Lecture 10 - Fisher LDA + Gaussian MLE

Friday, October 2, 2020 13:30

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

note for hwk 2:

Newton's: $w_{t+1} = w_t - \underbrace{H^{-1}(w_t)\nabla f(w_t)}_{\triangleq d_t}$

instead solve for $H_t d_t = \nabla f(w_t)$ for d_t

$\min_d \|H_t d - \nabla f(w_t)\|$

use numpy.linalg.lstsq

instead of checking $x = y$, $|x-y| < \delta$

$a \sim b$ $a \simeq b$
↑
16 digits of accuracy

$ax^2+bx+c=0$

$\frac{-b+\sqrt{b^2-4ac}}{2a}$ if b is huge, $b^2 \gg 4ac$

$-b+\sqrt{b^2 \sim\sim}$ → cancellation error

instead → $\frac{-b+\sqrt{\sim\sim}}{2a}\left(\frac{-b-\sqrt{\sim\sim}}{-b-\sqrt{\sim\sim}}\right)$

$= \frac{b^2-(b^2-4ac)}{2a(-b-\sqrt{\sim\sim})} = \frac{4ac}{-2a(b+\sqrt{\sim\sim})}$

$= \boxed{\frac{-2c}{b+\sqrt{b^2-4ac}}}$

generative model for classification: (Fisher) linear discriminant analysis

FLD (instead of LDA)

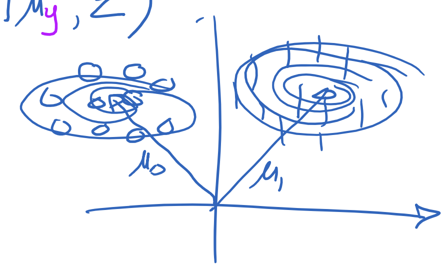
for classification $Y \in \{0,1\}$, $X \in \mathbb{R}^d$

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: 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 (μ_0,μ_1,Σ,π) (x: $\varphi(x)$)

[note: if you use Σ_0 & Σ_1, get "quadratic discriminant analysis" (QDA)]

(QDA)

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

↝ see hwk 2

✳ gen. approach: do joint MLE to estimate

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_i \log p(x_i, y_i; \theta) \quad \left[\text{vs. } \sum_i \log p(y_i | x_i) \text{ for log. reg.}\right]$$

side note: 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, $\Sigma \succ 0$ $\quad \mathbb{E}[(X-\mu)(X-\mu)^T] = \Sigma$

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

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

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

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

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

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

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

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

vector derivative review:

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

"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")

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

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

i^{th} component of f

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

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

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

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

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

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

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

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

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

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

2:41pm -> back at 2:51pm!

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

$g(x) = x^T A x$ $\qquad dg_{x_0} = x^T(A + A^T)$

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

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

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

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

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

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

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

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

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

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

$= \log |A|^{1/2} |I + A^{-1/2}\Delta A^{-1/2}| |A|^{1/2} - \log\det(A)$ $\qquad$ e-values of A

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

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

use $\det(A) = \prod_i \lambda_i(A)$ (product of eigenvalues of A)

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

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

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

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

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

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

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

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

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

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

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

back to log-likelihood of Gaussian

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

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

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

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

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

(the empirical covariance matrix)