_0}(\vec{d})]}{\delta}$$

$\to$ directional derivative at $x_0$ in direction $\vec{d}$

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

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

$df_{x_0}$ 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 you a way to get $df_{x_0}$ for "anything" (matrix, tensor, n-dim fct. etc...

2) be careful with dimensions

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

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

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

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

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

eg $f(\mu) = x - \mu$ $\qquad df_{\mu_0} = -I$

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

$g\circ f(\mu) = (x-\mu)^T A (x-\mu)$ $\qquad d(g\circ f)_{\mu_0} = dg_{f(\mu_0)} \circ df_{\mu_0}$

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

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

$\nabla_\mu$ $\quad (d(g\circ f)_\mu)^T$ $\quad \left[+\frac{1}{2}\sum_i 2\Sigma^{-1}(x_i-\mu)\right] \overset{\text{want}}{=} 0 \Rightarrow \boxed{\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^n x_i}$

15h38

**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

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

$$= \langle \underbrace{f'(A)}_{df_A}, \Delta \rangle + o(\|\Delta\|)$$

$A \succ 0 \Rightarrow$ 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 $\det(B) = \prod_i \lambda_i(B)$ ($\lambda_i$: e-values of $B$)

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

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

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

$\lambda$ is homogeneous fct. $f(\alpha v) = \alpha f(v)$

because $\lambda$ is homogeneous fct.

ie. $BV = \lambda V$

$(B)V = (\lambda)V$

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

sup $\|\Delta\|=1$ cst. with respect

$\|\Delta\| \, \varepsilon(A^{-1/2}\frac{\Delta}{\|\Delta\|}A^{-1/2})$ — $\sup_{\|\Delta\|=1}$ cst. with respect to $\|\Delta\|$

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

ie. $BV = \Delta V$, $(\frac{B}{b})V = (\frac{\Delta}{b})V$

$\frac{h(\Delta)}{\|\Delta\|} \le c\frac{\|\Delta\|^2}{\|\Delta\|} = c\|\Delta\|$

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

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

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

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

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. $\Sigma^{-1} = \Lambda$

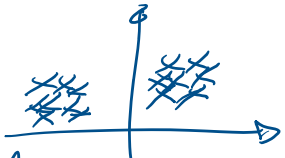
$\frac{n}{2}\underbrace{(\Sigma^{-1})^{-1}}_{\Sigma} - \frac{n}{2}\tilde{\Sigma}(\mu) \overset{want}{=} 0$

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

(the empirical covariance matrix)

## unsupervised learning

here $X$ without any label $Y$



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

[extension of FLD to multiple classes]

$Y \sim \mathrm{mult}(\pi)$ $\pi \in \Delta_K$

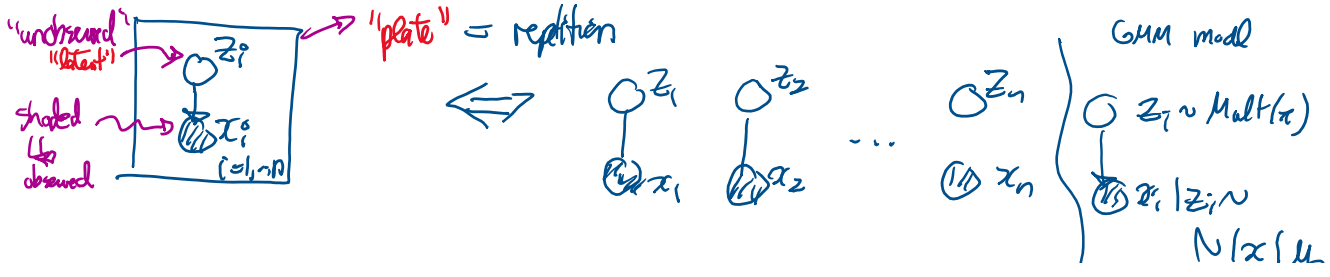
$X|Y=j \sim N(\mu_j, \Sigma)$

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

"GMM"

(more generally, can have $\Sigma_j$ per class)

graphical model for this "latent variable model"



observed