Lecture 9 - Fisher LDA + Gaussian MLE

October 1, 2021 3:01 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 $p(x, y; \theta) = \overbrace{p(x|y;\theta)}^{\text{class conditional}} p(y;\theta)$

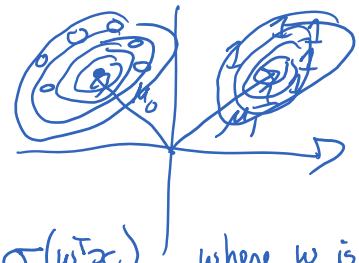
vs.

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

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

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

μ_0, μ_1: mean of class 0; Σ: shared; π: $p(y=1)$



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

($\phi(x)$)

[note: if you use Σ_0 & Σ_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

(*) gen. approach: do joint MLE to estimate

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

[vs. $\sum_i \log p(y_i|x_i;\theta)$ for logistic regression]

Sidenote: MLE for multivariate Gaussian

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

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

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

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

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

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

$tr(AB) = tr(BA)$

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

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

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

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

$+ \frac{n}{2}\log|\Sigma^{-1}|$

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

$\triangleq \tilde{\Sigma}(\mu)$

vector derivative review:

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

means that fct. $h(\|\Delta\|)$ s.t. $\lim_{\|\Delta\|\to 0} \frac{h(\|\Delta\|)}{\|\Delta\|} = 0$ ("little oh")

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"

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

$\vec{d} = \frac{\Delta}{\|\Delta\|}$

$df_{x_0}(\vec{d})$

df_{x_0} is linéen

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

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

- if $n=1$:

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

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

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

2) be careful with dimensions

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

df_{x_0} is a row vector $(1 \times m)$

$df_{x_0} = (Df(x_0))^T$

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

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

$g(f(x))$ — $=$ ( ) ( ) matrix product of Jacobians

e.g. $f(\mu) = x - \mu$ $\quad df_{\mu_0} = -I$

$g(w) = w^T A w$ $\quad dg_{w_0} = w_0^T (A + A^T)$

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

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

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

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

$(d(g \circ f)_{\mu_0})^T$ $\quad$ $\hookrightarrow x_i \in \mathbb{R}^q$

16h05

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

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

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

$\log \det(A+\Delta) - \log \det(A)$ $\quad A \succ 0 \Rightarrow$ invertible & has unique square root $A^{1/2}$

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

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

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

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

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

$\lambda(I + A) = 1 + \lambda(A)$ $\quad$ ($Av = \lambda(A)v$, $(I+A)v = (1+A)v$)

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

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

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

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

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

$O(\|\Delta\|)$

$Av = \lambda v$

$\frac{A}{b} v = \left(\frac{\lambda}{b}\right) v$

$\frac{\|\Delta\|^2}{\|\Delta\|} = \|\Delta\| \to 0$ as $\|\Delta\| \to 0$

$\operatorname{tr}(A) = \sum_j \lambda_j(A)$

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

$$= \underbrace{\operatorname{tr}(A^{-1}\Delta)}_{\langle A^{-1}, \Delta\rangle} + o(\|\Delta\|) \quad \Rightarrow \quad \boxed{\frac{d}{dA}\log\det(A) = A^{-1}}$$

(recall A is symmetric)

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

back to log-likelihood of Gaussians:

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

(concave fct. of $\Lambda = \Sigma^{-1}$)

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

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

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

(the empirical covariance matrix)