Discussion 03

2022.03.10

杨子健 yangzi@shanghaitech.edu.cn

p : number of features

sample x_1^T

X_0	X_1	X_2	…	X_{256}
1	0.156	0.432	⋯	0.824
1	0.671	0.014	⋯	0.969
⋯	⋯	⋯	⋯	⋯
1	0.523	0.142	⋯	0.718

N: number of instances

All-one vector for the intercept

Data matrix $\mathbf{X}$

G
0
1
⋯
9

One-hot coding

Y_1	Y_2	…	Y_{10}
1	0	⋯	0
0	1	⋯	0
⋯	⋯	⋯	⋯
0	0	⋯	1

N: number of instances

Coded output matrix $\mathbf{Y}$

Q $\min_{\mathbf{B}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 \longrightarrow \hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$

$$\|A\|_F^2 = tr(A^T A)$$

$$\|Y - XB\|_F^2 = tr[(Y-XB)^T(Y-XB)]$$

$$= tr[Y^TY - Y^TXB - B^TX^TY + B^TX^TXB]$$

$$tr(A) = tr(A^T)$$

$$= tr(Y^TY) + tr(B^TX^TXB) - 2tr(Y^TXB)$$

$$\nabla_B \|Y - XB\|_F^2 = \nabla_B tr(B^TX^TXB) - \nabla_B 2tr(Y^TXB)$$

$$\nabla_X tr(X^TAX) = (A + A^T)X \qquad \nabla_X tr(AXC) = A^TC^T$$

$$\nabla_B \|Y - XB\|_F^2 = (X^TX + X^TX)B - 2X^TY \stackrel{令}{=} 0$$

$$\Rightarrow \hat{B} = (X^TX)^{-1}X^TY$$

Binary classification

- Linear regression

$$f(x) = \beta_0 + x^T\beta$$

- Decision boundary

$$\{x : x^T\hat{\beta} = threshold\}$$

  - $threshold = 0$, if $y \in \{-1,1\}$
  - $threshold = 0.5$, if $y \in \{0,1\}$

Multi-class classification

- Linear regressions for K classes

$$f_k(x) = \beta_{k0} + x^T\beta_k, \qquad k = 1, \dots, K$$

- Decision boundary between classes k and ℓ:

$$\{x: \hat{f}_k(x) = \hat{f}_\ell(x)\}$$

- That is an affine set or hyperplane:

$$\{x: (\hat{\beta}_{k0} - \hat{\beta}_{\ell 0}) + x^T(\hat{\beta}_k - \hat{\beta}_\ell) = 0\}$$

Linear Regression of an Indicator Matrix

- Indicator response matrix

$$\mathcal{G} = \overbrace{\{0,1,2 \ldots, 9\}}^{K}$$

G
0
1
···
9

N

coding

K

Y_1	Y_2	···	Y_{10}
1	0	···	0
0	1	···	0
···	···	···	···
0	0	···	1

Indicator response matrix $\mathbf{Y} \in \mathbb{R}^{N \times K}$

- Our problem:

$$\hat{\mathbf{B}} = \underset{\mathbf{B}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2$$

$$\mathbf{B} = (\beta_1, \beta_2, \ldots, \beta_{10}) \in \mathbb{R}^{(p+1) \times K}$$

- The fitted values on $\mathbf{X}$:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

A new observation x is classified by

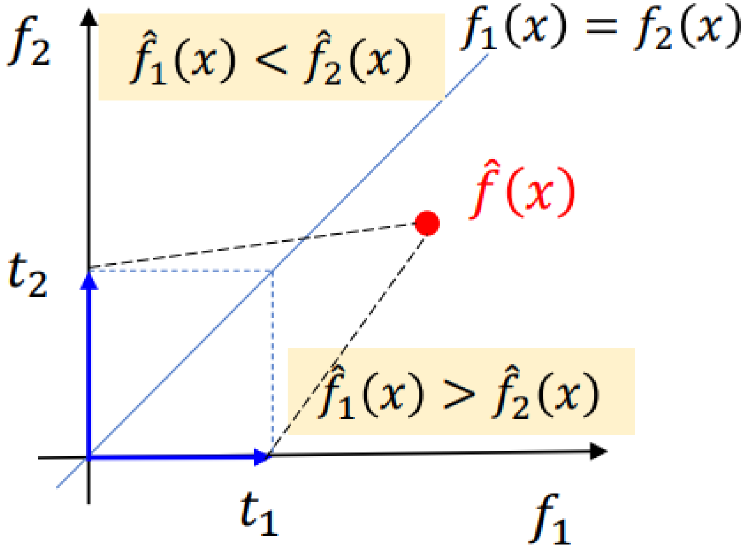
- Compute the fitted output

Q

$$\hat{f}(x) = \hat{\mathbf{B}}^T \binom{1}{x} = \begin{pmatrix} \hat{f}_1(x) \\ \hat{f}_2(x) \\ \vdots \\ \hat{f}_K(x) \end{pmatrix} \in \mathbb{R}^K$$

- Classify x according to

$$\hat{G}(x) = \underset{k \in \mathcal{G}}{\operatorname{argmax}}\, \hat{f}_k(x)$$



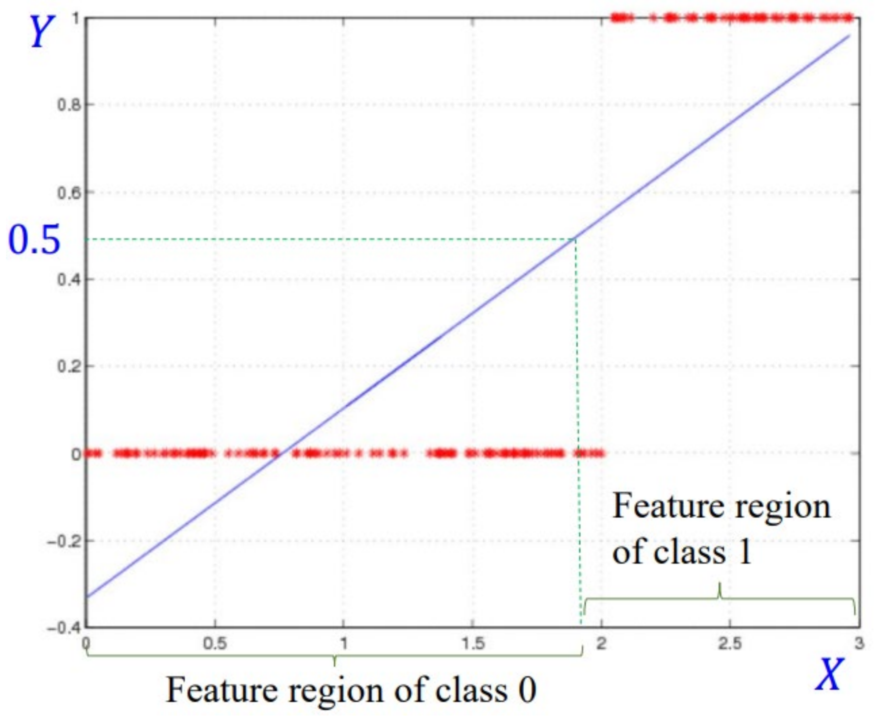
$$\hat{Y} = X\hat{B}$$

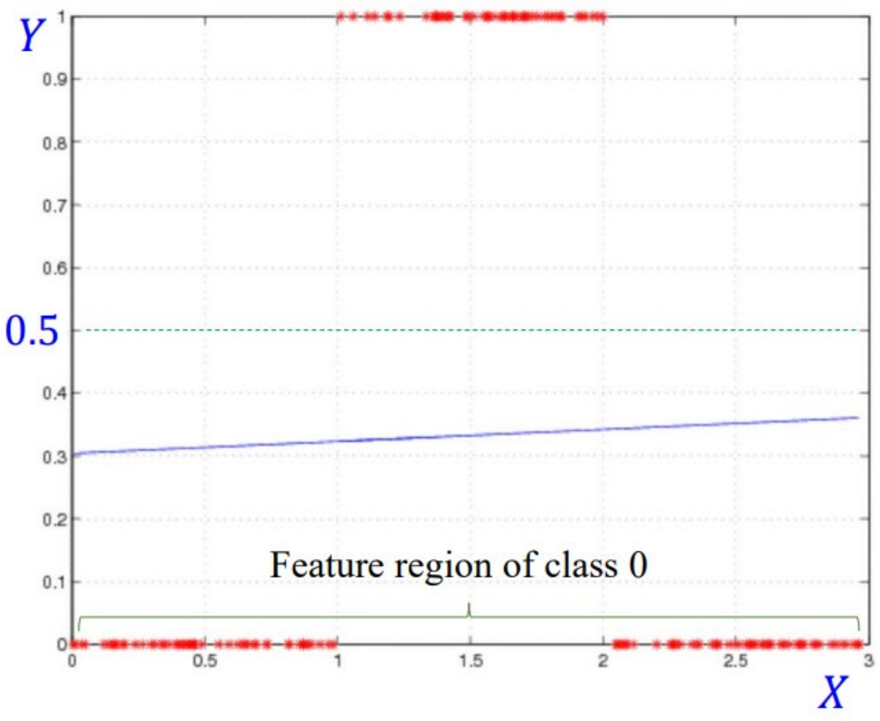
$$y_i^T \quad N \; \overset{K}{\boxed{Y}} = x_i^T \quad N \; \overset{p+1}{\boxed{X}} \cdot \quad p+1 \; \overset{K}{\boxed{\hat{B}}}$$

$$y_i^T = x_i^T \cdot \hat{B}$$

$$y_i = \hat{B}^T \cdot x_i = f(x_i)$$

The Phenomenon of Masking





Linear Discriminant Analysis

- Example: binary (two class) classification

Logit: $\log \frac{\Pr(G=1|X=x)}{1-\Pr(G=1|X=x)} = \log \frac{\Pr(G=1|X=x)}{\Pr(G=2|X=x)} = \beta_0 + x^T\beta$

$$\log \frac{\overset{p_1}{P(G=1|X=x)}}{\underset{p_2}{P(G=2|X=x)}} = \underbrace{\beta_0 + x^T\beta}_{\text{记为}t}$$

$$\log \frac{p_1}{p_2} = t \Rightarrow \frac{p_1}{p_2} = e^t$$

$$p_1 + p_2 = (e^t + 1)p_2 = 1$$

$$\Rightarrow p_2 = \frac{1}{e^t+1}, \quad p_1 = \frac{e^t}{e^t+1}$$

- The posterior probability

Q

$$\Pr(G=1|X=x) = \frac{\boxed{\exp(\beta_0 + x^T\beta)}}{1+\exp(\beta_0 + x^T\beta)},$$

$$\Pr(G=2|X=x) = \frac{1}{1+\exp(\beta_0 + x^T\beta)}$$

- Decision boundary

$$\{x|\beta_0 + x^T\beta = 0\}$$

- Posterior

$$\Pr(G=k|X=x)=\frac{\Pr(X=x|G=k)\Pr(G=k)}{\Pr(X=x)}=\frac{\boxed{\Pr(X=x|G=k)}\boxed{\Pr(G=k)}}{\sum_{\ell=1}^{K}\Pr(X=x|G=\ell)\Pr(G=\ell)}$$

  - Density of X in class $G=k$:

$$f_k(x)=\Pr(X=x|G=k)$$

  - Class prior:

$$\pi_k=\Pr(G=k)$$

$$\Pr(G=k|X=x)=\frac{f_k(x)\pi_k}{\sum_{\ell=1}^{K}f_\ell(x)\pi_\ell}$$

- Assumptions in LDA
  1. Model each class density as multivariate Gaussian

$$f_k(x)=\frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right)$$

  2. Assume that classes share a common covariance $\Sigma_k=\Sigma,\forall k$

- Compare two classes k and ℓ

Logit: $$\log\frac{\Pr(G=k|X=x)}{\Pr(G=\ell|X=x)}=\log\frac{f_k(x)}{f_\ell(x)}+\log\frac{\pi_k}{\pi_\ell}$$

$$=\boxed{\log\frac{\pi_k}{\pi_\ell}-\frac{1}{2}(\mu_k+\mu_\ell)^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_\ell)+x^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_\ell),}$$

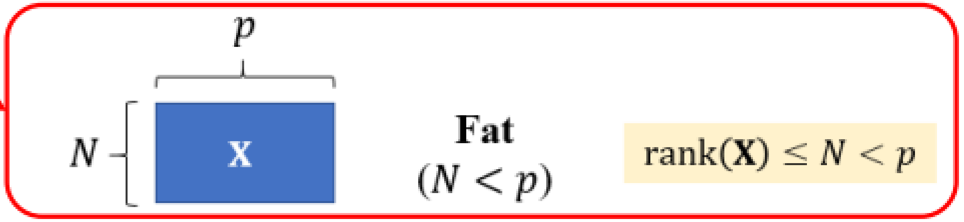
Quadratic term vanished due to the common covariance

Quadratic Discriminant Analysis

- Assumption: Each class has a specific covariance Σ_k

- Difference with LDA
  - Σ_k has to be estimated for each class
  - LDA need to estimate $\underbrace{K \times p}_{\mu_k,\, k=1,\dots,K} + \underbrace{p \times p}_{\Sigma}$ parameters
  - QDA need to estimate $K \times p + \underbrace{K \times p \times p}_{\Sigma_k,\, k=1,\dots,K}$ parameters

Regularized Discriminant Analysis



High dimensional problems ($p \gg N$)

- Regularization is necessary
  - No enough data to estimate feature dependencies
  - E.g., independent assumption on features
    - Diagonal within-class covariance matrix
      #paras: $K \times p \times p \rightarrow K \times p$

Regularized LDA (RLDA)

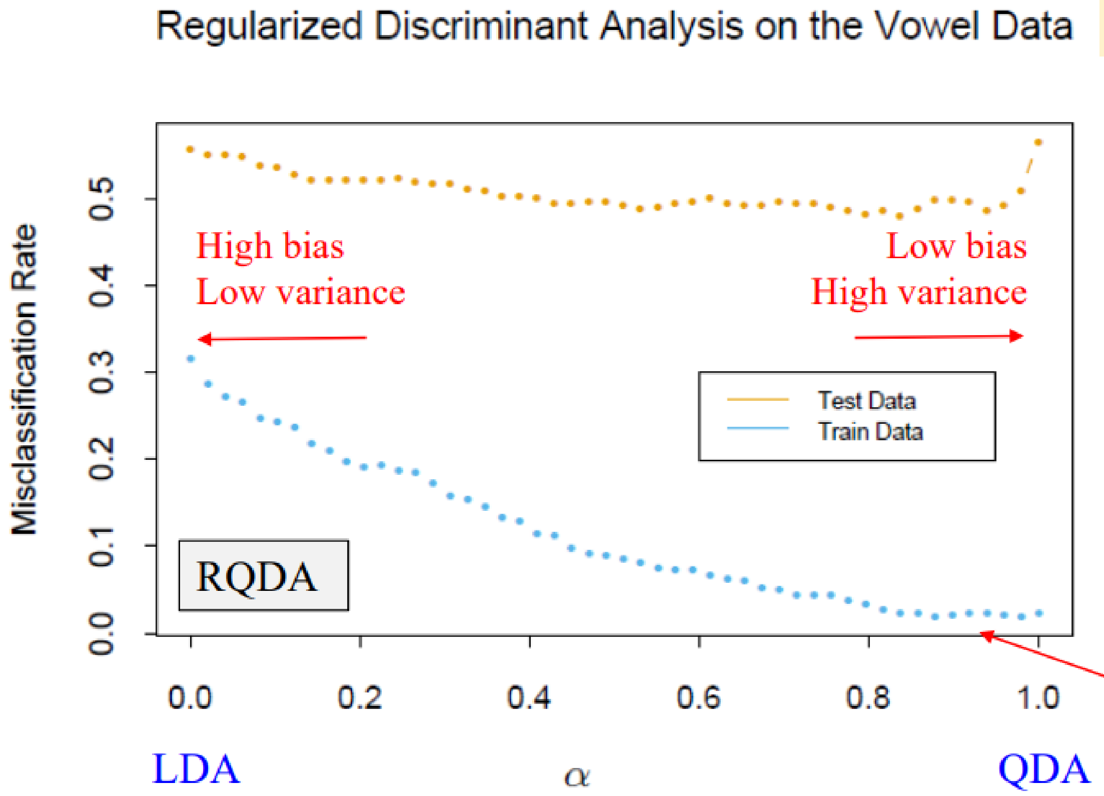
- Shrinks $\hat{\Sigma}$ towards its diagonal

$$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\mathrm{diag}(\hat{\Sigma}), \gamma \in [0,1]$$

where $\mathrm{diag}(\hat{\Sigma})$ denotes a diagonal matrix sharing the same diagonal elements with $\hat{\Sigma}$

Regularized Discriminant Analysis

https://hastie.su.domains/ElemStatLearn/



RQDA:

$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$

- $\alpha = 0$, LDA
- $\alpha = 1$, QDA

The optimal model
A compromise between QDA and LDA

FIGURE 4.7. *Test and training errors for the vowel data, using regularized discriminant analysis with a series of values of* $\alpha \in [0, 1]$. *The optimum for the test data occurs around* $\boxed{\alpha = 0.9,}$ *close to quadratic discriminant analysis.*

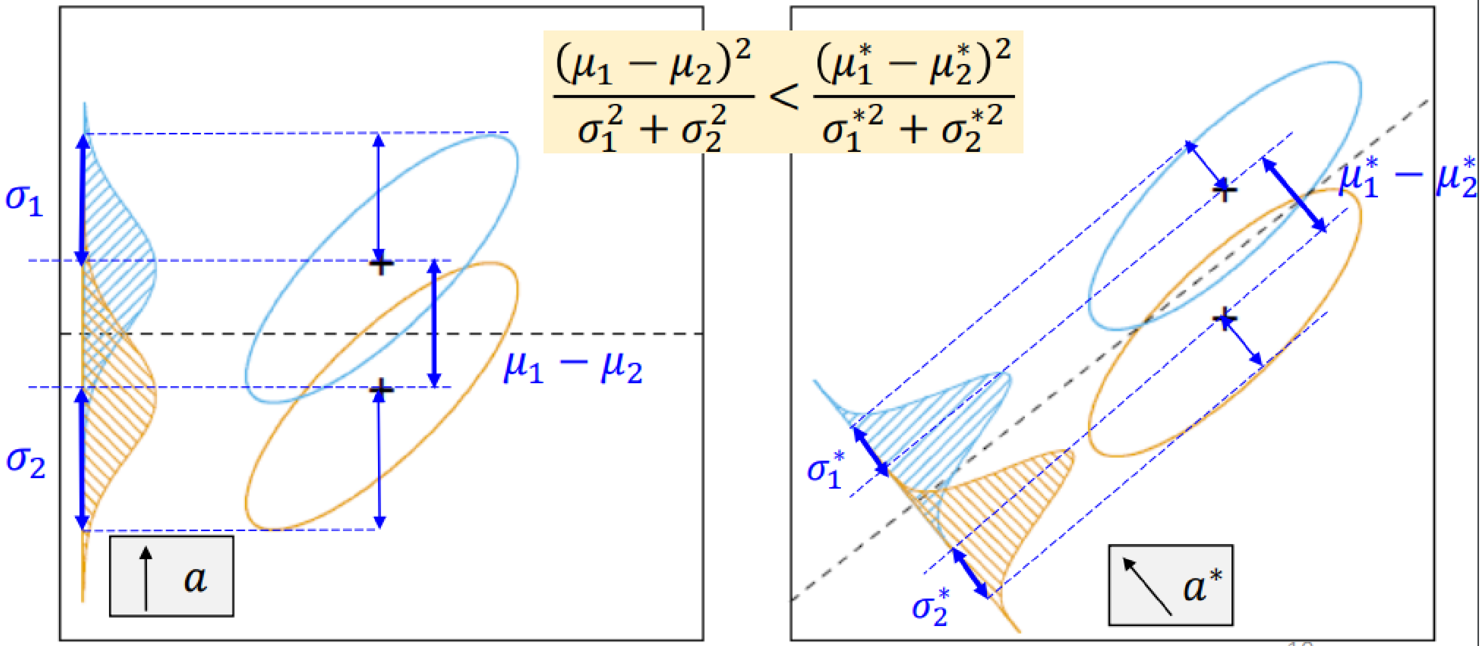
Fisher's Formulation of Discriminant Analysis

- Find $z = x^T a$ such that the between class variance is maximized relative to the within class variance.

Between class variance

Our goal: $\max\limits_{a} \dfrac{\boxed{(\mu_1^* - \mu_2^*)^2}}{\boxed{\sigma_1^{*2} + \sigma_2^{*2}}}$

Within class variance



THANKS