Optimization and Machine Learning SI151

Lu Sun

School of Information Science and Technology

ShanghaiTech University

March 18, 2021

Today:
- Linear Methods for Classification II
  - Generalization of LDA
  - Logistic Regression
  - Summary

Readings:
- The Elements of Statistical Learning (ESL), Chapters 4.3, 4.4, 18.1, 18.2 and 18.3

Linear Methods for Classification II

- Generalization of LDA
    - Regularized Discriminant Analysis
    - Fisher's Formulation of Discriminant Analysis
- Logistic Regression
- Summary

Regularized Discriminant Analysis
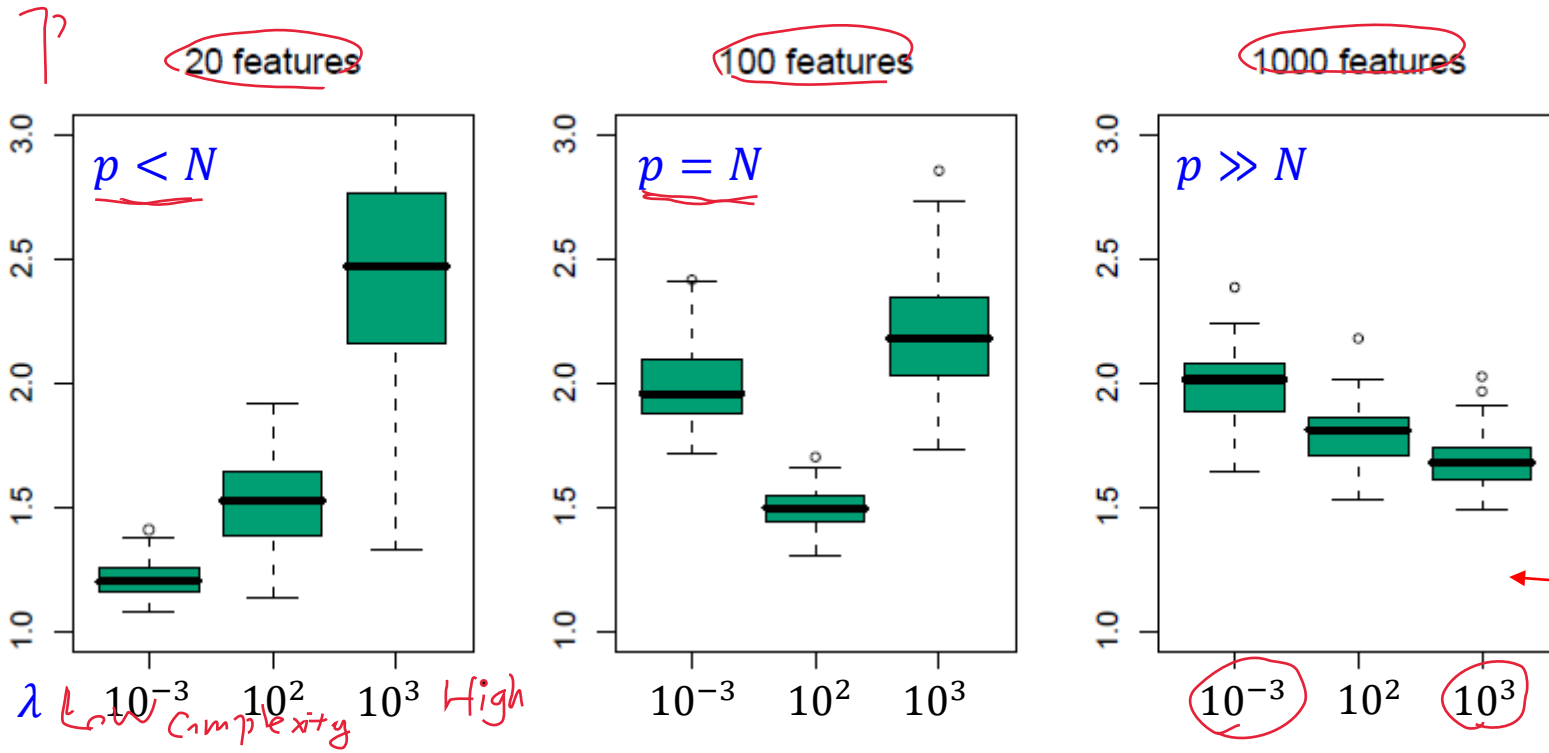
High dimensional problems ($p \gg N$)
- genomics problem, signal/image analysis
- Less fitting is better

$$\min_{\beta} \| y - X\beta \|_2^2 + \lambda \| \beta \|_2^2$$

Example $N = 100$
- 100 samples are generated by a linear model
- Ridge regression
- Relative error (divide by Bayes error)

λ Low Complexity High

No enough information to estimate the high-dimensional covariance matrix

3

Regularized Discriminant Analysis

$$\Sigma = \begin{matrix} \text{features} \downarrow i \end{matrix} \boxed{\begin{matrix} \rightarrow \text{features} \\ \quad j \end{matrix}} \rightarrow G_{ij} \rightarrow x_i^T x_j \equiv \langle x_i, x_j \rangle$$

$$\Sigma = \frac{1}{n} X^T X \quad \begin{cases} = 0, & \vartheta \neq j \\ 1, & i = j \end{cases}$$

High dimensional problems ($p \gg N$)

$\delta_k(x)$

- Cannot fit LDA to the data
  - inversion of a $p \times p$ covariance matrix Σ
  - Σ is singular, due to $\text{rank}(\Sigma) < N \ll p$

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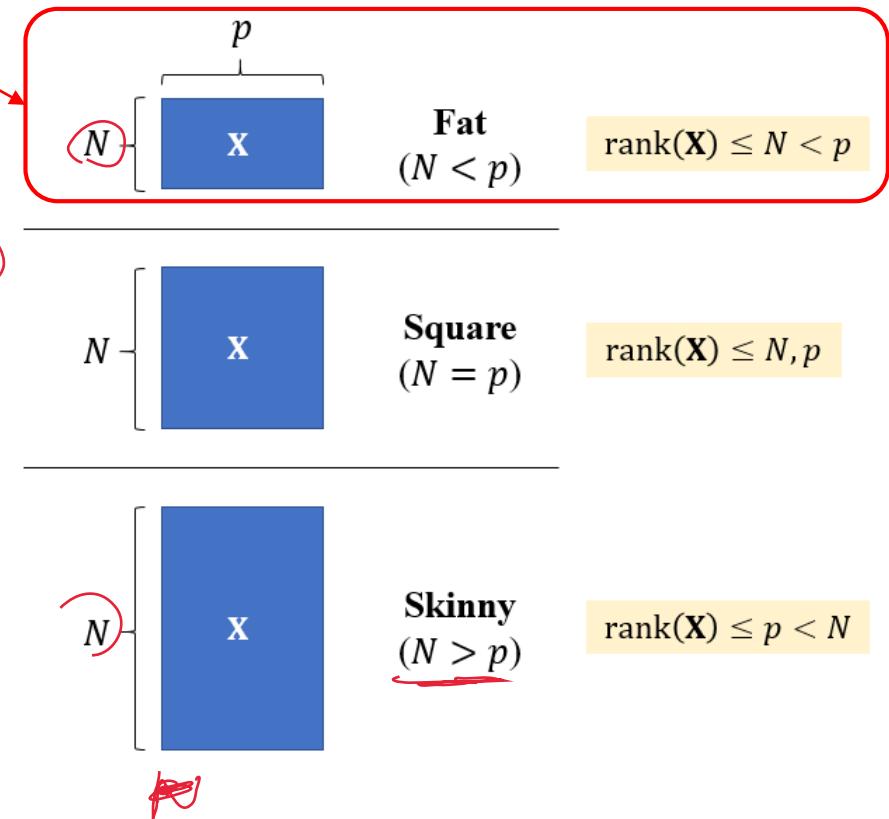
$\Sigma = diag(\sigma)$.
$p \times p$ p

N	X	**Fat** $(N < p)$ $\text{rank}(\mathbf{X}) \leq N < p$
N	X	**Square** $(N = p)$ $\text{rank}(\mathbf{X}) \leq N, p$
N	X	**Skinny** $(N > p)$ $\text{rank}(\mathbf{X}) \leq p < N$

Model complexity

Regularized Discriminant Analysis

Regularized LDA (RLDA)

- Shrinks $\widehat{\Sigma}$ towards its diagonal
$$\widehat{\Sigma}(\gamma) = \gamma\widehat{\Sigma} + (1 - \gamma)\text{diag}(\widehat{\Sigma}), \gamma \in [0, 1]$$
where $\text{diag}(\widehat{\Sigma})$ denotes a diagonal matrix sharing the same diagonal elements with $\widehat{\Sigma}$

Diagonal LDA

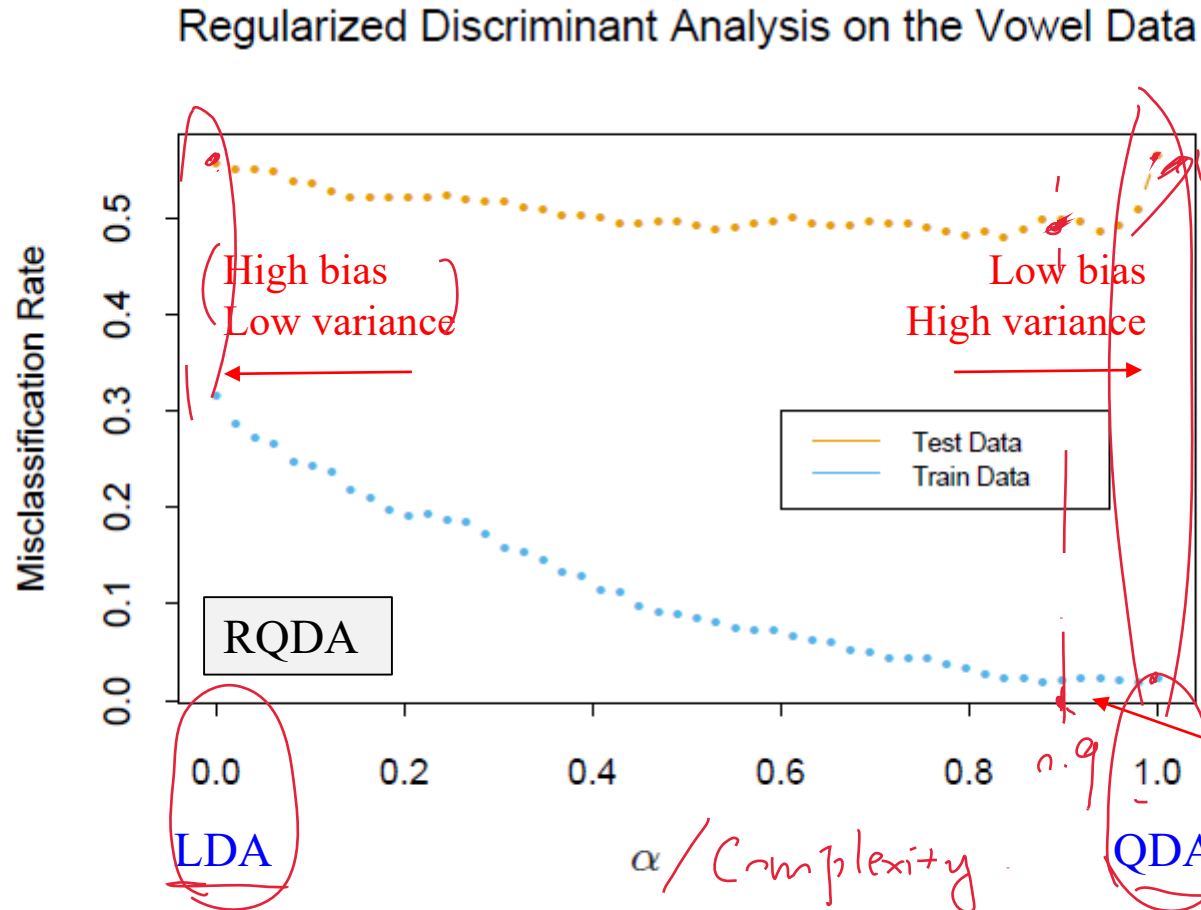- Independent assumption on feature dependencies
$$\widehat{\Sigma} = \text{diag}(\widehat{\Sigma})$$

Regularized Discriminant Analysis

A brief summary of generalized LDA ($\alpha, \gamma \in [0, 1]$)

	Method	Covariance matrix	Effect
Linear	Regularized LDA (RLDA)	$\widehat{\boldsymbol{\Sigma}}(\gamma) = \gamma\widehat{\boldsymbol{\Sigma}} + (1 - \gamma)\text{diag}(\widehat{\boldsymbol{\Sigma}})$	Shrink $\widehat{\boldsymbol{\Sigma}}$ towards $\text{diag}(\widehat{\boldsymbol{\Sigma}})$
	Diagonal LDA	$\widehat{\boldsymbol{\Sigma}} = \text{diag}(\widehat{\boldsymbol{\Sigma}})$	Make features independent
Quadratic	Regularized QDA (RQDA)	$\widehat{\boldsymbol{\Sigma}}_k(\alpha) = \alpha\widehat{\boldsymbol{\Sigma}}_k + (1 - \alpha)\widehat{\boldsymbol{\Sigma}}$	Shrink $\widehat{\boldsymbol{\Sigma}}_k$ towards $\widehat{\boldsymbol{\Sigma}}$ (LDA + QDA)
	Variant of RQDA	$\widehat{\boldsymbol{\Sigma}}_k(\alpha, \gamma) = \alpha\widehat{\boldsymbol{\Sigma}}_k + (1 - \alpha)\widehat{\boldsymbol{\Sigma}}(\gamma)$	Shrink $\widehat{\boldsymbol{\Sigma}}_k$ towards $\widehat{\boldsymbol{\Sigma}}(\gamma)$ (RLDA + QDA)

Regularized Discriminant Analysis

Regularized Discriminant Analysis on the Vowel Data

High bias
Low variance

Low bias
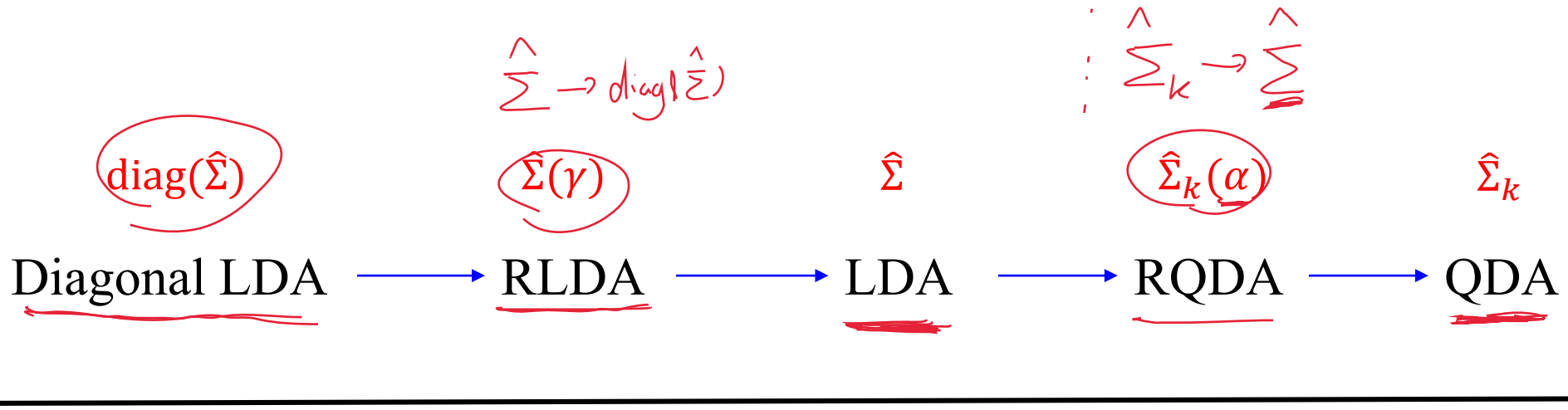High variance

RQDA

LDA

QDA

α / Complexity

0.9

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* $\alpha = 0.9$, *close to quadratic discriminant analysis.*

7

Regularized Discriminant Analysis

$$\hat{\Sigma}_k \rightarrow \hat{\Sigma}(\gamma)$$

$$\hat{\Sigma} \rightarrow diag(\hat{\Sigma})$$

$$\hat{\Sigma}_k \rightarrow \hat{\Sigma}$$

$diag(\hat{\Sigma})$ $\hat{\Sigma}(\gamma)$ $\hat{\Sigma}$ $\hat{\Sigma}_k(\alpha)$ $\hat{\Sigma}_k$

Diagonal LDA $\longrightarrow$ RLDA $\longrightarrow$ LDA $\longrightarrow$ RQDA $\longrightarrow$ QDA

High bias
Low variance

Low bias
High variance

(Low complexity)

(High Complexity)

8

Fisher's Formulation of Discriminant Analysis

LDA: Approach 1

1. Estimating $\widehat{\boldsymbol{\Sigma}}$, $\hat{\mu}_k$ and $\hat{\pi}_k$

2. Discriminant function
$$\delta_k(x) = x^T \widehat{\boldsymbol{\Sigma}}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \widehat{\boldsymbol{\Sigma}}^{-1} \hat{\mu}_k + \log \hat{\pi}_k$$

3. Classify to class k that maximizes the discriminant function
$$\hat{G}(x) = \underset{k \in \mathcal{G}}{\operatorname{argmax}} \, \delta_k(x)$$

$$\widehat{\Sigma} = D^{-\frac{1}{2}} U^T \qquad \left(\widehat{\Sigma}\right)^{-\frac{1}{2}T} = U \cdot D^{-\frac{1}{2}}$$

$\boldsymbol{Q}$: why data sphering makes $\widehat{\boldsymbol{\Sigma}}^* = \mathbf{I}$?

Hint: $\widehat{\boldsymbol{\Sigma}} = \dfrac{\sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T}{N - K}$

$$I = \Sigma^{-\frac{1}{2}} \Sigma \left(\Sigma^{-\frac{1}{2}}\right)^T$$

LDA: Approach 2

1. Estimating $\widehat{\boldsymbol{\Sigma}}$, $\hat{\mu}_k$ and $\hat{\pi}_k$

2. Eigen-decomposition:
$$\widehat{\boldsymbol{\Sigma}} = \mathbf{U}\mathbf{D}\mathbf{U}^T$$

$$\widehat{\Sigma}^{\frac{1}{2}} = U D^{\frac{1}{2}}$$

3. Data sphering ($\widehat{\Sigma}^* = \mathbf{I}$)
- $x^* = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T x = \widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} x$
- $\hat{\mu}_k^* = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T \hat{\mu}_k = \widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \hat{\mu}_k$

4. Classify to its closest class centroid in the transformed space
$$G(x) = \underset{k \in \mathcal{G}}{\operatorname{argmin}} \frac{1}{2} \|x^* - \hat{\mu}_k^*\|^2 - \ln \hat{\pi}_k$$

$$x_o \rightarrow \Sigma^{-\frac{1}{2}} x_o \qquad x_o^*$$

Fisher's Formulation of Discriminant Analysis

LDA: Approach 1

1. Estimating $\widehat{\boldsymbol{\Sigma}}$, $\hat{\mu}_k$ and $\hat{\pi}_k$

2. Discriminant function
$$\delta_k(x) = x^T \widehat{\boldsymbol{\Sigma}}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \widehat{\boldsymbol{\Sigma}}^{-1} \hat{\mu}_k + \log \hat{\pi}_k$$

3. Classify to class k that maximizes the discriminant function
$$\hat{G}(x) = \underset{k \in \mathcal{G}}{\operatorname{argmax}} \, \delta_k(x)$$

LDA: Approach 2

1. Estimating $\widehat{\boldsymbol{\Sigma}}$, $\hat{\mu}_k$ and $\hat{\pi}_k$

2. Eigen-decomposition:
$$\widehat{\boldsymbol{\Sigma}} = \mathbf{U}\mathbf{D}\mathbf{U}^T$$

3. Data sphering ($\hat{\Sigma}^* = \mathbf{I}$)
   - $x^* = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T x = \widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} x$
   - $\hat{\mu}_k^* = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T \hat{\mu}_k = \widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \hat{\mu}_k$

4. Classify to its closest class centroid in the transformed space
$$\hat{G}(x) = \underset{k \in \mathcal{G}}{\operatorname{argmin}} \frac{1}{2} \|x^* - \hat{\mu}_k^*\|^2 - \ln \hat{\pi}_k$$

Fisher's Formulation of Discriminant Analysis

1. $\log \frac{\Pr(G=k|X=x)}{\Pr(G=\ell|X=x)} = \delta_k(x) - \delta_\ell(x)$

$\mathcal{N}(\hat{\mu}_k, \widehat{\Sigma})$ $\hat{\pi}_k$

2. $\delta_k(x) \propto \log \Pr(G = k|X = x)$ $\Pr(G = k|X = x) = \dfrac{\Pr(X = x|G = k)\,\Pr(G = k)}{\Pr(X = x)}$

3. $\log \Pr(G = k|X = x) = -\frac{1}{2}(x - \hat{\mu}_k)^T \widehat{\Sigma}^{-1}(x - \hat{\mu}_k) + \log \hat{\pi}_k + \boxed{C}$ Constant

$\qquad = -\frac{1}{2}(x - \hat{\mu}_k)^T \mathbf{U}\mathbf{D}^{-\frac{1}{2}}\left(\mathbf{U}\mathbf{D}^{-\frac{1}{2}}\right)^T (x - \hat{\mu}_k) + \log \hat{\pi}_k + C$

$\qquad = -\frac{1}{2}\left(\mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T x - \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T \hat{\mu}_k\right)^T \left(\mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T x - \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T \hat{\mu}_k\right) + \log \hat{\pi}_k + C$

$\qquad = -\frac{1}{2}(x^* - \hat{\mu}_k^*)^T (x^* - \hat{\mu}_k^*) + \log \hat{\pi}_k + C$

$\qquad = -\frac{1}{2}\|x^* - \hat{\mu}_k^*\|^2 + \ln \hat{\pi}_k + C$

4. $\hat{G}(x) = \underset{k \in \mathcal{G}}{\mathrm{argmax}}\, \delta_k(x) = \underset{k \in \mathcal{G}}{\mathrm{argmax}}\, \log \Pr(G = k|X = x) = \underset{k \in \mathcal{G}}{\mathrm{argmin}}\, \frac{1}{2}\|x^* - \hat{\mu}_k^*\|^2 - \ln \hat{\pi}_k$

Fisher's Formulation of Discriminant Analysis

LDA: Approach 1

1. Estimating $\widehat{\boldsymbol{\Sigma}}$, $\hat{\mu}_k$ and $\hat{\pi}_k$

2. Discriminant function
$$\delta_k(x) = x^T \textcolor{red}{\widehat{\boldsymbol{\Sigma}}^{-1}} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \textcolor{red}{\widehat{\boldsymbol{\Sigma}}^{-1}} \hat{\mu}_k + \log \hat{\pi}_k$$

3. Classify to class k that maximizes the discriminant function
$$\hat{G}(x) = \underset{k \in \mathcal{G}}{\text{argmax}}\, \delta_k(x)$$

> Complexity
> $\mathcal{O}(p^3)$

- Two approaches have almost the same time and storage complexity
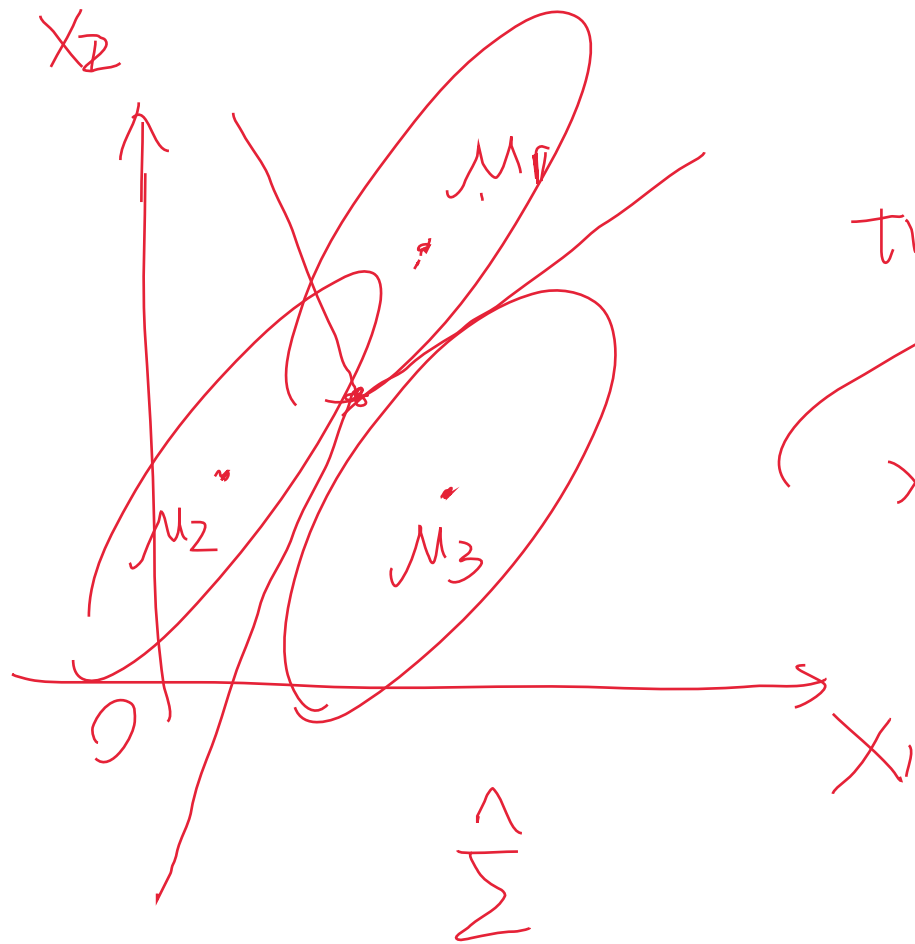- Approach 2 shows the potential of LDA for dimension reduction

$(\#classes = k)$

$LDA \Rightarrow (k-1)$

LDA: Approach 2

1. Estimating $\widehat{\boldsymbol{\Sigma}}$, $\hat{\mu}_k$ and $\hat{\pi}_k$

2. Eigen-decomposition:
$$\textcolor{red}{\widehat{\boldsymbol{\Sigma}} = \mathbf{U}\mathbf{D}\mathbf{U}^T}$$

3. Data sphering ($\hat{\Sigma}^* = \mathbf{I}$)
   - $x^* = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T x = \widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} x$
   - $\hat{\mu}_k^* = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T \hat{\mu}_k = \widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \hat{\mu}_k$

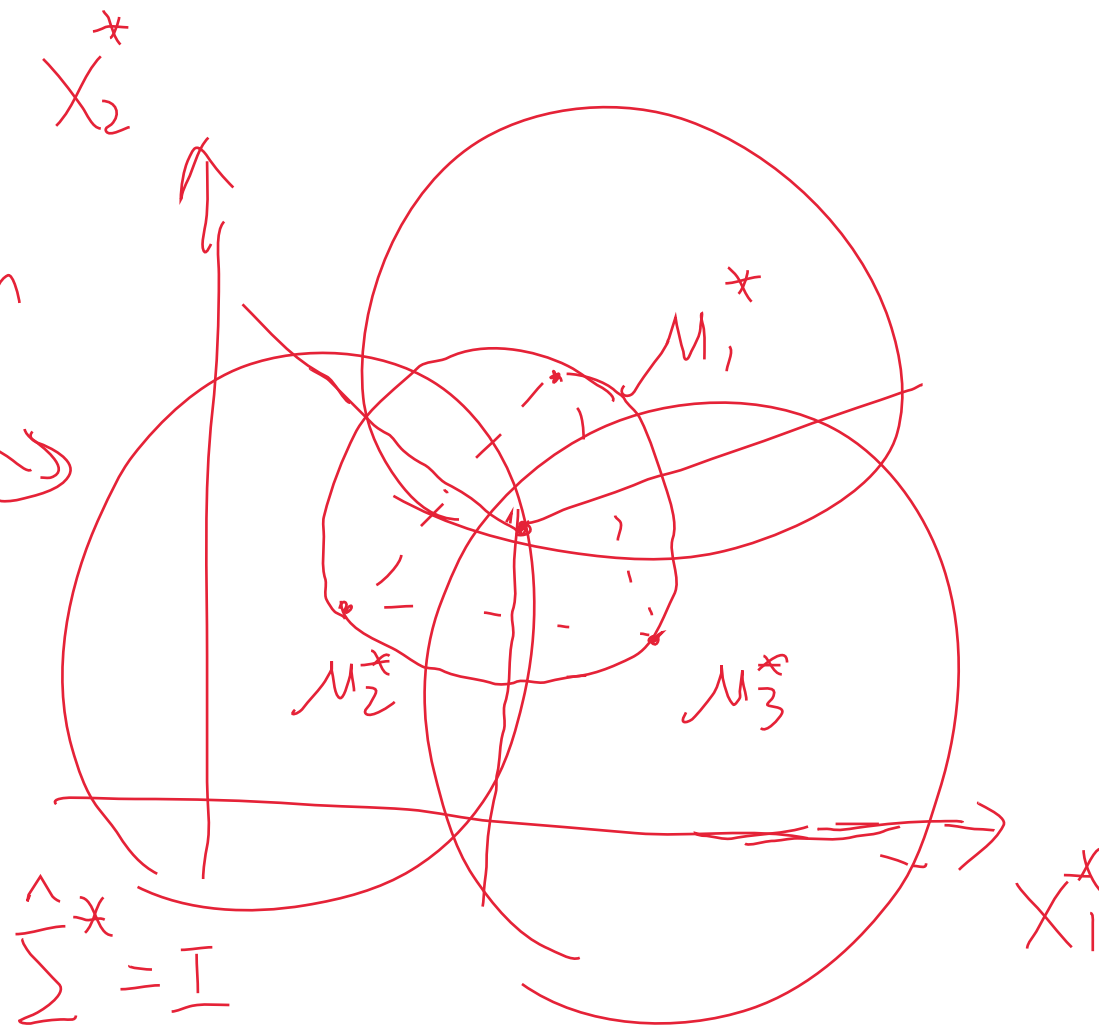4. Classify to its closest class centroid in the transformed space
$$\hat{G}(x) = \underset{k \in \mathcal{G}}{\text{argmin}}\, \frac{1}{2}\|x^* - \hat{\mu}_k^*\|^2 - \ln \hat{\pi}_k$$

X_2

M_1

M_2

M_3

O

X_1

$\hat{\Sigma}$

transform

$x^* \leftarrow \hat{\Sigma}^{-\frac{1}{2}} x$

X_2^*

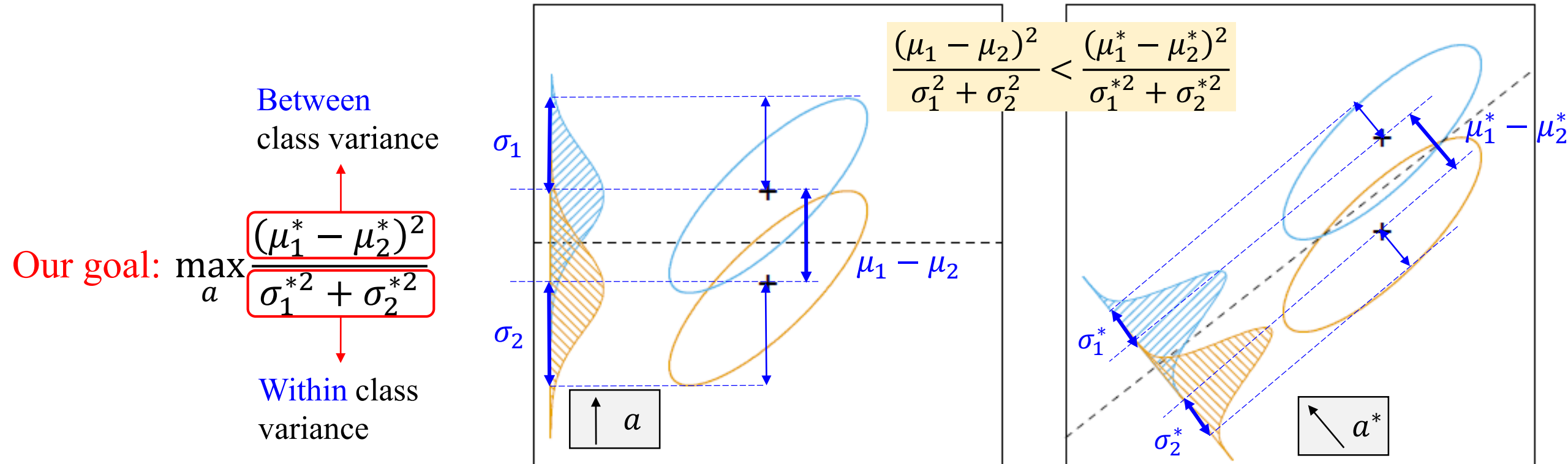M_1^*

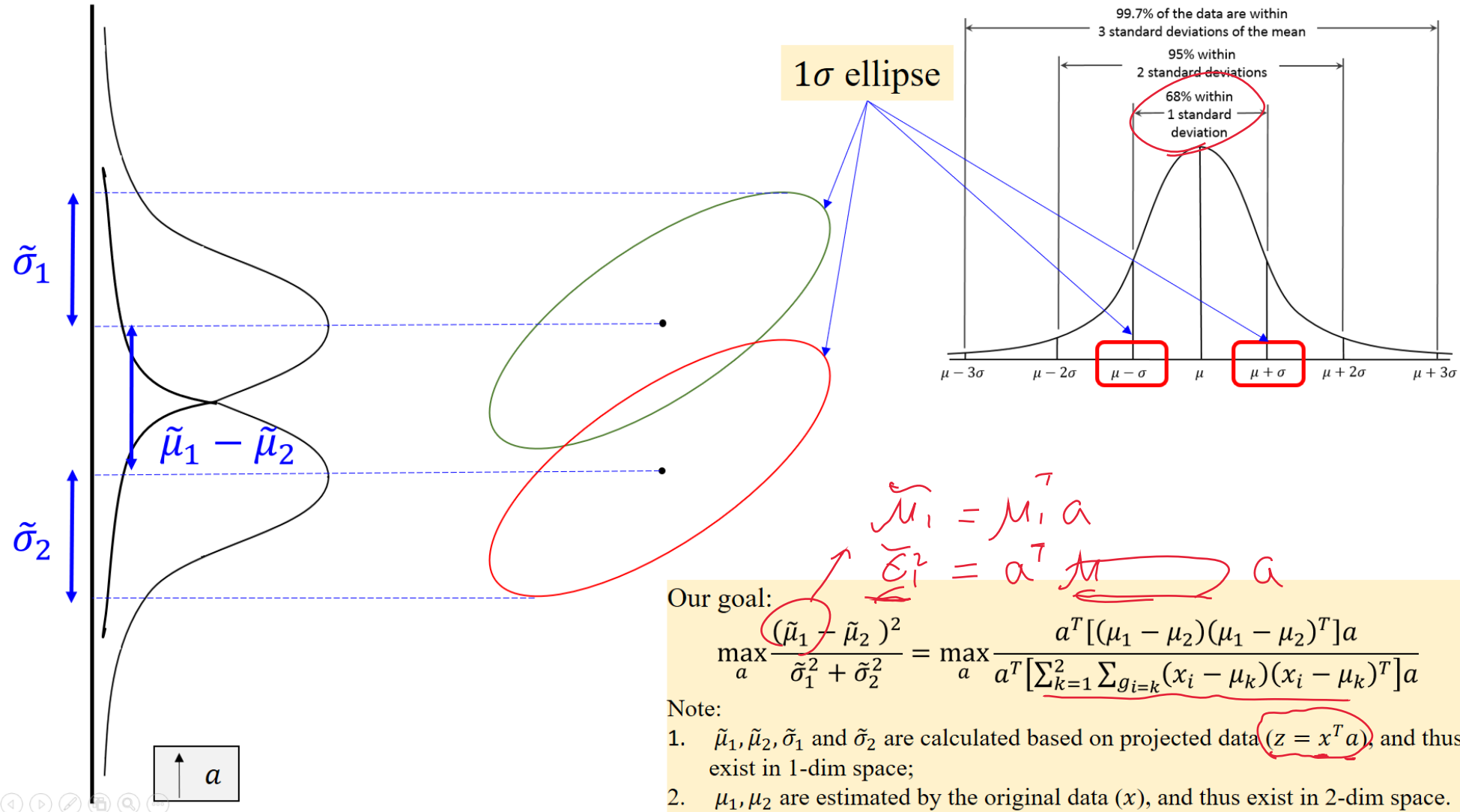M_2^*

M_3^*

X_1^*

$\hat{\Sigma}^* = I$

K classes

LDA $\rightarrow$ at most $(K-1)$-dim space

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.

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

Between class variance

Within class variance

$$\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} < \frac{(\mu_1^* - \mu_2^*)^2}{\sigma_1^{*2} + \sigma_2^{*2}}$$

Fisher's Formulation of Discriminant Analysis

1σ ellipse

99.7% of the data are within
3 standard deviations of the mean

95% within
2 standard deviations

68% within
1 standard
deviation

$\mu - 3\sigma \quad \mu - 2\sigma \quad \mu - \sigma \quad \mu \quad \mu + \sigma \quad \mu + 2\sigma \quad \mu + 3\sigma$

$\tilde{\sigma}_1$

$\tilde{\mu}_1 - \tilde{\mu}_2$

$\tilde{\sigma}_2$

$\tilde{\mu}_1 = M_i^T a$

$\tilde{\mathcal{E}}_1^2 = a^T M a$

Our goal:

$$\max_a \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{\sigma}_1^2 + \tilde{\sigma}_2^2} = \max_a \frac{a^T[(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T]a}{a^T[\sum_{k=1}^2 \sum_{g_{i=k}}(x_i - \mu_k)(x_i - \mu_k)^T]a}$$

Note:
1. $\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\sigma}_1$ and $\tilde{\sigma}_2$ are calculated based on projected data $(z = x^T a)$, and thus exist in 1-dim space;
2. μ_1, μ_2 are estimated by the original data (x), and thus exist in 2-dim space.

a

14

Fisher's Formulation of Discriminant Analysis

$T = B + W$

- Maximize the Rayleigh quotient:

$T = \sum_{i=r}^{N} (x - \mu)(x - \mu)^T$

$$\max_a \frac{a^T \mathbf{B} a}{a^T \mathbf{W} a}$$

□ Between class variance

$$\mathbf{B} = \sum_{k=1}^{K} N_k (\mu_k - \bar{\mu})(\mu_k - \bar{\mu})^T$$

□ Within class variance
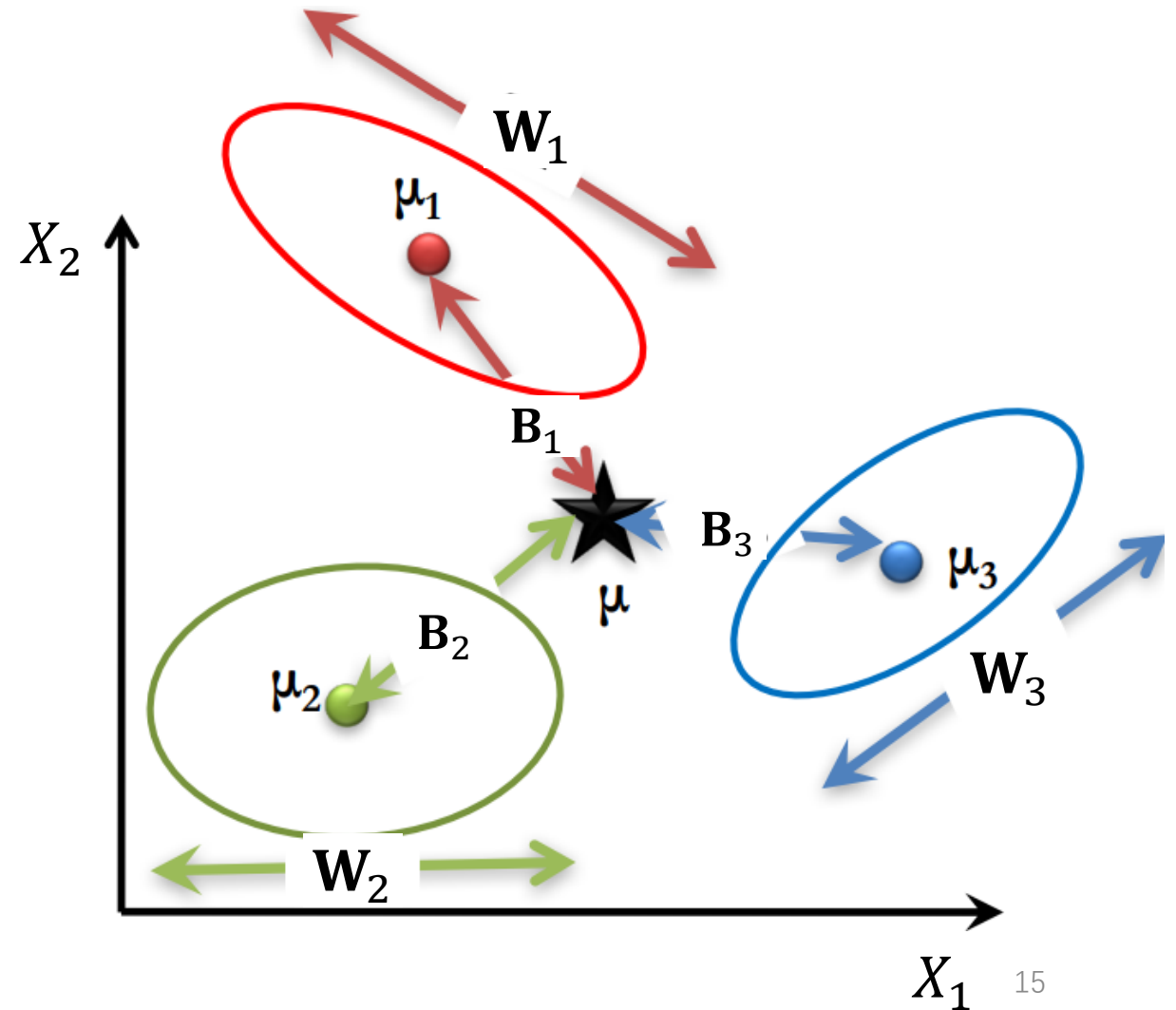
$$\mathbf{W} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \bar{\mu}_k)(x_i - \bar{\mu}_k)^T$$

Fisher's Formulation of Discriminant Analysis

- Maximize the Rayleigh quotient:

$$\max_{a} \frac{a^T \mathbf{B} a}{a^T \mathbf{W} a}$$

  - Between class variance

$$\mathbf{B} = \sum_{k=1}^{K} N_k (\mu_k - \bar{\mu})(\mu_k - \bar{\mu})^T$$

  - Within class variance

$$\mathbf{W} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \bar{\mu}_k)(x_i - \bar{\mu}_k)^T$$

- Equivalently,

$$\max_{a} a^T \mathbf{B} a$$

$$s.t. \ a^T \mathbf{W} a = 1$$

$$L(a, \lambda) = a^T B a - \lambda (a^T W a - 1.)$$

$$\frac{\partial L(a,\lambda)}{\partial a} = 0$$

  - a is discriminant coordinates (canonical variates)
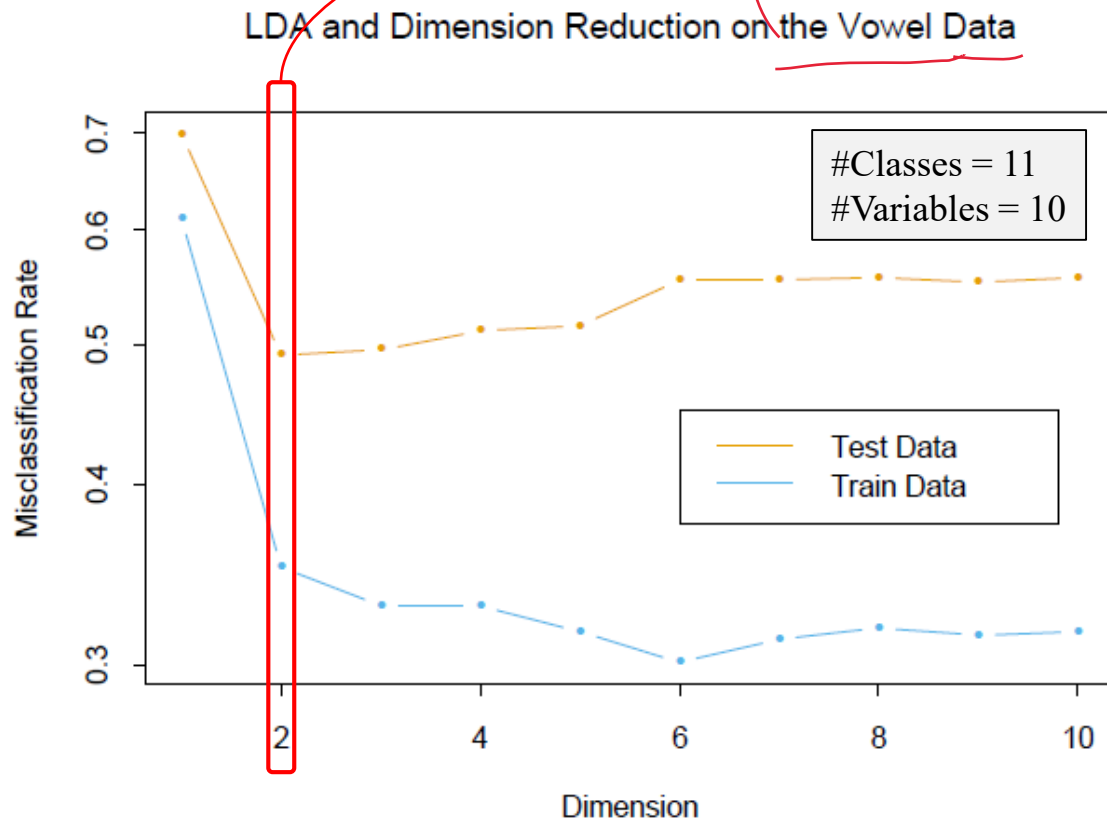
  - Generalized eigenvalue problem

$$\mathbf{B} a = \lambda \mathbf{W} a$$

  which can be efficiently solved
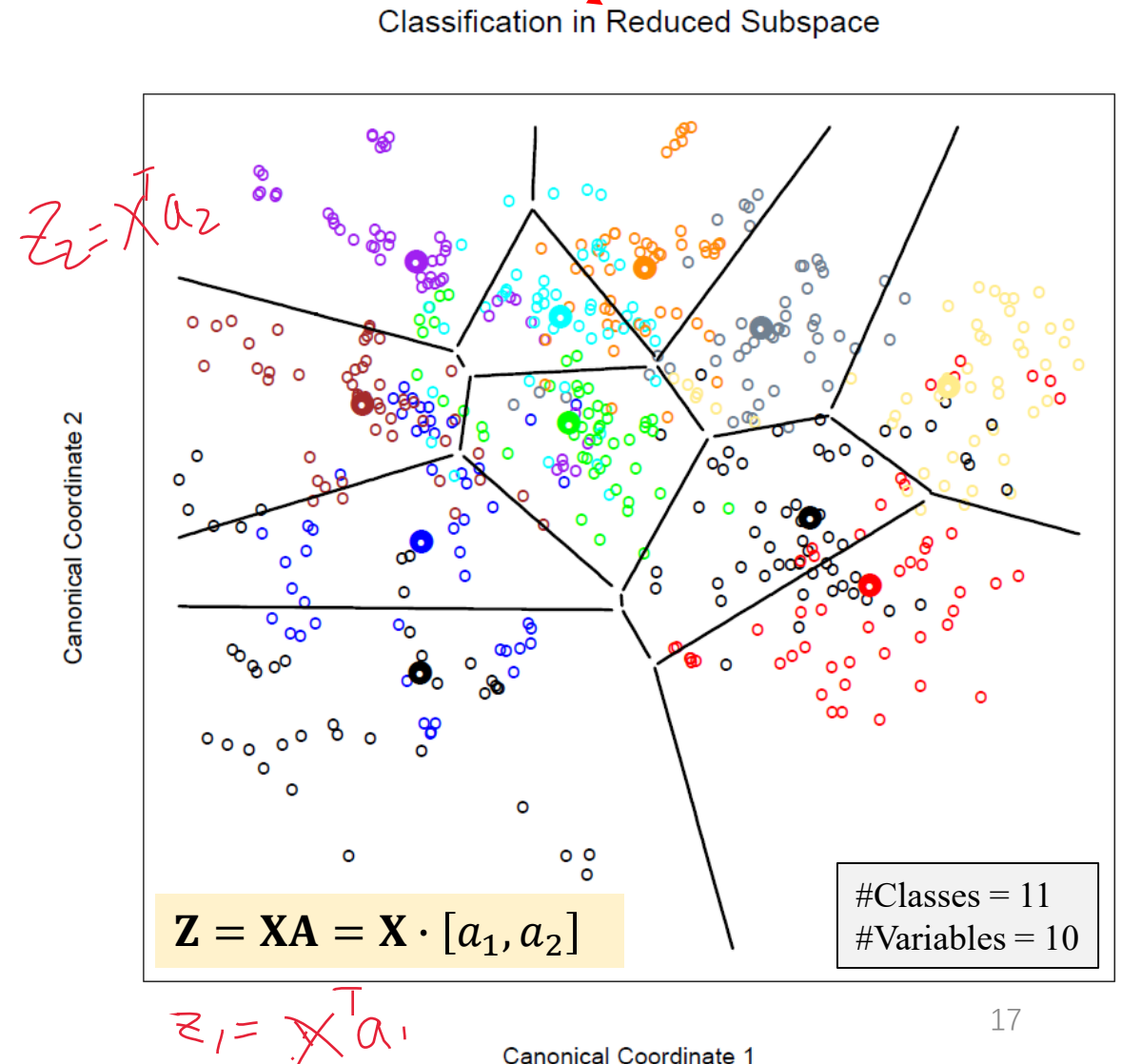
  Ex. 4.1.
  Hint: Lagrangian multipliers

16

Fisher's Formulation of Discriminant Analysis

LDA and Dimension Reduction on the Vowel Data

$(K = 11, \; p = 10)$

#Classes = 11
#Variables = 10

Test Data
Train Data

$$\max_{a} \frac{a^T \mathbf{B} a}{a^T \mathbf{W} a}$$

Classification in Reduced Subspace

$z_2 = X^T a_2$

$z_1 = X^T a_1$

$\mathbf{Z} = \mathbf{XA} = \mathbf{X} \cdot [a_1, a_2]$

#Classes = 11
#Variables = 10
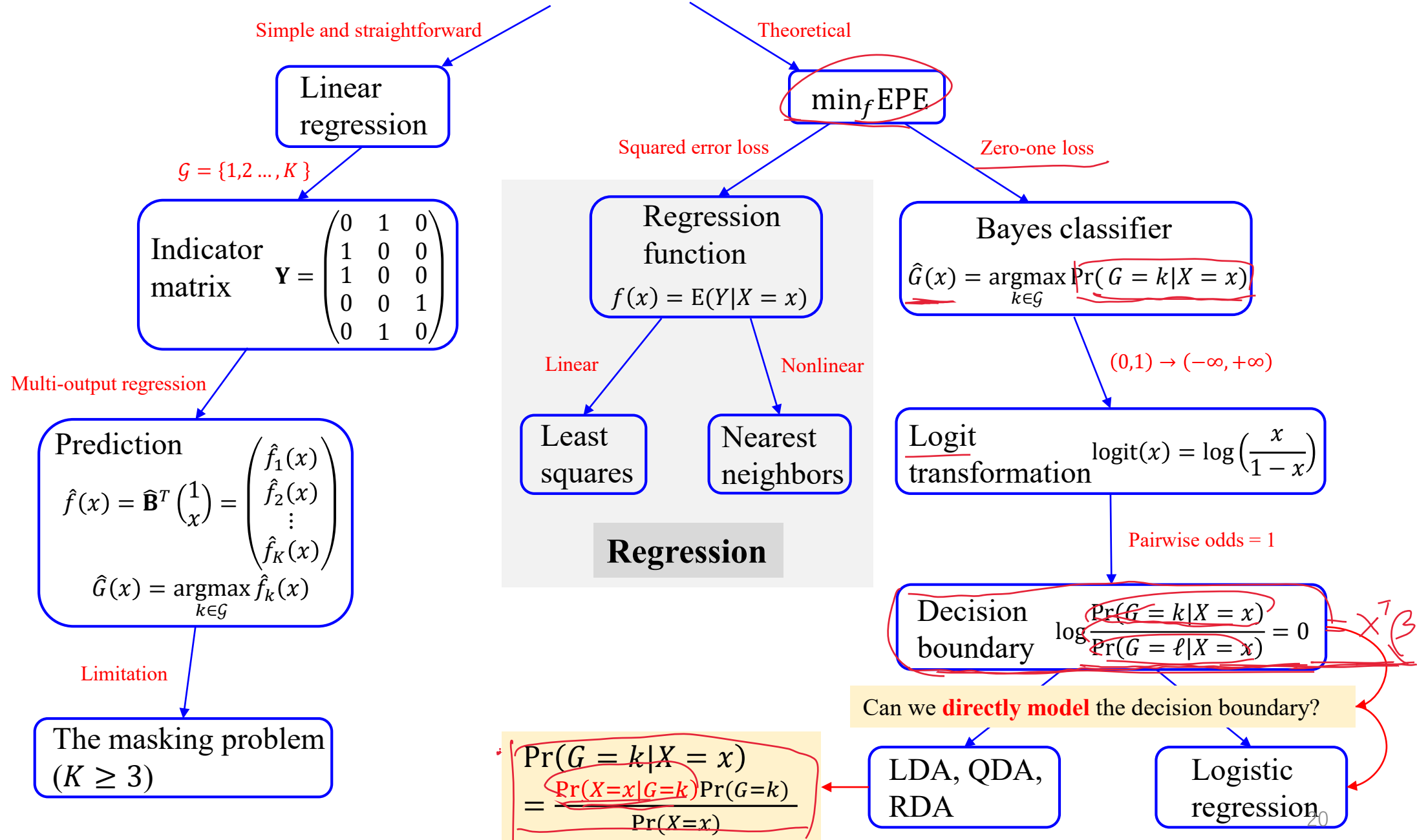
Fisher's Formulation of Discriminant Analysis

We will discuss an example in our **live lecture** according to:

http://www.sci.utah.edu/~shireen/pdfs/tutorials/Elhabian_LDA09.pdf

Linear Methods for Classification II

- Generalization of LDA
    - Regularized Discriminant Analysis
    - Fisher's Formulation of Discriminant Analysis
- **Logistic Regression**
- Summary

Classification

Simple and straightforward → **Linear regression**

Theoretical → $\min_f \text{EPE}$

Linear regression

$\mathcal{G} = \{1, 2, \dots, K\}$ → **Indicator matrix** $\mathbf{Y} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$

Multi-output regression → **Prediction**
$$\hat{f}(x) = \hat{\mathbf{B}}^T \begin{pmatrix} 1 \\ x \end{pmatrix} = \begin{pmatrix} \hat{f}_1(x) \\ \hat{f}_2(x) \\ \vdots \\ \hat{f}_K(x) \end{pmatrix}$$
$$\hat{G}(x) = \underset{k \in \mathcal{G}}{\text{argmax}}\, \hat{f}_k(x)$$

Limitation → **The masking problem** $(K \geq 3)$

$\min_f \text{EPE}$

Squared error loss → **Regression function** $f(x) = \text{E}(Y|X=x)$

Zero-one loss → **Bayes classifier** $\hat{G}(x) = \underset{k \in \mathcal{G}}{\text{argmax}}\, \text{Pr}(G=k|X=x)$

Regression function $f(x) = \text{E}(Y|X=x)$

Linear → **Least squares**

Nonlinear → **Nearest neighbors**

Regression

Bayes classifier

$(0,1) \to (-\infty, +\infty)$ → **Logit transformation** $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$

Pairwise odds = 1 → **Decision boundary** $\log\dfrac{\text{Pr}(G=k|X=x)}{\text{Pr}(G=\ell|X=x)} = 0$

Can we **directly model** the decision boundary?

$$\text{Pr}(G=k|X=x) = \frac{\text{Pr}(X=x|G=k)\text{Pr}(G=k)}{\text{Pr}(X=x)}$$

LDA, QDA, RDA

Logistic regression

20

Linear Logistic Regression

logistic: (sigmoid)

$$\sigma(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

$$\sigma(-z) = 1 - \sigma(z)$$

- Example: binary (two class) classification

Logit: $\quad \log\frac{\Pr(G=1|X=x)}{1-\Pr(G=1|X=x)} = \log\frac{\Pr(G=1|X=x)}{\Pr(G=2|X=x)} = \beta_0 + x^T\beta \quad = 0$
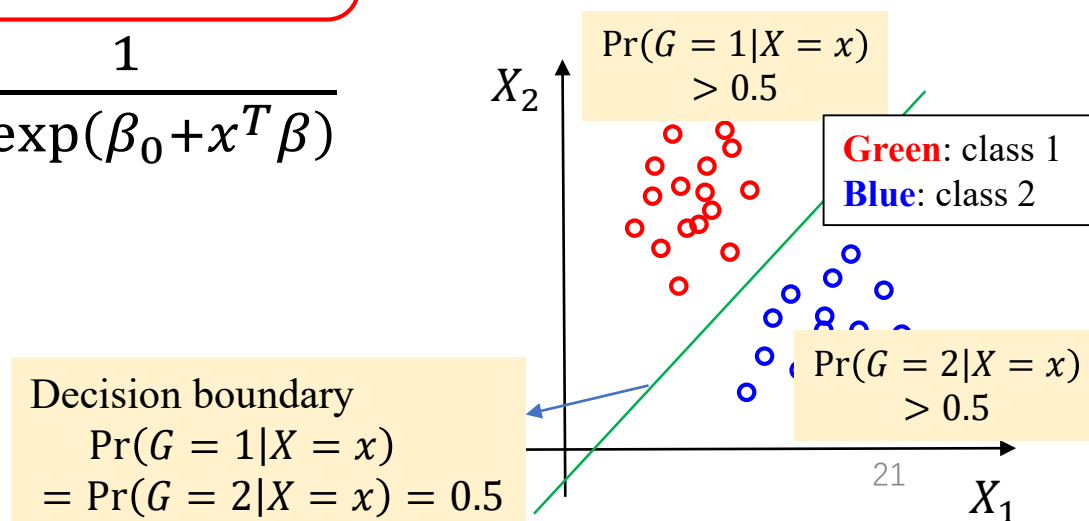
- The posterior probability

x_0

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

logistic$(x^T\beta)$
$(-\infty, +\infty) \to (0,1)$

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

- Decision boundary

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

$\Pr(G = 1|X = x) > 0.5$

Green: class 1
Blue: class 2

X_2

$\Pr(G = 2|X = x) > 0.5$

Decision boundary
$\Pr(G = 1|X = x)$
$= \Pr(G = 2|X = x) = 0.5$

X_1

Linear Logistic Regression

- Model the posterior probabilities of the K classes via linear function in x.

$$\log \frac{\Pr(G = 1 | X = x)}{\Pr(G = K | X = x)} = \beta_{10} + x^T \beta_1$$

$$\log \frac{\Pr(G = 2 | X = x)}{\Pr(G = K | X = x)} = \beta_{20} + x^T \beta_2$$

$$\vdots$$

$$\log \frac{\Pr(G = K - 1 | X = x)}{\Pr(G = K | X = x)} = \beta_{(K-1)0} + x^T \beta_{K-1}$$

- $K - 1$ log-odds or logit function

$$\text{logitPr}(x) = \log \frac{\Pr(x)}{1 - \Pr(x)}$$

- The inverse of logit is logistic function

Logit

$f(x) = \log \frac{x}{1-x}$

A map of values from $(0,1)$ to $(-\infty, +\infty)$

Logistic

$f(x) = \frac{e^x}{1 + e^x}$

A map of values from $(-\infty, +\infty)$ to $(0,1)$

Linear Logistic Regression

- Model the posterior probabilities of the K classes via linear function in x.

$$\log \frac{\Pr(G = 1 | X = x)}{\Pr(G = K | X = x)} = \beta_{10} + x^T \beta_1$$

$$\log \frac{\Pr(G = 2 | X = x)}{\Pr(G = K | X = x)} = \beta_{20} + x^T \beta_2$$

$$\vdots$$

$$\log \frac{\Pr(G = K - 1 | X = x)}{\Pr(G = K | X = x)} = \beta_{(K-1)0} + x^T \beta_{K-1}$$

- $K - 1$ log-odds or logit function

$$\mathrm{logit} \Pr(x) = \log \frac{\Pr(x)}{1 - \Pr(x)}$$

- The inverse of logit is logistic function

- A simple calculation yields

$$\Pr(G = k | X = x) = \frac{\exp(\beta_{k0} + x^T \beta_k)}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + x^T \beta_\ell)},$$

$$k = 1, \dots, K - 1$$

$$\Pr(G = K | X = x) = \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + x^T \beta_\ell)}$$

- Parameter set

$$\theta = \{\beta_{10}, \beta_1, \dots, \beta_{(K-1)0}, \beta_{K-1}\}$$
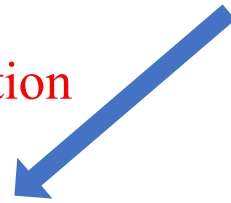
- #parameters $= (p + 1) \times (K - 1)$

Linear Logistic Regression

$$\log \frac{\Pr(G = 1 | X = x)}{\Pr(G = K | X = x)} = \beta_{10} + x^T \beta_1$$

$$\vdots$$

$$\log \frac{\Pr(G = K - 1 | X = x)}{\Pr(G = K | X = x)} = \beta_{(K-1)0} + x^T \beta_{K-1}$$

$$\Pr(G = 1 | X = x) = \Pr(G = K | X = x) \exp(\beta_{10} + x^T \beta_1)$$

$$\vdots$$

$$\Pr(G = K - 1 | X = x) = \Pr(G = K | X = x) \exp(\beta_{(K-1)0} + x^T \beta_{K-1})$$

summation

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

$$\sum_{\ell=1}^{K-1} \Pr(G = \ell | X = x) = \Pr(G = K | X = x) \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + x^T \beta_\ell)$$

$$\Pr(G = K | X = x) = \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + x^T \beta_\ell)}$$

$$\Pr(G = k | X = x) = \frac{\exp(\beta_{k0} + x^T \beta_k)}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + x^T \beta_\ell)}, k = 1, \ldots, K - 1$$

Linear Logistic Regression

$$\log \frac{Pr(G=1|X,\beta)}{Pr(G=2|X,\beta)} = X^T\beta = 0$$

$$\Rightarrow \begin{cases} Pr(G=1|X,\beta) = \sigma(X^T\beta) \\ P_?(G=2|X,\beta) = 1 - \sigma(X^T\beta) \\ \qquad\qquad\qquad = \sigma(-X^T\beta) \end{cases}$$

- Estimating parameter set $\theta = \{\beta_{10}, \beta_1, \dots, \beta_{(K-1)0}, \beta_{K-1}\}$
  - Maximum likelihood estimation (MLE)

  $$\mathcal{L}(\theta) = p(\underline{D}|\theta) = \prod_{i=1}^{N} P(x_i, y_i|\theta)$$

  $$= \prod_{i=1}^{n} P(y_i|x_i,\theta) \, p(x_i)$$

- Log-likelihood for N observations

  $$\max_\theta \quad \ell(\theta) = \log Pr(\mathbf{g}|\mathbf{X};\theta) = \sum_{i=1}^{N} \log Pr(g_i|x_i;\theta) \propto \prod_{i=1}^{N} P(y_i|x_i,\theta)$$

- **Two classes**
  - Bernoulli distribution
  - $Pr(g = y|x;\theta) = p(x;\theta)^y (1 - p(x;\theta))^{1-y}$

  $$Pr(G=1|X,\beta) \qquad Pr(G=2|X,\beta)$$

Class	$g = 1$	$g = 2$
Code	$y = 1$	$y = 0$
Probability	$p(x;\theta)$	$1 - p(x;\theta)$

$G(z) = \dfrac{1}{1+e^{-z}}$, $\log G(z) = 1 - G(-z)$

Linear Logistic Regression

• Two classes

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

$$\ell(\theta) = \sum_{i=1}^{N} \left\{ y_i \log \boxed{p(x_i; \theta)} + (1 - y_i)\log(1 - p(x_i; \theta)) \right\}$$

concave function

$$= \sum_{i=1}^{N} \left\{ y_i \left[x^T\beta - \log\left(1 + e^{x_i^T\beta}\right)\right] - (1 - y_i)\log\left(1 + e^{x_i^T\beta}\right) \right\}$$

$$f(\beta) = \sum_{i=1}^{N} \left\{ y_i x_i^T\beta - \log\left(1 + e^{x_i^T\beta}\right) \right\}$$

$$x_i \leftarrow \begin{pmatrix} 1 \\ x_i \end{pmatrix}$$

$$\beta \leftarrow \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix}$$

$y \in \{-1, 1\}$

$y \in \{0, 1\}$

cross-entropy / logistic-loss $\rightarrow$ $\log(1 + G(y_i x_i^T\beta))$

26

Linear Logistic Regression

- The *first* derivative of $\ell(\theta)$

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{N} \left(y_i x_i - \frac{\exp(x^T \beta)}{1 + \exp(x^T \beta)} \right)$$

$$= \sum_{i=1}^{N} x_i (y_i - p(x_i))$$

- The *second* derivative (Hessian)

$$\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} = \sum_{i=1}^{N} -x_i \left(\frac{\partial p(x_i)}{\partial \beta^T} \right)$$

$$= -\sum_{i=1}^{N} x_i x_i^T \, p(x_i)(1 - p(x_i))$$

- In matrix form

$$\frac{\partial \ell(\beta)}{\partial \beta} = \mathbf{X}^T (\mathbf{y} - \mathbf{p})$$

$$\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} = -\mathbf{X}^T \mathbf{W} \mathbf{X}$$

where $\mathbf{W} \in \mathbb{R}^{N \times N}$ is a diagonal matrix with the i-th diagonal element $p(x_i)(1 - p(x_i))$

The Newton-Raphson algorithm:
find the minimum or maximum iteratively by

$$x^{\text{new}} = x^{\text{old}} - \frac{f'(x^{\text{old}})}{f''(x^{\text{old}})}$$

- The Newton-Raphson step:

$$\beta^{\text{new}} = \beta^{\text{old}} - \left(\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} \right)^{-1} \frac{\partial \ell(\beta)}{\partial \beta}$$

$$= \beta^{\text{old}} + (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{y} - \mathbf{p})$$

$$= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \left(\mathbf{X} \beta^{\text{old}} + \mathbf{W}^{-1} (\mathbf{y} - \mathbf{p}) \right)$$

$$= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{z}$$

$$\beta^{\text{old}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X} \beta^{\text{old}}$$

LS: $\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

- Given the response

$$\mathbf{z} = \mathbf{X} \beta^{\text{old}} + \mathbf{W}^{-1} (\mathbf{y} - \mathbf{p}),$$

- it is represented as a <u>weighted least squares</u> problem:

$$\beta^{\text{new}} \leftarrow \operatorname{argmin}_{\beta} (\mathbf{z} - \mathbf{X}\beta)^T \mathbf{W} (\mathbf{z} - \mathbf{X}\beta)$$

Linear Logistic Regression

- Iteratively reweighted least squares (IRLS) algorithm

Global opt.

1. Initialize β
2. *Repeat*
3.   Form linearized responses
$$z_i = x_i^T \beta + \frac{y_i - p_i}{p_i(1 - p_i)} \qquad \boxed{\mathbf{z} = \mathbf{X}\beta^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})}$$
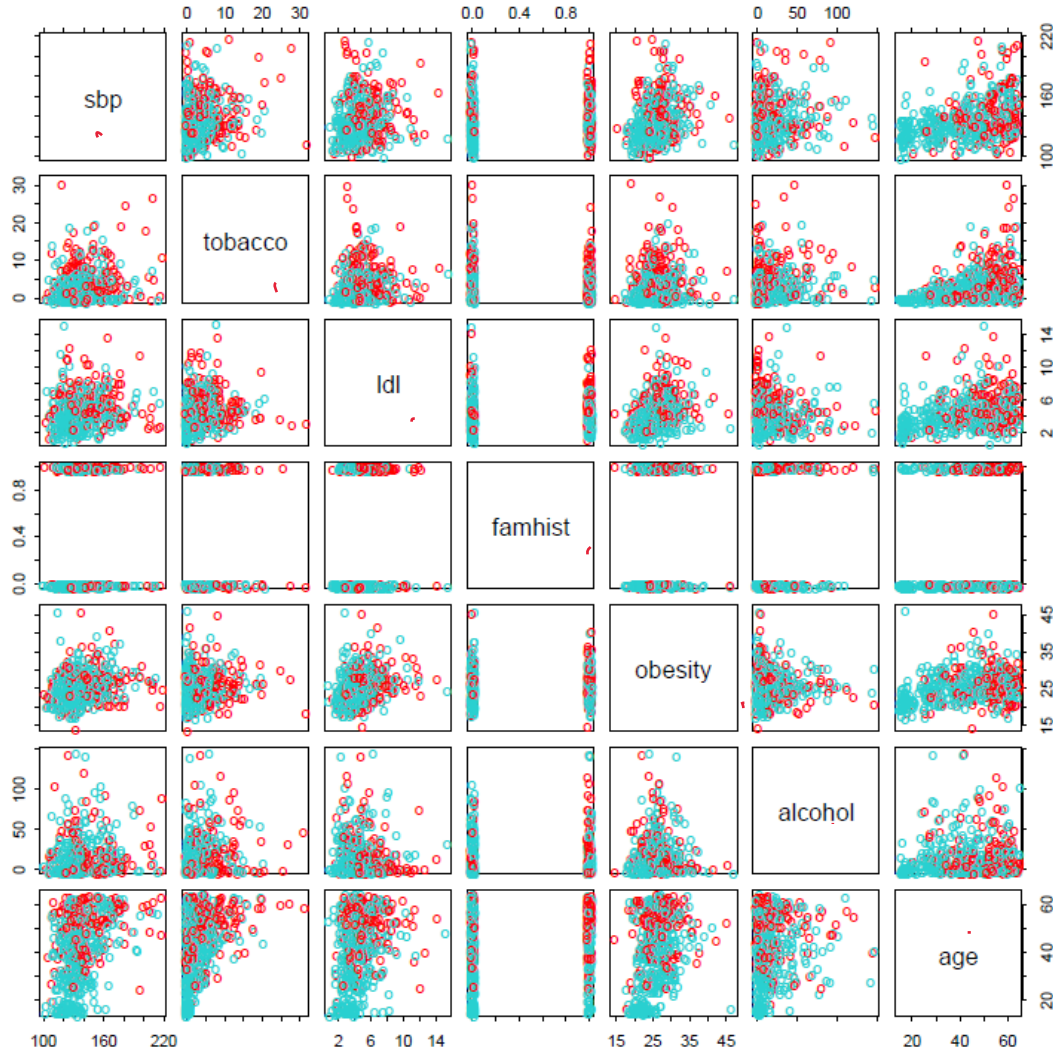4.   Form weights $w_i = p_i(1 - p_i)$
5.   Update β by weighted least squares of z_i on x_i with w_i, $\forall i$
6. *Until convergence*

$$\boxed{\beta^{\text{new}} \leftarrow \text{argmin}_\beta (\mathbf{z} - \mathbf{X}\beta)^T \mathbf{W}(\mathbf{z} - \mathbf{X}\beta)} + \lambda \|\beta\|_1$$

$$\frac{|\ell(\theta^{\text{new}}) - \ell(\theta^{\text{old}})|}{|\ell(\theta^{\text{old}})|} < \varepsilon = 10^{-5}$$

Linear Logistic Regression

Example: South African Heart Disease
- Red: 160 cases
- Green: 302 controls
- Z score measures the significance of a coefficient

feature	Coefficient	Std. Error	Z Score	
sbp	0.006	0.006	1.023	收缩压
tobacco	0.080	0.026	3.034	
ldl	0.185	0.057	3.219	
famhist	0.939	0.225	4.178	
obesity	−0.035	0.029	−1.187	肥胖
alcohol	0.001	0.004	0.136	饮酒
age	0.043	0.010	4.184	

The data is fitted by logistic regression

Linear Logistic Regression

- L_1 regularized logistic regression

log-likelihood

$$\max_{\beta_0, \beta} \left\{ \sum_{i=1}^{N} \left[y_i(\beta_0 + \beta^T x_i) - \log(1 + e^{\beta_0 + \beta^T x_i}) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- Standardize the inputs, and penalize without β_0

- Solved by the Newton algorithm
  - Replace the weighted least squares by the weighted lasso.

- L_2 regularized logistic regression? Algorithm?

Step 5 → the weighted Ridge regression.

Handwritten annotations at top:

Discriminative v.s. Generative

$P(Y|X)$ $P(X,Y)$

$Y = X^T\beta + \varepsilon,$

$(P(Y|X) = N(X^T\beta, \sigma^2))$

$P(A|B) = P(B|A)P(A)$

(Logistic,

Ridge, lasso)

$\begin{pmatrix} LDA \\ NB \end{pmatrix}$

$P(Y|X) = G(z)$

Connection between LDA and Logistic Regression

- The linear logistic model only specifies the conditional distribution, while the LDA model specifies the joint distribution

  Handwritten: $P(Y|X) = \dfrac{P(X|Y)P(Y)}{P(X)}$ Gaussian

  LDA:

- If the additional assumption made by LDA is appropriate, LDA tends to estimate the parameters more efficiently.

  Handwritten: $\textcircled{T} = B + W$

- Another advantage of LDA is that samples without class labels can be used under the model of LDA. On the other hand, LDA is not robust to gross outliers. Because logistic regression relies on fewer assumptions, it seems to be more robust to the non-Gaussian type of data.

- In practice, logistic regression and LDA often give similar results.

Linear Methods for Classification II

- Generalization of LDA
  - Regularized Discriminant Analysis
  - Fisher's Formulation of Discriminant Analysis
- Logistic Regression
- Summary