Lecture 9 - Fisher LDA + Gaussian MLE

Tuesday, October 4, 2022 2:56 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 $p(x, y; \theta) = \underbrace{p(x|y;\theta)}_{\text{class conditional}} p(y;\theta)$

vs.

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

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

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

mean of class 0 — shared — $p(y=1)$

as before (see exponential family argument)

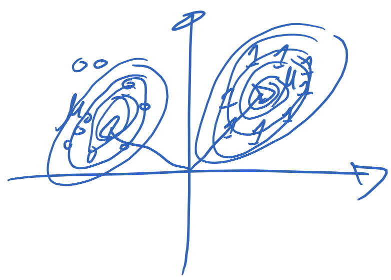
can show that $p(y|x;\theta) = \sigma(w^T x)$ where w is a fct. of $(\mu_0, \mu_1, \Sigma, \pi)$

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

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

gen. approach: do joint MLE to estimate

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



[vs. $\underset{w}{\arg\max} \sum_i \log p(y_i | x_i; w)$

for logistic regression]

sidenote: MLE for multivariate Gaussian

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

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

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

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

$\Sigma^T = \quad '' \quad = \Sigma$

$\mu \in \mathbb{R}$

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

Σ is symmetric

$\Sigma \succeq 0$

$\Sigma^T = \quad '' \quad = \Sigma$

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

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

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

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

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

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

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

<u>vector derivative review</u>:

$o(\|\Delta\|)$ "little oh" means that it's fct. $h(\|\Delta\|)$ s.t.

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

$$\boxed{\begin{array}{l} f \text{ is } \underline{\text{differentiable at}}\ x_0 \text{ iff } \exists \text{ a } \underline{\text{linear operator}}\ df_{x_0}: \mathbb{R}^m \to \mathbb{R}^n \\ \text{s.t. } \forall \Delta \in \mathbb{R}^m \quad f(x_0+\Delta) - f(x_0) = df_{x_0}(\Delta) + o(\|\Delta\|) \end{array}}$$

"differential"

"derivative"

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

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

→ directional derivative of f at x_0 in direction $\vec{d}$

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 <u>Jacobian matrix</u>

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

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

• 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 fct, etc.)

2) be careful with dimensions

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

df_{x_0} is a <u>row vector</u> $(1\times m)$

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

<u>chain rule</u>:

$f: \mathbb{R}^m \to \mathbb{R}^n$ $\quad d(g\circ f) \quad dg \ldots df$

chain rule:

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

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

$g(f(x_0))$

$= (\;)(\;)$ matrix product of Jacobians

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

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

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

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

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

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

---

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

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

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

15h58

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

$$f(A+\Delta) - f(A) = tr(f'(A)^T\Delta) + o(\|\Delta\|)$$
$$= \langle \underbrace{f'(A)}_{df_A}, \Delta \rangle + o(\|\Delta\|)$$

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

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

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

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

$= \log |I + A^{-1/2}\Delta A^{-1/2}|$ $\qquad$ use $\det(B) = \prod_i \lambda_i(B)$ e-value of B

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

$\lambda(I+B) = 1 + \lambda(B)$ $\qquad (I+B)v = (1+\lambda)v$

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

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

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

$= \|\Delta\|^2 \underbrace{\lambda_i\left(A^{-1/2}\frac{\Delta}{\|\Delta\|}A^{-1/2}\right)^2}_{\text{cst.}}$

λ is homogeneous fct.

i.e. $Bv = \lambda v$

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

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

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

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

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

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

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

(recall A is symmetric)

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

back to log-likelihood of Gaussian:

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

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

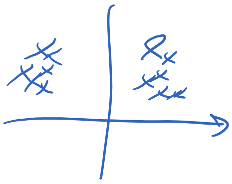
$$\frac{n}{2} \underbrace{(\Sigma^{-1})^{-1}}_{\Sigma} - \frac{n}{2} \tilde{\Sigma}(\mu) \overset{\text{want}}{=} 0$$

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

(the empirical covariance matrix)

unsupervised learning

here X without any label Y



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

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

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

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

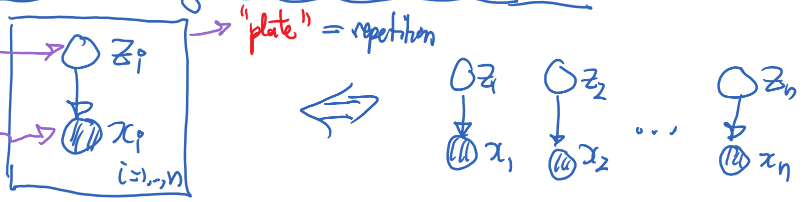
"GMM model" $\quad$ (more generally, can have Σ_j per class)

graphical model for this "latent variable model"

"unobserved" "latent"

shaded ↳ observed

"plate" = repetition

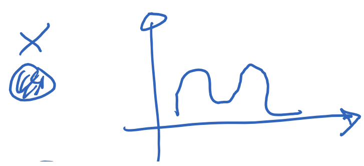


GMM model:

$z_i \sim \text{Mult}(\pi)$

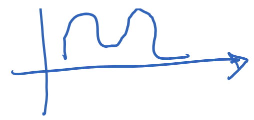
$x_i | z_i \sim N(x_i | \mu_{z_i}, \Sigma)$

two views on $p(x)$ $\quad$ unstructured view



mixture distribution $p_\theta(x)$

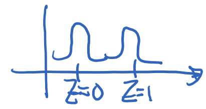
two views on $p(x)$



mixture distribution $p_\theta(x)$

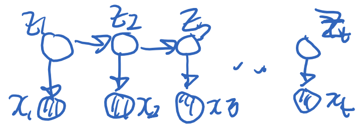
structure

latent variable model

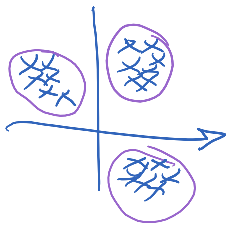


$p(x) = \sum_z p_\theta(x|z)\, p_\theta(z)$

(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 pt. x_i:

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

$j = 1, \dots, k$ ← # of clusters (specified in advance for K-means)

applications: • vector quantization (compression)

• in computer vision: use K-means to get "bag of visual words" representation of image patches

• many many others!