Final Review Side A

2023年4月14日 2:52

- Single Value Decomposition svd()

L_2 Euclidean norm : $\|x\| = (\sum x_i^2)^{1/2}$
L_1 Manhattan norm : $\|x\| = \sum |x_i|$
matrix norm : $\|A\| = \{\text{trace}(A^TA)\}^{1/2}$

$X = UDV^T$, U $n \times p$ with orth col, V $p \times p$ with orth col

$D = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_p \end{pmatrix}$ $d_1 \geq d_2 \cdots \geq d_p = 0$

$X = \sum_1^p d_j u_j v_j^T$ where $u_j v_j$ col of U, V.

$X^* = O_1 X O_2$, O_1 $n \times n$, O_2 $p \times p$ $\Rightarrow$ svd(X) = svd(X^*)
$\|X\|_2 = d_1$ $\|X\|_F = (d_1^2 + \cdots d_p^2)^{1/2}$

- $\boxed{\text{PCA : reduce dimensions}}$

find $r < p$, that has same structure.

$sd = \sqrt{\sum \lambda_i} = 5$
$\sum PC(i)^2 = 1$ $\sum PC(j)^2 =$ prop var
principal component is eigen vector of C times x

PCA define $y_i = Ax_i$ A $r \times p$ matrix princomp() PCs are orthogonal
$X = UDV^T$, $X^* = UD^*V^T$, $D^* = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_r \end{pmatrix}$ Comp $= \sum$ loading$(i) \cdot x_i$
Let $\hat{X} = X - \bar{X}$, $\tilde{X} = UDV^T$, $S = \frac{1}{n-1} VD^2V^T$ PCs are uncorrelated

Define $Y = \hat{X}V = (\hat{X}v_1 \cdots \hat{X}v_p)$ pc-scores: column of Y (UD or $\hat{X}V$) $\ell_j \cdot y_i$
pc-loadings: vector $v_1 \cdots v_p$

first pc: overall strength
second pc: contrast

Andrew Curve (identify outliers)
$\phi_1 = \frac{1}{\sqrt{2}}$ $\phi_2 = \sin(2\pi t)$ $\phi_3 = \cos(2\pi t)$ $\phi_4 = \sin(4\pi t)$ $\phi_5 = \cos(4\pi t)$ $\int_0 \phi_k \phi_\ell = 0$

- $\boxed{\text{ICA}}$ replace assumption $Cov(Y_i) = I$ with Y_i independent
$X_i = \mu + AY_i$ Y_i independent $E(Y_i) = 0$ $Cov(Y_i) = I$ A is mixing matrix.

two steps ① prewhitening L ② component extraction W
- to estimate A, $\hat{A} = (WL)^{-1} = L^{-1}W^T$, $y_i = \hat{A}^{-1}(x_i - \bar{x})$

pre-white : find L that $Cov(L(x_i - \mu)) = Cov(LX_i) = I$
comp extract : choose W that comp of $Y_i = WZ_i$ independent
(pratically impossible)

Entropy - $H(X) = -\int_{-\infty}^{\infty} f(x) \ln(f(x)) dx = -E(\ln f(x))$

fastICA() X: pre-whitened centered W: estimated orthogonal
K: pre-whiten matrix A: estimated transpose of A

- $\boxed{\text{Factor Analysis}}$ represent X in terms of unobserved factors.

$C = V \Lambda V^T$, Λ diagonal $\lambda_1 \geq \lambda_2 \cdots \lambda_r > 0$
$C = (\lambda^{1/2} v_1 \cdots)(\lambda^{1/2} v_1 \cdots)^T = LL^T$ L is $p \times r$

Factor Analysis Model: factanal()
$X = \mu + LF + \varepsilon = \mu + \sum^r F_k \ell_k + \varepsilon$. F_k: factors, ℓ_k: loadings

$Cov(F) = I$, $Cov(\varepsilon) = \psi = \begin{pmatrix} \psi_1 & & \\ & \ddots & \\ & & \psi_p \end{pmatrix}$, $\psi_i > 0$, $C = LL^T + \psi$
L is loading matrix (not unique)

$Y = DX + a \Rightarrow Y = \underbrace{D\mu + a}_{\text{mean}} + \underbrace{DL}_{\text{loadings}} F + D\varepsilon$

- Estimate L, ψ

① Iterative principal factor - until convergence
$\hat{L}\hat{L}^T + \hat{\psi} \approx S = \frac{1}{n-1} \sum (x_i - \bar{x})(x_i - \bar{x})^T$

② Maximum likelyhood estimation
Compute $\mathcal{L}(L, \psi)$. test $H_0 : C = LL^T + \psi$
Likelyhood-Ratio test
large p-value : r factor is appropriate

- Scaling with rotation - easier to interpret ($\ell_{jk} = 0$ for most j)

Orthogonal - LO where O orth
Oblique - LO where O invertible (all factors correlated)
↳ $X = \mu + [LO][O^{-1}F] + \varepsilon$, $Cov(F^*) = [O^TO]^{-1}$

Varimax - max variance of squared elements of LO for O orth
Promax - power up variance loading - result in non-orth loadings

$\boxed{\text{Cluster Analysis}}$ - identify similar groups of observations

① Hierarchical clustering : use distance hclust()
Start with n clusters, then combine clusters

distance matrix $D = \begin{pmatrix} 0 & d(x_1, x_2) & \cdots \\ \vdots & \ddots & \vdots \\ d(x_n, x_1) & \cdots & 0 \end{pmatrix}$

$d_s \leq d_a \leq d_c$

single linkage	$d_s(U, V) = \min\{d(x_i, x_j)\}$	produce less compact clusters
complete linkage	$d_c(U, V) = \max\{d(x_i, x_j)\}$	favour more compact clusters
average linkage	$d_a(U, V) = \text{average}\{d(x_i, x_j)\}$	compromise between extremes

$d(x, y) = d(y, x)$, $d(x, y) \leq d(x, z) + d(z, y)$, $d(x, y) = 0$ iff $x = y$
$d(U, V) \leq d(U, W) + d(U, V)$

② Model-based clustering : assume a function

Mixture model: $f(x) = \lambda_1 f_1(x; \theta_1) + \cdots + \lambda_k f_k(x; \theta_k)$
λ_i unknown, $\sum \lambda_i = 1$, θ_i unknown
Ideally $f_i(x) f_j(x) \approx 0$ for $i \neq j$

Example : $f_j(x) = \frac{1}{(2\pi)^{p/2} \det(C_j)^{1/2}} \cdot \exp\left(-\frac{1}{2}(x - \mu_j)^T C_j^{-1} (x - \mu_j)\right)$

k-means $\sum \min\{\|x_i - \mu_1\|^2, \cdots, \|x_i - \mu_k\|^2\}$ kmeans()
μ_i is the center of clusters

MLE estimation $\mathcal{L} = \sum_1^n \ln\{\sum_1^k \lambda_j f_j(x_i; \theta_j)\}$
EM-algorithm compute $E_{(\theta, \lambda)}(\Delta_{ij} | X_i = x_i) = \hat{\delta}_{ij}(\theta, \lambda)$
assign observation i to cluster j if $\hat{\delta}_{ij} > \hat{\delta}_{i\ell}$ for all $\ell \neq j$
Ⓔ $\hat{\delta}_{ij} = \frac{\hat{\lambda}_i f_j(x_i; \hat{\theta}_j)}{\sum_1^k \hat{\lambda}_\ell f_\ell(x_i; \hat{\theta}_\ell)}$ Ⓜ $\hat{\lambda}_j = \frac{1}{n} \sum \hat{\delta}_{ij}$ with $\{\hat{\theta}_j\}$ updated

Final Review Side B

2023年4月14日 8:35

Basics

$x_1 \cdots x_n$ observations. $x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{in} \end{bmatrix}$ p-dimension

- mean $\mu = E(X) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{pmatrix}$ $E(AX+b) = AE(X)+b$ $a^TCa \geq 0$ for $N_p(0,C)$
- covariance matrix $C = Cov(X) = E(X-\mu)(X-\mu)^T = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_p^2 \end{pmatrix}$ $\sigma_{jk} = Cov(X_j, X_k)$ $\sigma_j^2 = Var(X_j)$

  $Cov(AX+b) = ACA^T$

  C non singular $\Rightarrow (X-\mu)^TC^{-1}(X-\mu) \sim \chi^2(p)$ → assess multinormal
- Estimations: $\hat{\mu} = \bar{x} = \frac{1}{n}\sum x_i$ $\hat{C} = S = \frac{1}{n-1}\sum (x_i-\bar{x})(x_i-\bar{x})^T$
- correlation matrix $R = D^{-1/2}CD^{-1/2}$, $D = diag(C)$ corr( ) $Corr(X_1,X_2) = \frac{Cov(X_1,X_2)}{\sqrt{V(X_1)V(X_2)}}$
- concentration matrix: $K = C^{-1}$ solve(corr( ))

  K determines dependency $k_{ij} \neq 0 \Rightarrow$ dependence $V(X_1 \pm X_2) = V(X_1) + V(X_2) \pm 2Cov(X_1,X_2)$

  $\Rightarrow f(x) = \frac{1}{(2\pi)^{p/2}\det(C)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^TC^{-1}(x-\mu)\right)$

Multivariate Normal $X \sim N_p(\mu, C)$, $Y = AX+b \Rightarrow Y \sim N_r(A\mu+b, ACA^T)$

$X \sim N_p(\mu, C) \Rightarrow a^TX \sim N(a^T\mu, a^TCa)$

Conditional $X_1 | X_2 = x_2 \sim N_r(\mu_{1|2}, C_{1|2})$ $\boxed{\begin{aligned} \mu_{1|2} &= \mu_1 + C_{12}C_{22}^{-1}(x_2-\mu_2) \\ C_{1|2} &= C_{11} - C_{12}C_{22}^{-1}C_{21} \end{aligned}}$

- **Assess Normality**

  Mahalanobois distance: $d_i = (x_i-\bar{x})^TS^{-1}(x_i-\bar{x})$ if χ^2 then normal

  normal ⇒ straight line

  Shapiro-Wilk test: if normal $corr(x_i, z_i) \approx 1$, if not $corr(x_i, t_i) < 1$ H_0: is normal < 0.05 reject

  Kurtosis $K(X) = \frac{E((X-\mu)^4)}{Var(X)^2}$ $K(X) \geq 1$, $K(aX+b) = K(X)$

  $X \sim N(\mu, \sigma^2) \Rightarrow K(X) = 3$

  $|K(X) - 3|$ measures non-normality

Scatterplot matrix pairs( )

p increase, effectiveness decrease.

Histogram hist( )

Optimal bandwidth ① $3.49 \times SD \times n^{-1/3}$ ② $2 \times IQR \times n^{-1/3}$

curse of dimension - $x_1 \cdots x_n$ in $\mathbb{R}^p$

as p increase, points become sparse.

Bi-plot: PC1 vs PC2

vector indicate how x_i correlated with first two PCs

$\cos(\theta) = corr(x_i, x_j)$

AIC = $-2 \times$ max loglikelyhood $+ 2 \times$ number of parameter

BIC = $-2 \times$ max loglikelyhood $+ \ln(2) \times$ number of parameter

choose model that minimize AIC and BIC

Generalized Linear Model

Take $H(x)$ cts. $0 < H(x) < 1$. $H(x)$ strictly increasing

Define $P(G=1|X=x) = H(\beta_0 + x^T\beta)$

$P(G=0|X=x) = 1 - H(\beta_0 + x^T\beta)$ $\ln\left(\frac{P(G=1|X=x)}{P(G=0|X=x)}\right) = \beta_0 + x^T\beta$

Then $\underbrace{H^{-1}}_{\text{link function}}(P(G=1|X=x)) = \beta_0 + x^T\beta$

Link example Logit $H^{-1}(x) = \ln(x/(1-x))$

Probit $H^{-1}(x) = \Phi^{-1}(x)$ inverse normal

Log-log $H^{-1}(x) = -\ln(-\ln(x))$

MLE : $\mathcal{L}(\beta_0, \beta) = \sum\{g_i(\beta_0 + x_i^T\beta) - \ln(1+\exp(\beta_0 + x_i^T\beta))\}$

If satisfy $\sum\left\{g_i - \frac{\exp(\beta_0 + x_i^T\beta)}{1+\exp(\beta_0 + x_i^T\beta)}\right\} = 0$ $(\sum \times x_i = 0)$

- LDA k=2 is logistic regression

  LDA k=2, $P(G=1|X=x) = \frac{\lambda_1 f_1/\lambda_0 f_0}{(1+\lambda_1 f_1)/\lambda_0 f_0}$ $\frac{\lambda_1 f_1}{\lambda_0 f_0} = \exp(\beta_0 + x^T\beta)$

  $\beta = C^{-1}(\mu_1 - \mu_0)$ $\beta_0 = \frac{1}{2}(\mu_0^TC^{-1}\mu_0 - \mu_1^TC^{-1}\mu_1) + \ln(\lambda_1/\lambda_0)$

  LDA estimate mean covariance

  Logistic regression use maximum likelyhood

Supervised Learning

Classification

Given data, find a rule $\phi: \mathcal{X} \to \mathcal{G}$

$\mathcal{X}$ is range of possible x, $\mathcal{G}$ is set of all possible classes

① **Logistic Regression** model $P(G=g|X=x) = \psi_g(x;\beta)$ glm( )

$P(G=1|X=x) = \frac{\exp(x^T\beta)}{1+\exp(x^T\beta)}$ $G = \{0,1\}$

$P(G=g|X=x) = \frac{\exp(x^T\beta_g)}{1+\exp(x^T\beta_1)+\cdots+\exp(x^T\beta_k)}$ $G = \{0,1,\cdots,k\}$

Optimal classification rule $R_g = \{x: \lambda_g f_g(x) > \lambda_j f_j(x)\ \forall j \neq g\}$

$\phi(x) = \arg\max_g\{\lambda_g f_g(x): g = 1, \cdots, k\}$

$(= g$ if $\lambda_g f_g > \lambda_j f_j$ for all $j \neq g)$

Posterior distribution $P(G=g|X=x) = \frac{\lambda_g f_g(x)}{\lambda_1 f_1 + \cdots + \lambda_k f_k}$

- Application: multivariate normal

  $\phi(x) = g$ if $\ln\left(\frac{f_g(x)}{f_j(x)}\right) > \ln\left(\frac{\lambda_j}{\lambda_g}\right)$ for all $j \neq g$.

  $\ln\left(\frac{f_g(x)}{f_j(x)}\right) = \{x^TC^{-1}\mu_g - \frac{1}{2}\mu_g^TC^{-1}\mu_g\} - \{x^TC^{-1}\mu_j - \frac{1}{2}\mu_j^TC^{-1}\mu_j\}$

② **Discriminant Analysis**

- LDA linear discriminant lda( ) class: predicted class. posterior: possibility estimate

  $\hat{\mu}_g = \{\sum I(g_i=g)\}^{-1}\sum x_i I(g_i=g)$ $\hat{C} = \frac{1}{n-k}\sum(x_i - \hat{\mu}_{g_i})(x_i - \hat{\mu}_{g_i})^T$

  $\hat{\lambda}_g = \frac{1}{n}\sum I(g_i=g)$ $\hat{\phi}(x) = \arg\max_g \hat{d}_g(x)$

  discriminant scores: $\hat{d}_g(x) = x^T\hat{C}^{-1}\hat{\mu}_g - \frac{1}{2}\hat{\mu}_g^T\hat{C}^{-1}\hat{\mu}_g + \ln(\hat{\lambda}_g)$

  LDA assumes $C_1 = C_2 \cdots = C_k = C$
- QDA quadratic discriminant qda( )

  $X|G=g \sim N_p(\mu_g, C_g)$ $\phi(x) = \arg\max_g d_g(x)$

  discriminant score: $d_g(x) = \ln(\lambda_g) - \frac{1}{2}(x-\mu_g)^TC_g^{-1}(x-\mu_g) - \frac{1}{2}\ln(\det(C_g))$

  QDA assumes $C_1 \cdots C_k$ are arbitrary
- QDA allows for flexible boundries

  QDA estimate more parameters, variance of $\hat{C}_1 \cdots \hat{C}_k > \hat{C}$ for LDA

  QDA typically has lower bias due to flexibility

  note: bias-variance trade-off
- Cross Validation - leave out m obs - training set - estimate rule

  - remain n-m obs - test set - estimate error rate

③ Tree-based classification, use half spaces

Example $B_j = [x_2 < 5] \times [x_1 > 3]$

Advantage: no assumptions - flexibility - any

Disadvantage: complex, depends on stopping rule

MANOVA - assess feasibility of LDA, $H_0: \mu_1 = \mu_2 = \cdots \mu_k$ manova( )

$S_T = \underbrace{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T}_{S_B} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T}_{S_W}$

high F → reject H_0, low F → fail to reject H_0

Test statistics: Wilks Lambda / Pillai

Hotelling-Lawley trace / Roy's maximal root