Introduction to Machine Learning CS182

Lu Sun

School of Information Science and Technology

ShanghaiTech University

October 24, 2023

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

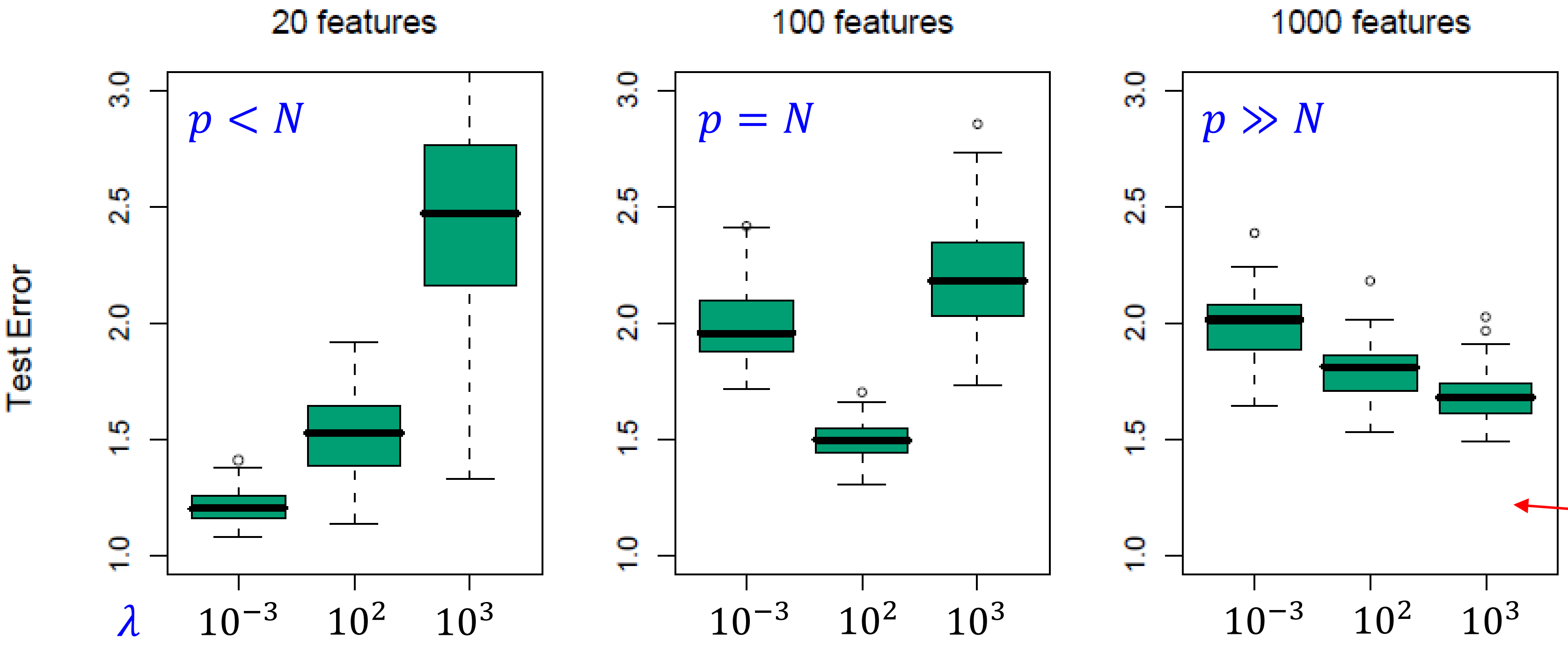
Linear Methods for Classification II

- 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



Example

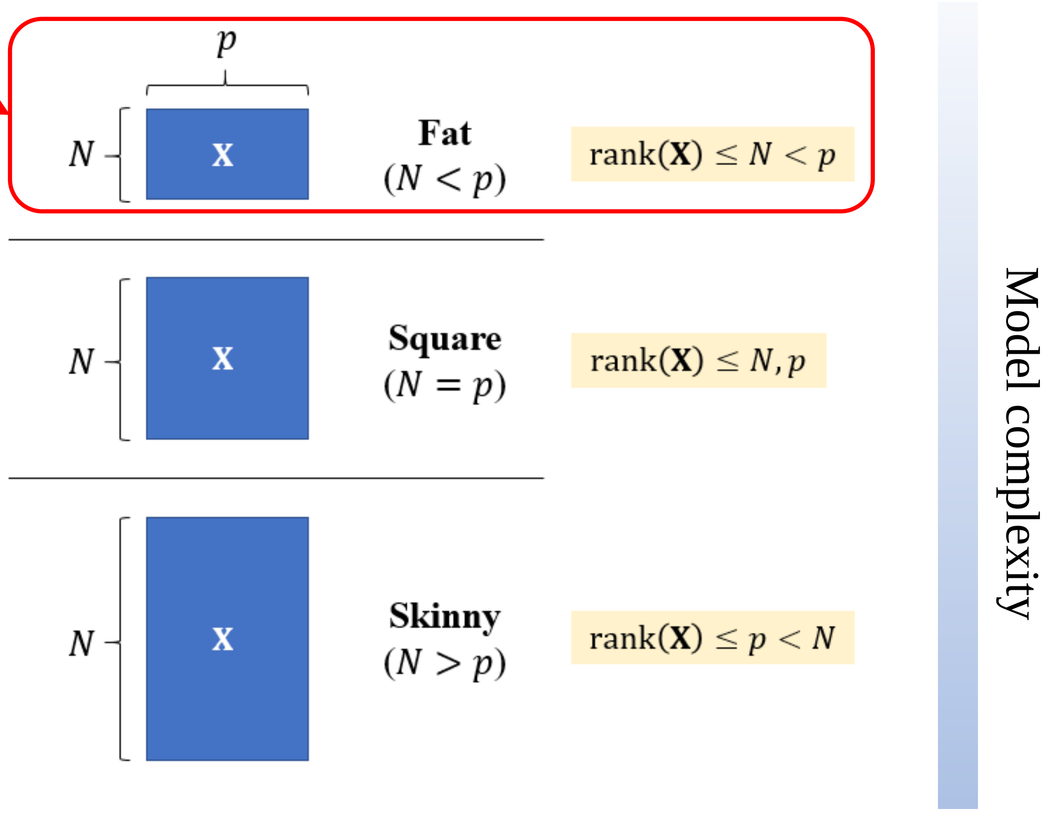
- 100 samples are generated by a linear model
- Ridge regression
- Relative error (divide by Bayes error)

No enough information to estimate the high-dimensional covariance matrix

Regularized Discriminant Analysis

High dimensional problems ($p \gg N$)

- Cannot fit LDA to the data
  - inversion of a $p \times p$ covariance matrix Σ
  - Σ is singular, due to $\text{rank}(\Sigma) \leq N \ll p$
- Regularization is necessary
  - No enough data to estimate feature dependencies
  - E.g., independent assumption on features
    - Diagonal within-class covariance matrix
      #paras: $K \times p \times p \rightarrow K \times p$



Regularized Discriminant Analysis

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 $\hat{\Sigma}$

Diagonal LDA

- Independent assumption on feature dependencies

$$\hat{\Sigma} = \mathrm{diag}(\hat{\Sigma})$$

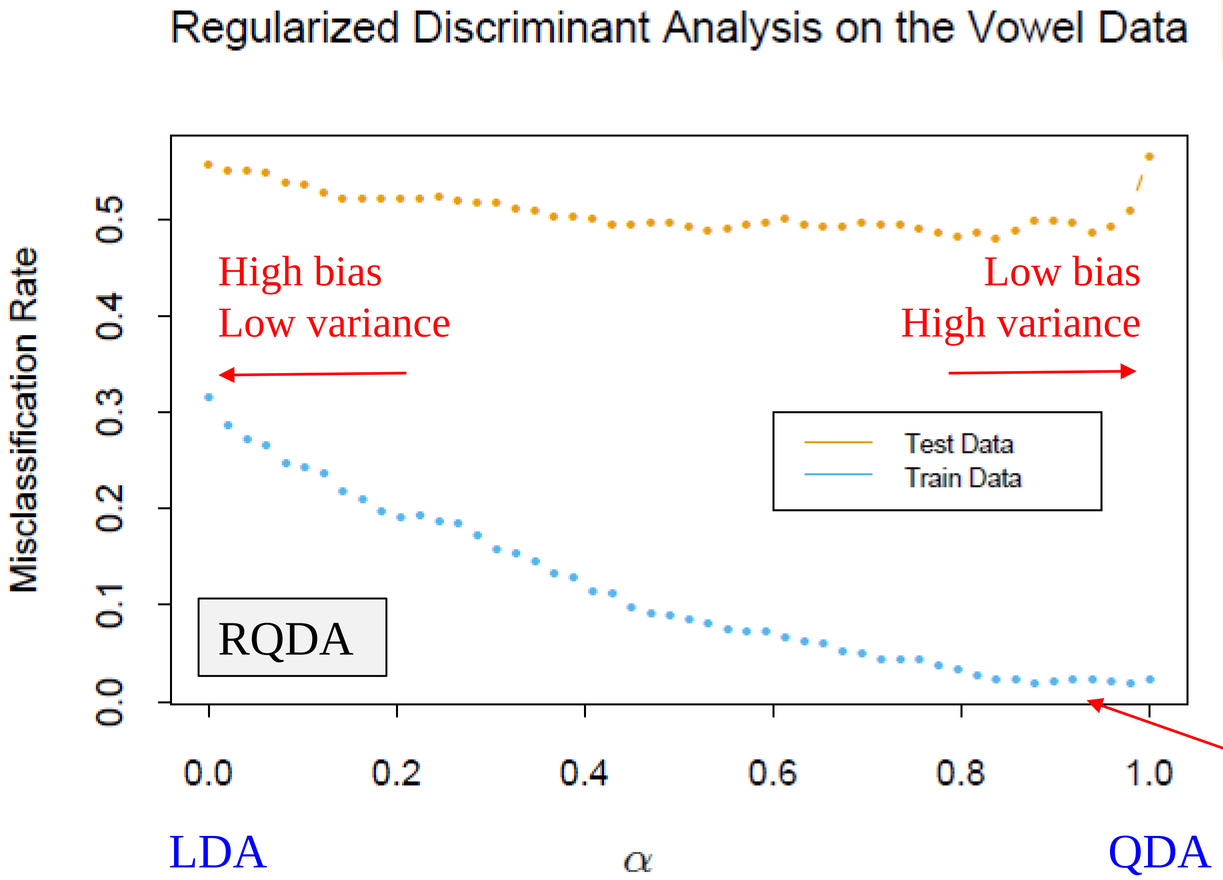
Regularized Discriminant Analysis

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

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

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

Regularized Discriminant Analysis

$\mathrm{diag}(\hat{\Sigma})$		$\hat{\Sigma}(\gamma)$		$\hat{\Sigma}$		$\hat{\Sigma}_k(\alpha)$		$\hat{\Sigma}_k$
Diagonal LDA	$\longrightarrow$	RLDA	$\longrightarrow$	LDA	$\longrightarrow$	RQDA	$\longrightarrow$	QDA

High bias
Low variance

Low bias
High variance

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

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

Fisher's Formulation of Discriminant Analysis

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

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

$$\Pr(G = k|X = x) = \frac{\overbrace{\boxed{\Pr(X = x|G = k)}}^{\mathcal{N}(\hat{\mu}_k, \widehat{\boldsymbol{\Sigma}})}\overbrace{\boxed{\Pr(G = k)}}^{\hat{\pi}_k}}{\Pr(X = x)}$$

3. $$\begin{aligned}\log \Pr(G = k|X = x) &= -\frac{1}{2}(x - \hat{\mu}_k)^T\widehat{\boldsymbol{\Sigma}}^{-1}(x - \hat{\mu}_k) + \log \hat{\pi}_k + \underbrace{\boxed{C}}_{\text{Constant}} \\ &= -\frac{1}{2}(x - \hat{\mu}_k)^T\mathbf{UD}^{-\frac{1}{2}}\left(\mathbf{UD}^{-\frac{1}{2}}\right)^T(x - \hat{\mu}_k) + \log \hat{\pi}_k + C \\ &= -\frac{1}{2}\left(\mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T x - \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T\hat{\mu}_k\right)^T\left(\mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T x - \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T\hat{\mu}_k\right) + \log \hat{\pi}_k + C \\ &= -\frac{1}{2}(x^* - \hat{\mu}_k^*)^T(x^* - \hat{\mu}_k^*) + \log \hat{\pi}_k + C \\ &= -\frac{1}{2}\|x^* - \hat{\mu}_k^*\|^2 + \ln \hat{\pi}_k + C\end{aligned}$$

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

Fisher's Formulation of Discriminant Analysis

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

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

LDA: Approach 2

1. Estimating $\widehat{\boldsymbol{\Sigma}}$, $\hat{\mu}_k$ and $\hat{\pi}_k$
2. Eigen-decomposition:
$$\widehat{\boldsymbol{\Sigma}} = \mathbf{UDU}^T$$
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
$$\hat{G}(x) = \underset{k \in \mathcal{G}}{\operatorname{argmin}} \frac{1}{2} \|x^* - \hat{\mu}_k^*\|^2 - \ln \hat{\pi}_k$$

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

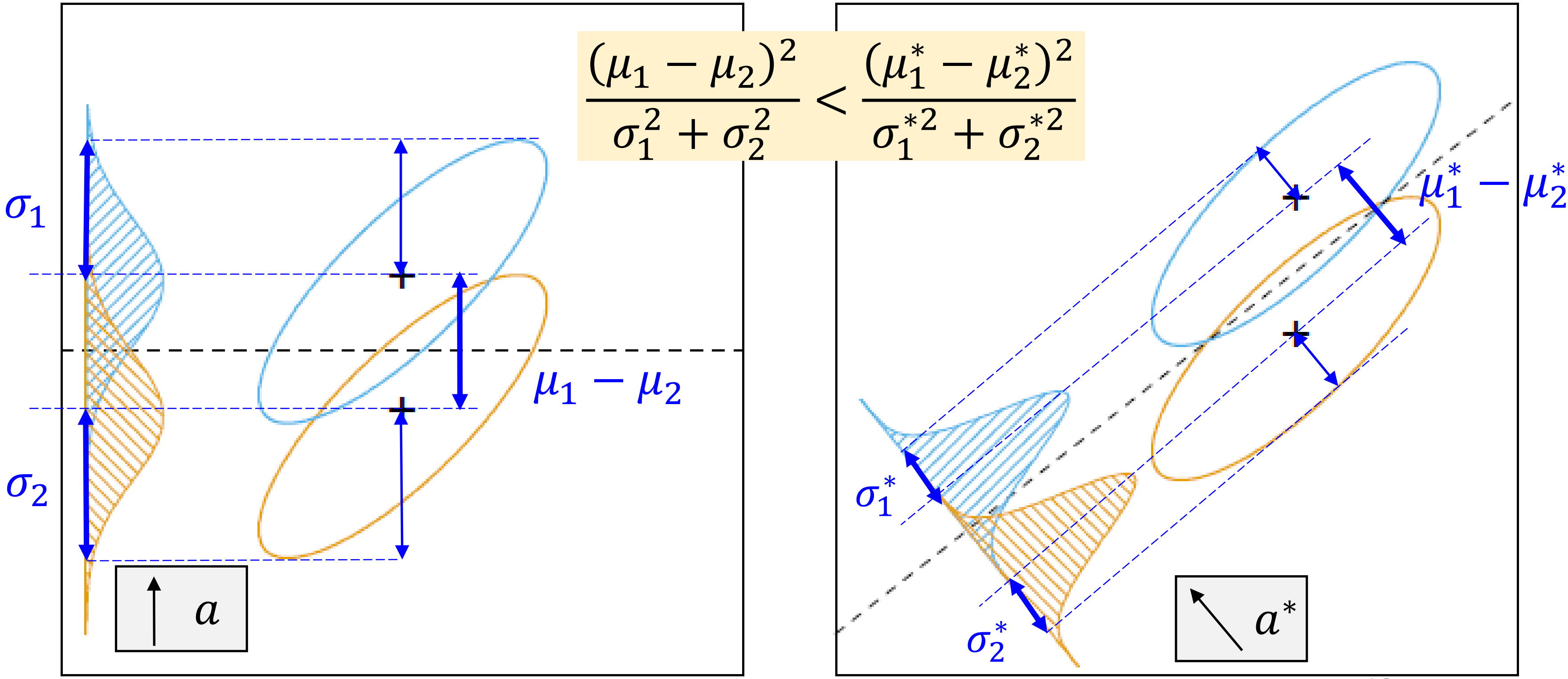
Fisher's Formulation of Discriminant Analysis

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

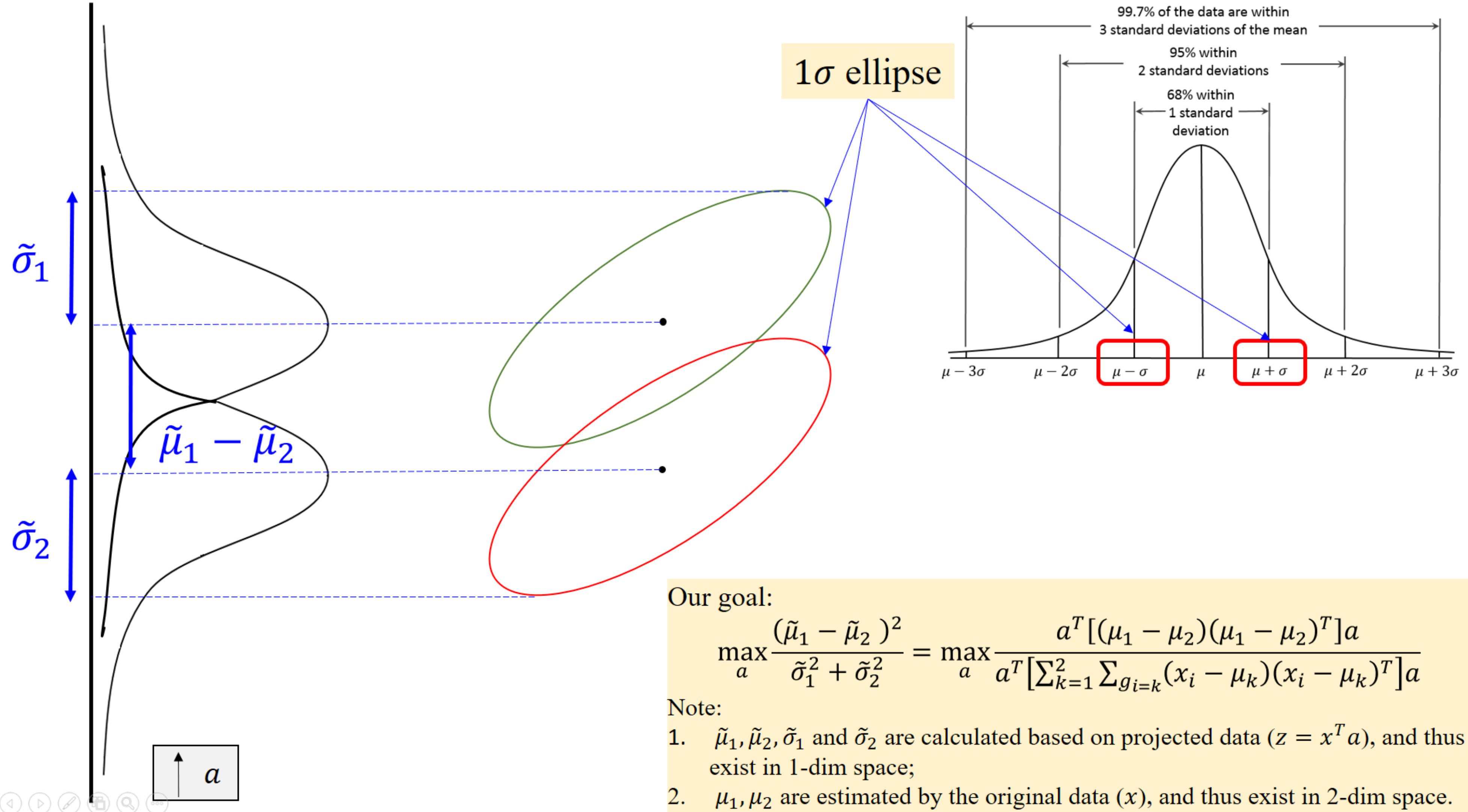
Between class variance

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

Within class variance



Fisher's Formulation of Discriminant Analysis



Our goal:

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

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.

Fisher's Formulation of Discriminant Analysis

- Maximize the Rayleigh quotient:

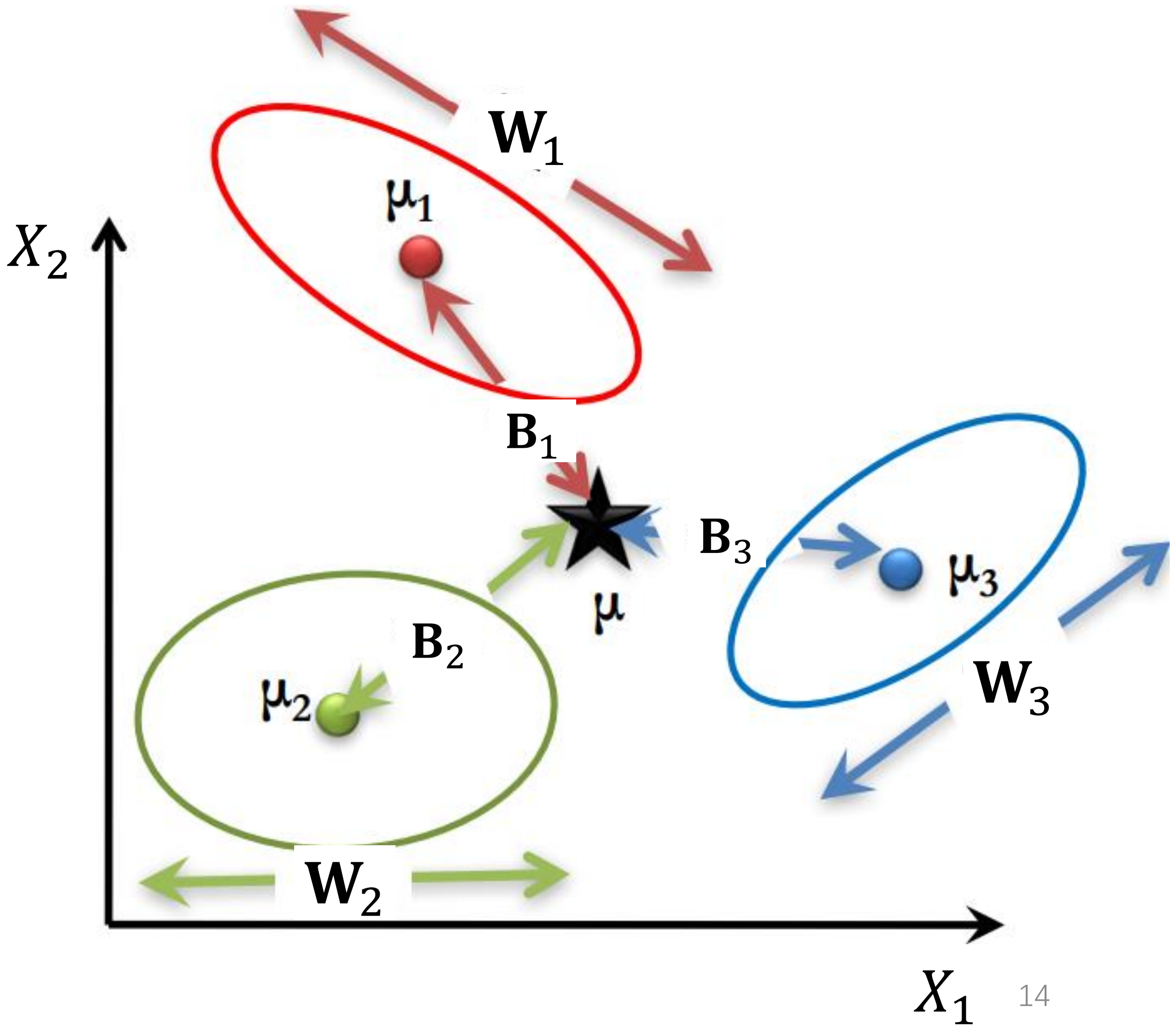
$$\max_a \frac{a^T \mathbf{B} a}{a^T \mathbf{W} a}$$

  - Between class variance

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

  - Within class variance

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



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,

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

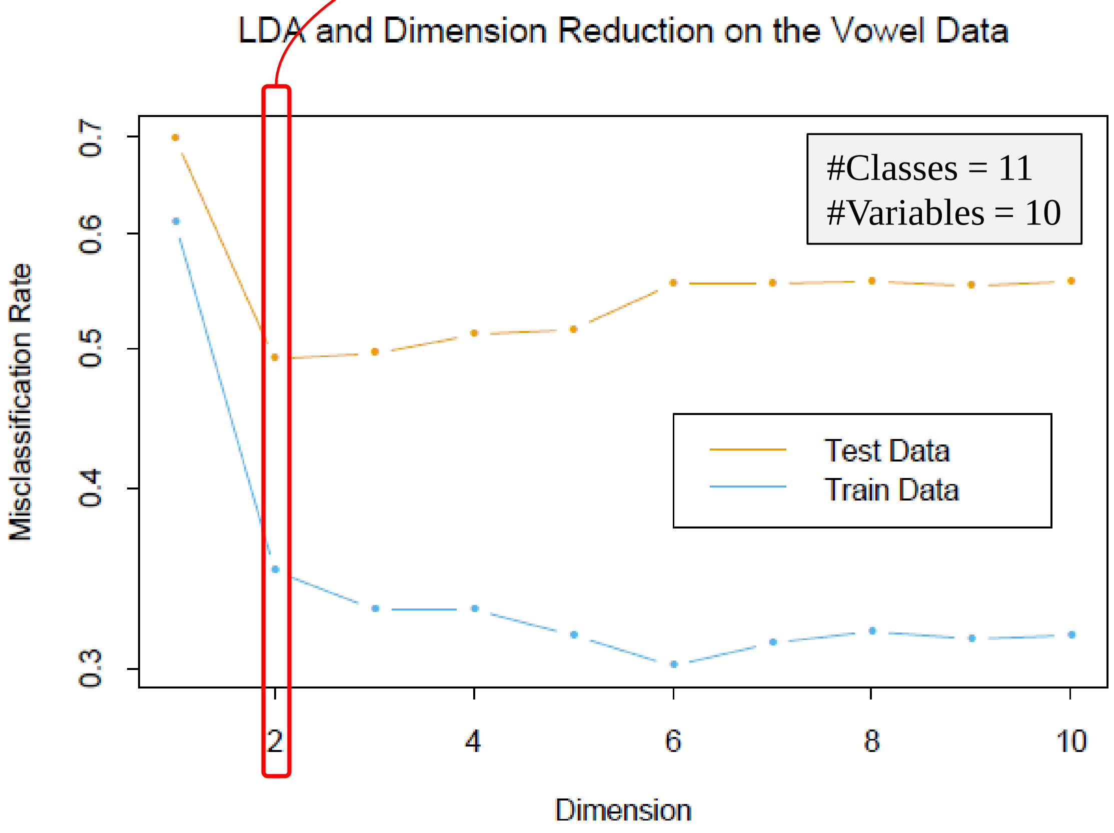
  - a is discriminant coordinates (canonical variates)
  - Generalized eigenvalue problem

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

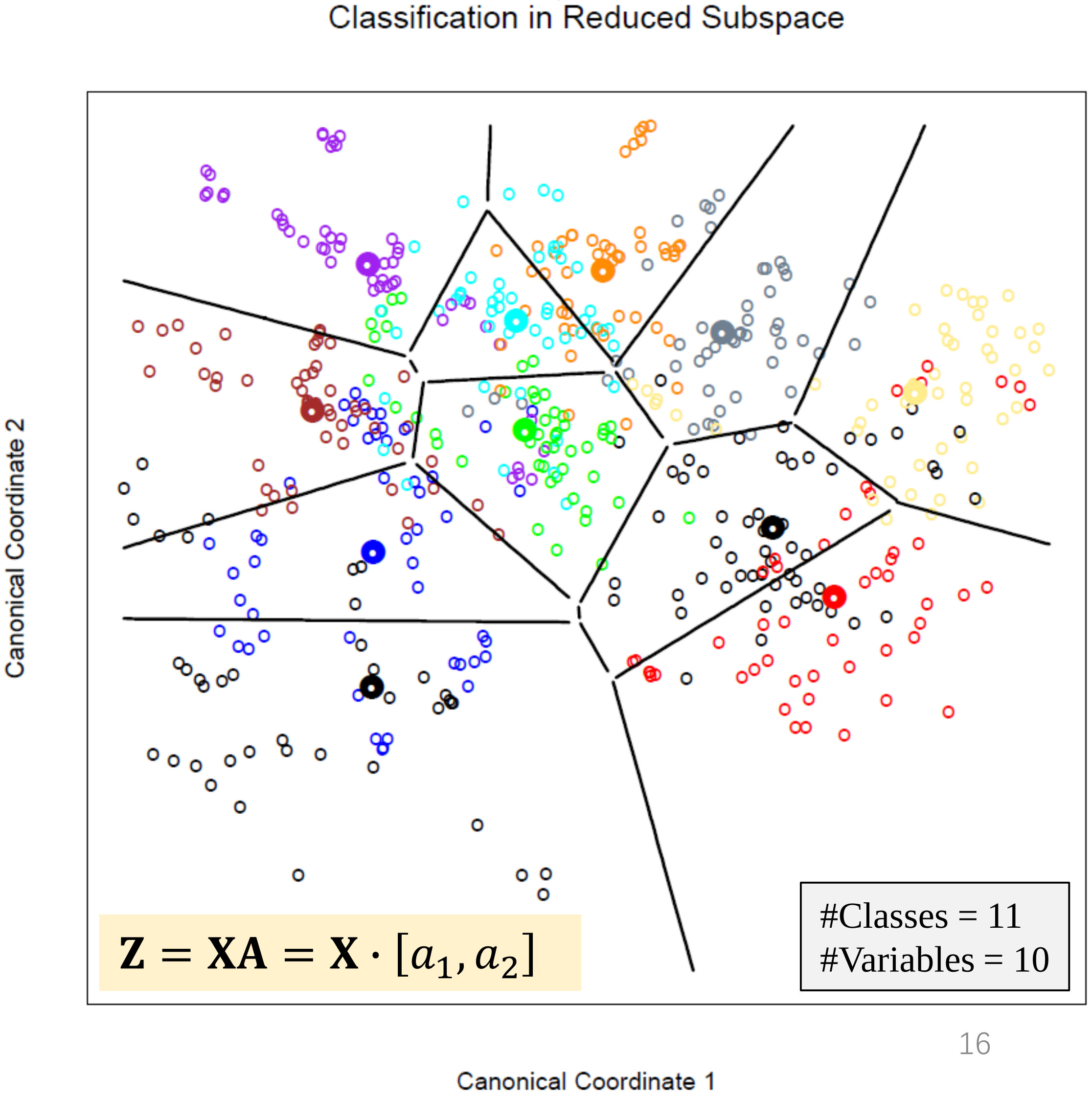
which can be efficiently solved

Ex. 4.1.
Hint: Lagrangian multipliers

Fisher's Formulation of Discriminant Analysis



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

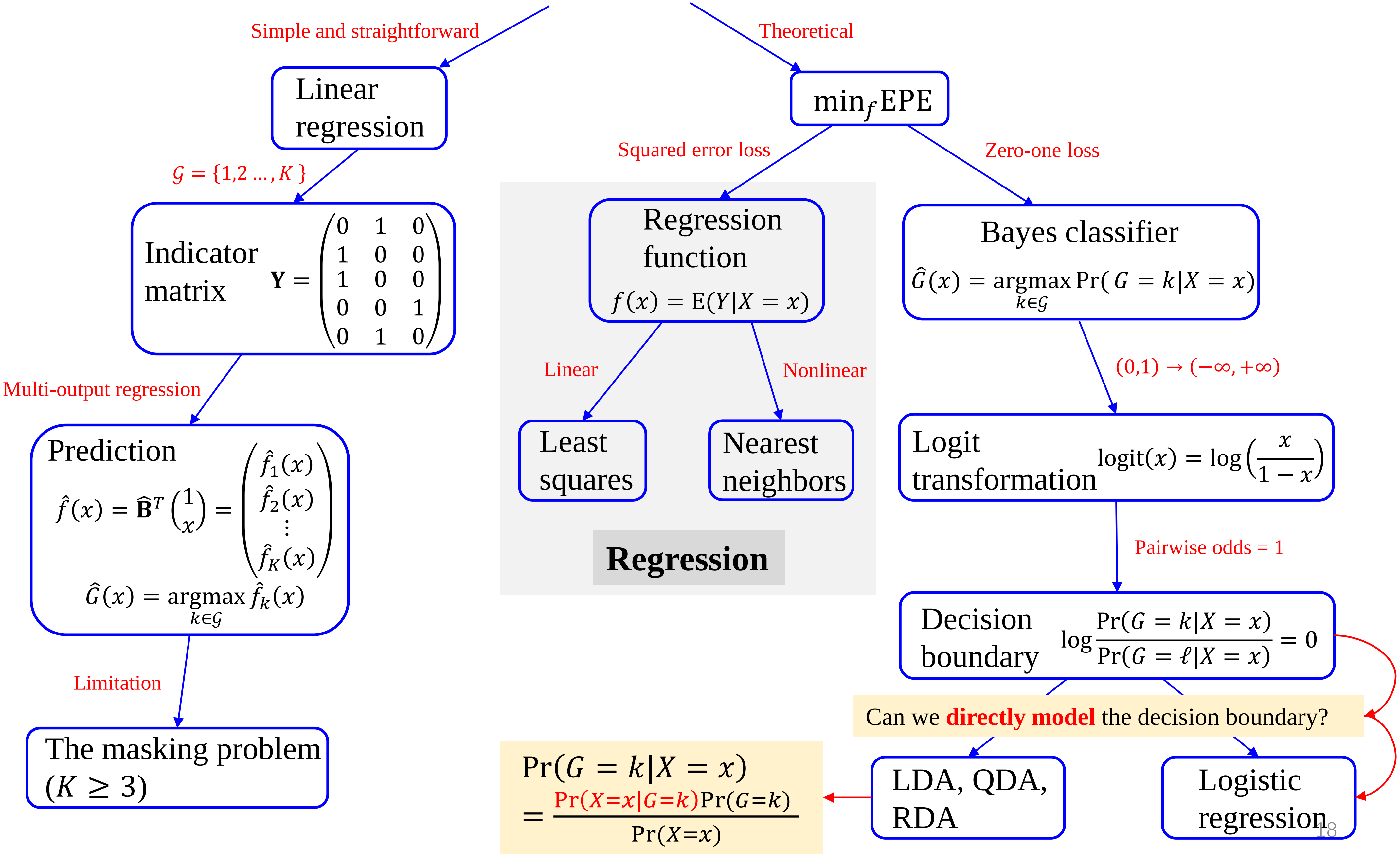


$$\mathbf{Z} = \mathbf{XA} = \mathbf{X} \cdot [a_1, a_2]$$

Linear Methods for Classification II

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

Classification



Linear Discriminant Analysis

- Recall our discussion on linear regression of an indicator matrix

Linear classification:

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

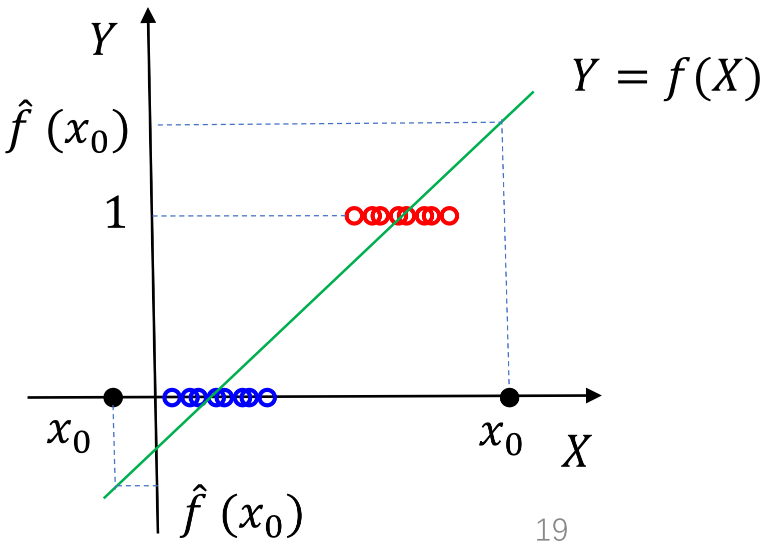
Minimizing EPE:

$$\hat{G}(x) = \underset{k \in \mathcal{G}}{\operatorname{argmax}} \boxed{\Pr(G = k | X = x)}$$

- It is inappropriate to represent a posterior directly by a linear function.

- Solution: make some monotone transformation of the posterior be linear in X

Linear decision boundary



Linear Discriminant Analysis

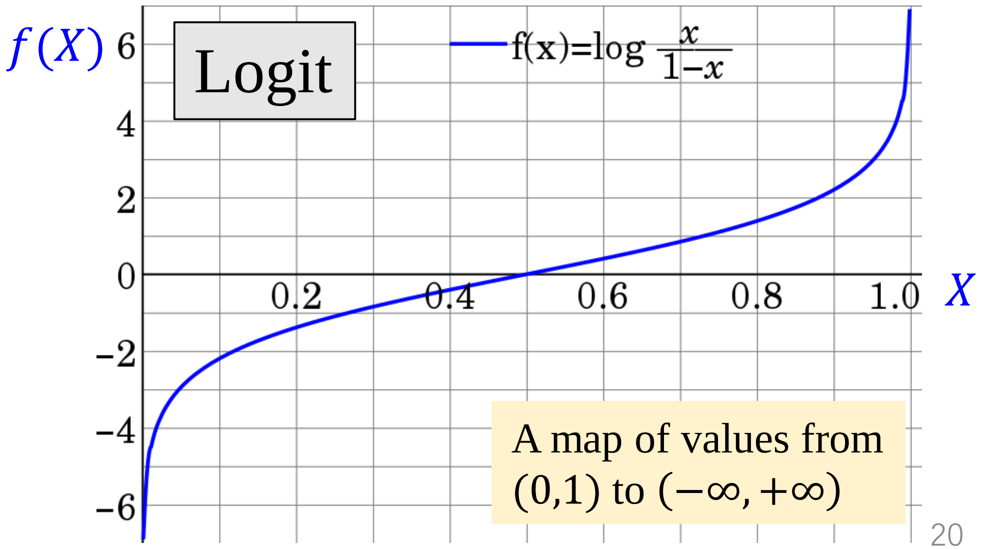
- Logit transform

Odds (发生比)

$$\operatorname{logit}(\Pr(x)) = \log\left(\boxed{\frac{\Pr(x)}{1-\Pr(x)}}\right)$$

It maps $\Pr(x) \in (0,1)$ to $\operatorname{logit}(\Pr(x)) \in (-\infty, +\infty)$

- Decision boundary
  - Odds equals to 1
  - Or, logit equals to 0



Linear Discriminant Analysis

- Example: binary (two class) classification

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

- The posterior probability

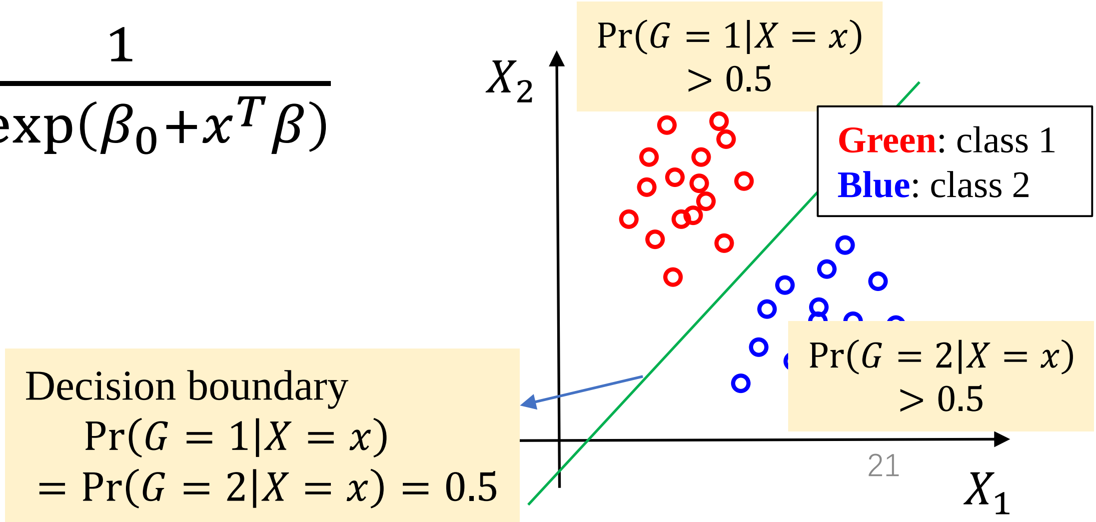
$$\Pr(G=1|X=x) = \frac{\boxed{\exp(\beta_0 + x^T\beta)}}{1+\exp(\beta_0 + x^T\beta)},$$

$\exp(x) = e^x$

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

- Decision boundary

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



Linear Logistic Regression

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

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

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

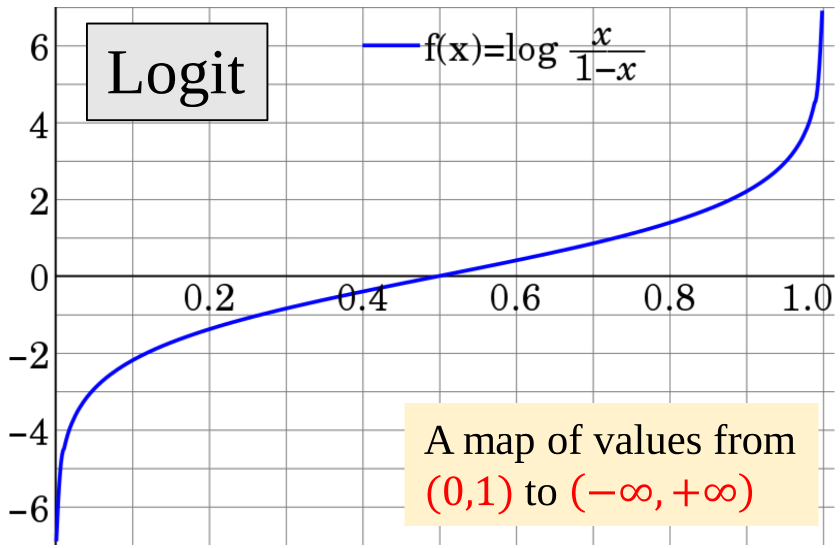
$$\vdots$$

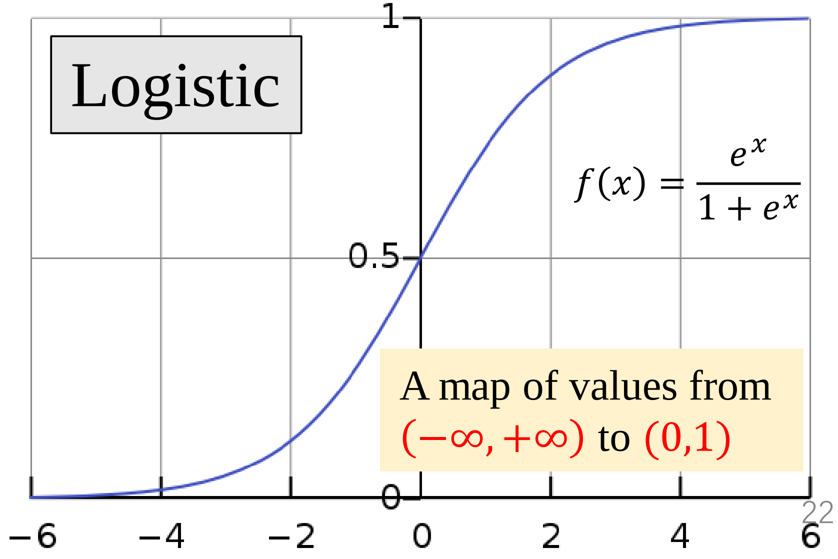
$$\log \frac{\Pr(G=K-1|X=x)}{\Pr(G=K|X=x)} = \beta_{(K-1)0} + x^T\beta_{K-1}$$

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

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

- The inverse of logit is logistic function





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

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

- The inverse of logit is logistic function

?

- A simple calculation yields

$$\Pr(G=k|X=x)=\frac{\exp(\beta_{k0}+x^T\beta_k)}{1+\sum_{\ell=1}^{K-1}\exp(\beta_{\ell0}+x^T\beta_\ell)},$$

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

$$\Pr(G=K|X=x)=\frac{1}{1+\sum_{\ell=1}^{K-1}\exp(\beta_{\ell0}+x^T\beta_\ell)}$$

- Parameter set

$$\theta=\{\beta_{10},\beta_1,\ldots,\beta_{(K-1)0},\beta_{K-1}\}$$

- #parameters = $(p+1)\times(K-1)$

Linear Logistic Regression

$$\log \frac{\Pr(G=1|X=x)}{\Pr(G=K|X=x)} = \beta_{10} + x^T\beta_1$$
$$\vdots$$
$$\log \frac{\Pr(G=K-1|X=x)}{\Pr(G=K|X=x)} = \beta_{(K-1)0} + x^T\beta_{K-1}$$

$$\Pr(G=1|X=x) = \Pr(G=K|X=x)\exp(\beta_{10} + x^T\beta_1)$$
$$\vdots$$
$$\Pr(G=K-1|X=x) = \Pr(G=K|X=x)\exp\left(\beta_{(K-1)0} + x^T\beta_{K-1}\right)$$

summation

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

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

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

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

Linear Logistic Regression

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

- Log-likelihood for N observations

$$\ell(\theta) = \log \Pr(\mathbf{g}|\mathbf{X};\theta) = \sum_{i=1}^{N} \log \Pr(g_i|x_i;\theta)$$

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

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

Linear Logistic Regression

- Two classes

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

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

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

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

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

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

Please refer to:
https://en.wikipedia.org/wiki/Cross_entropy#Cross-entropy_loss_function_and_logistic_regression

Linear Logistic Regression *

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 *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 step:

$$\beta^{\text{new}} = \beta^{\text{old}} - \left(\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} \right)^{-1} \frac{\partial \ell(\beta)}{\partial \beta}$$
$$= \beta^{\text{old}} + (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{p})$$
$$= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\left(\mathbf{X}\beta^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})\right)$$
$$= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}$$

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

- Given the response

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

- it is represented as a weighted least squares problem:

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

Linear Logistic Regression *

- Iteratively reweighted least squares (IRLS) algorithm

1. Initialize β
2. *Repeat*
3. Form linearized responses

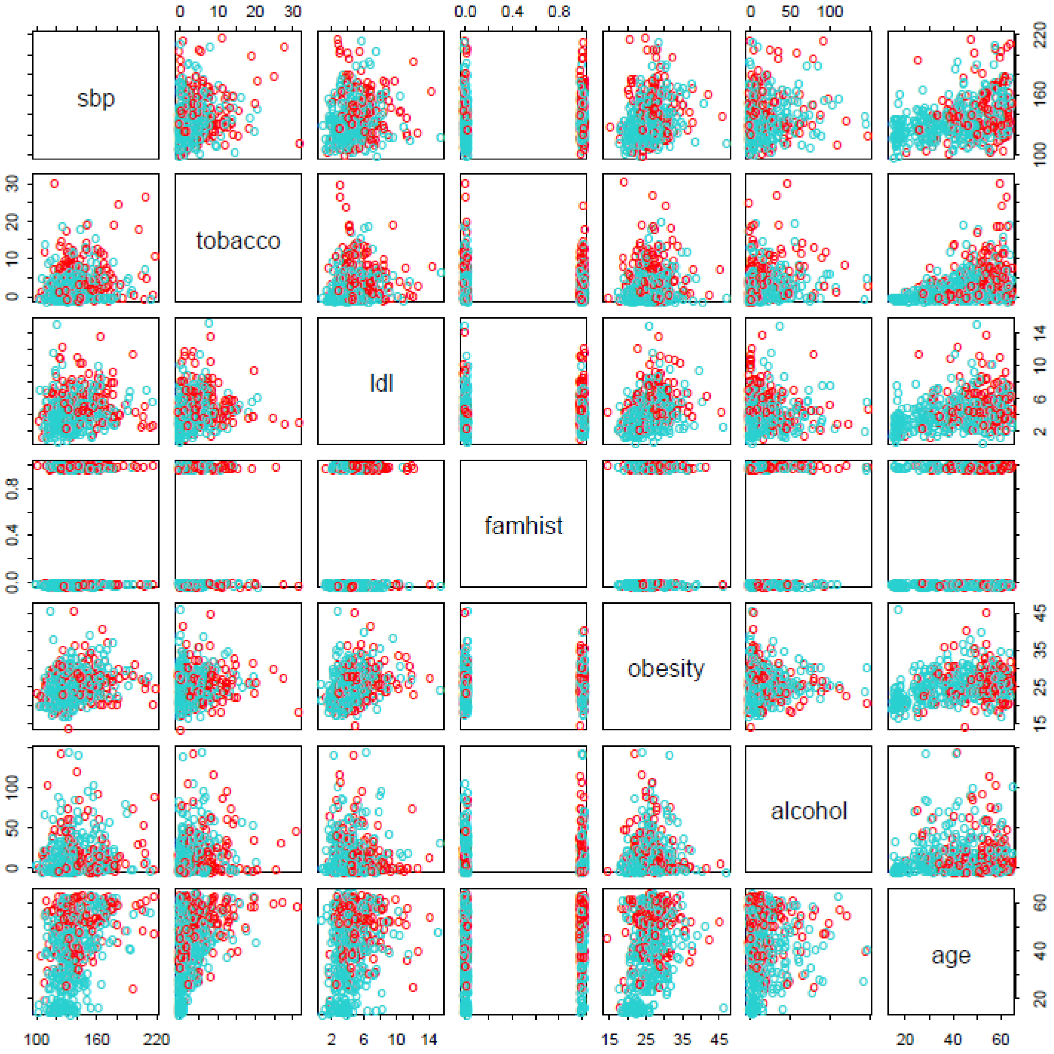
$$z_i = x_i^T\beta + \frac{y_i - p_i}{p_i(1 - p_i)}$$

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

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*

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

Linear Logistic Regression



Example: South African Heart Disease

- Red: 160 cases
- Green: 302 controls
- Z score measures the significance of a coefficient

	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

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

- 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?

Connection between LDA and Logistic Regression

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

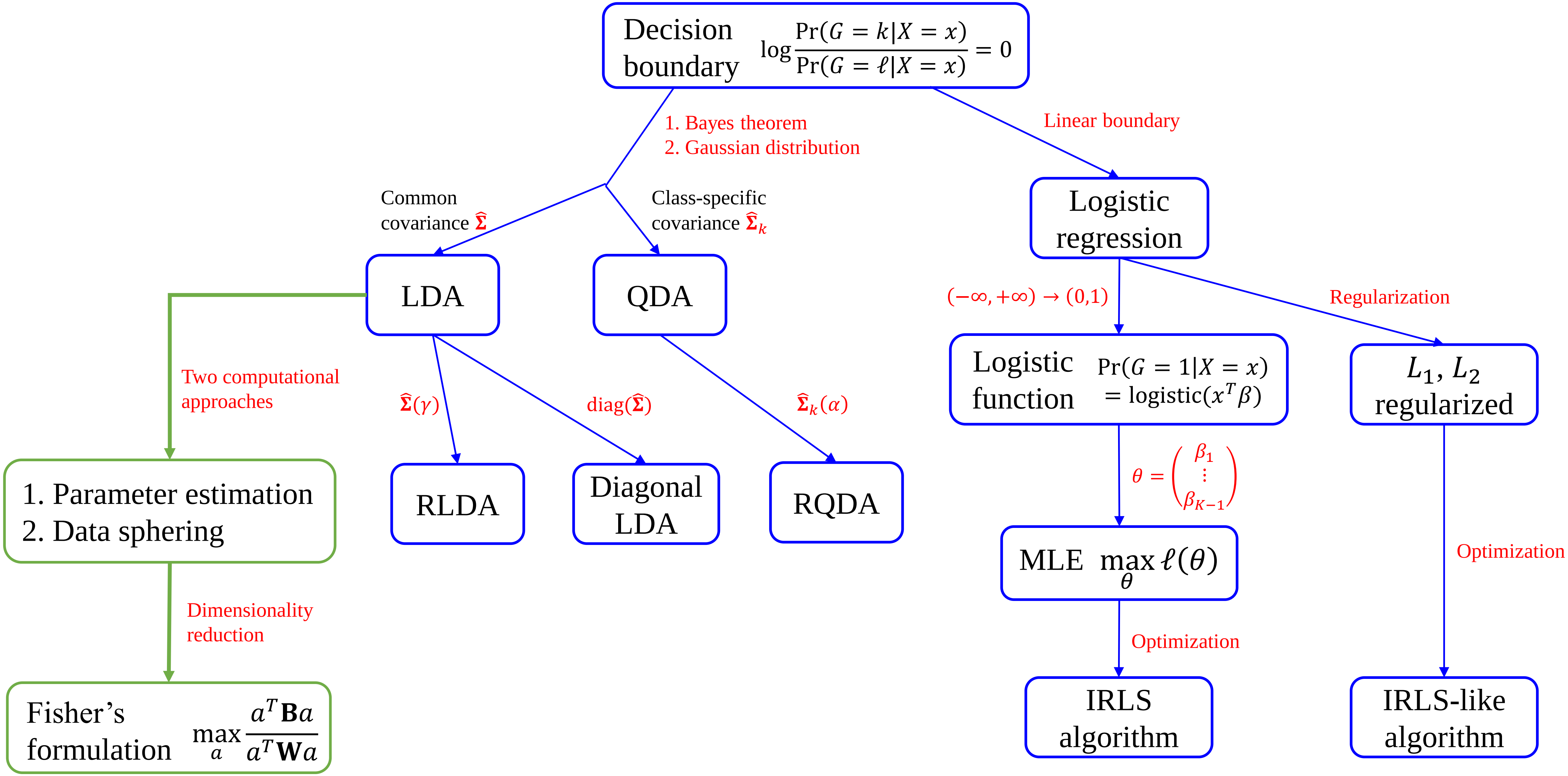
- If the additional assumption made by LDA is appropriate, LDA tends to estimate the parameters more efficiently.

- 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



Decision boundary $\log \frac{\Pr(G=k|X=x)}{\Pr(G=\ell|X=x)} = 0$
1. Bayes theorem
2. Gaussian distribution
Linear boundary
Common covariance $\hat{\boldsymbol{\Sigma}}$
Class-specific covariance $\hat{\boldsymbol{\Sigma}}_k$
Logistic regression
LDA
QDA
$(-\infty, +\infty) \rightarrow (0,1)$
Regularization
Two computational approaches
$\hat{\boldsymbol{\Sigma}}(\gamma)$
diag($\hat{\boldsymbol{\Sigma}}$)
$\hat{\boldsymbol{\Sigma}}_k(\alpha)$
Logistic function $\Pr(G=1|X=x) = \text{logistic}(x^T\beta)$
L_1, L_2 regularized
1. Parameter estimation
2. Data sphering
RLDA
Diagonal LDA
RQDA
$\theta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{K-1} \end{pmatrix}$
Optimization
MLE $\max_\theta \ell(\theta)$
Dimensionality reduction
Optimization
Fisher's formulation $\max_a \frac{a^T \mathbf{B} a}{a^T \mathbf{W} a}$
IRLS algorithm
IRLS-like algorithm