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 → 实质是分类
  - 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 正则化
  - Fisher's Formulation of Discriminant Analysis
- Logistic Regression
- Summary

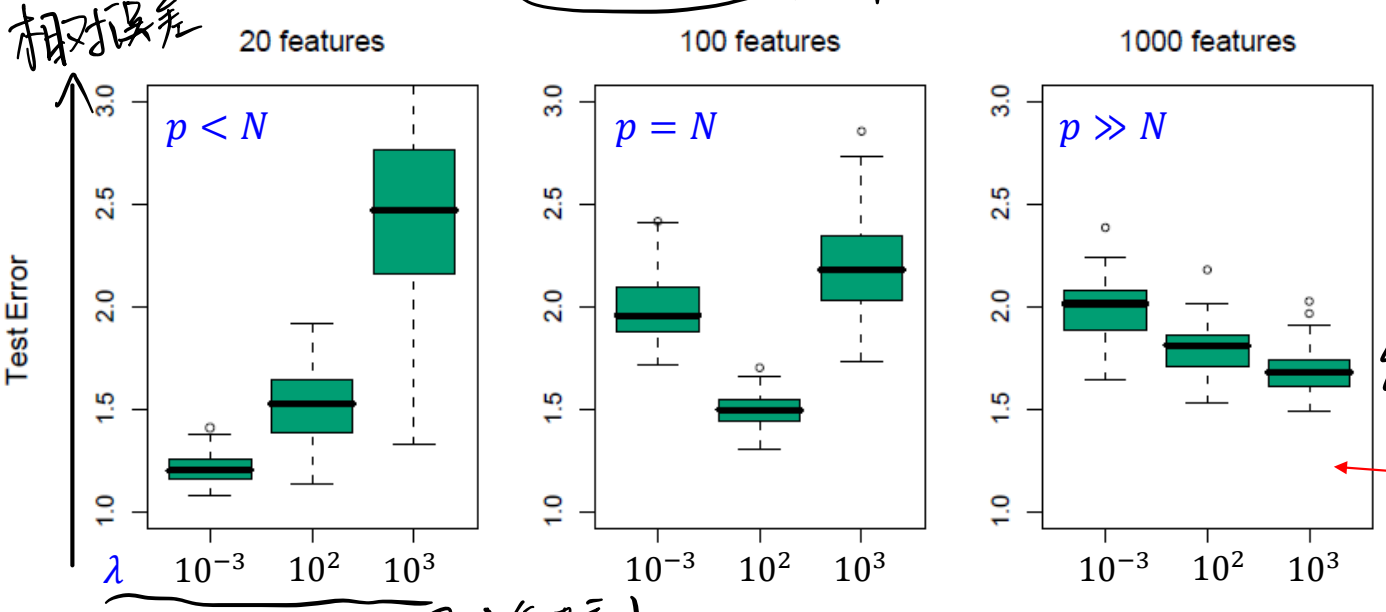
Fisher公式降维.

Regularized Discriminant Analysis

高维少数据
高奇异矩阵.

High dimensional problems ($p \gg N$)

- genomics problem, signal/image analysis
- Less fitting is better → 训练集小，防止过拟合.



Example

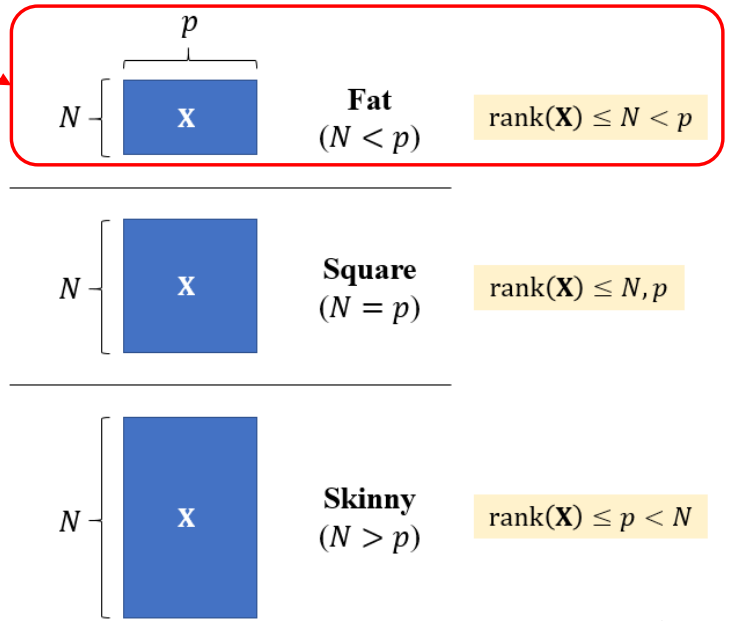
- 100 samples are generated by a linear model
- Ridge regression
- Relative error (divide by Bayes error)

维度越高模型越简单越好

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: $K \times p \times p \rightarrow K \times p$

Model complexity

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 ($\alpha, \gamma \in [0, 1]$)

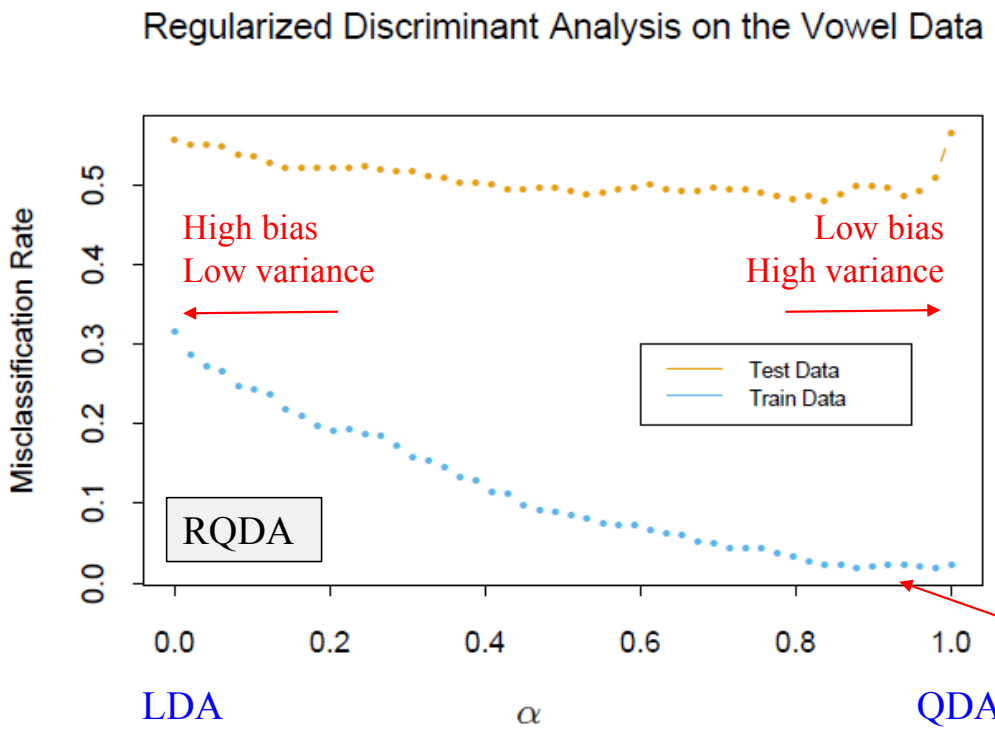
	Method	Covariance matrix	Effect
Linear	Regularized LDA (RLDA)	$\hat{\boldsymbol{\Sigma}}(\gamma) = \gamma\hat{\boldsymbol{\Sigma}} + (1-\gamma)\text{diag}(\hat{\boldsymbol{\Sigma}})$	Shrink $\hat{\boldsymbol{\Sigma}}$ towards $\text{diag}(\hat{\boldsymbol{\Sigma}})$
	Diagonal LDA	$\hat{\boldsymbol{\Sigma}} = \text{diag}(\hat{\boldsymbol{\Sigma}})$	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)

认为Σ_i相同(均为Σ)则令$\alpha=0$

认为Σ_i均不同则$\alpha=1$.

$\alpha\in(0,1)$表示介于两者之间

Regularized Discriminant Analysis



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

RQDA:

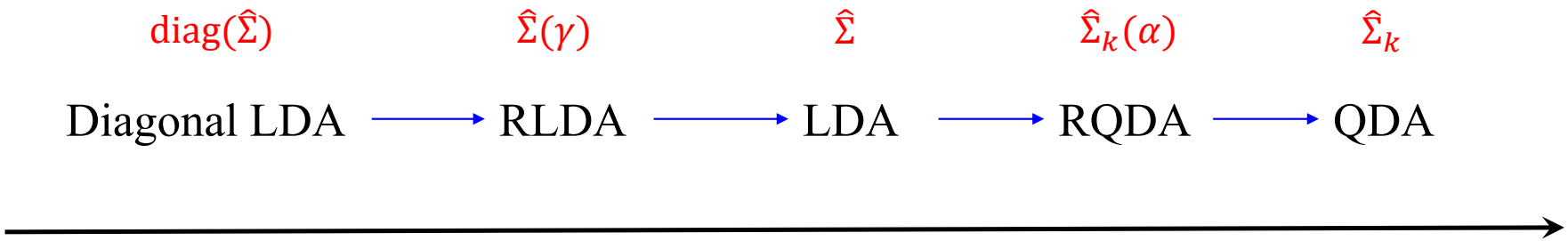
$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$

- $\alpha = 0$, 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

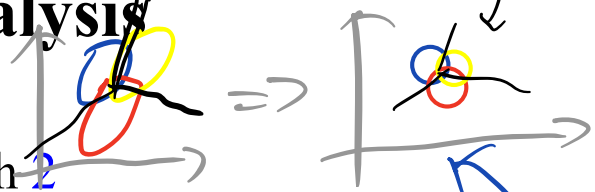


计算 LDA

要看处于哪个范围内.

只考虑距圆心距离即可

Fisher's Formulation of Discriminant Analysis



LDA: Approach 1

1. Estimating $\hat{\boldsymbol{\Sigma}}$, $\hat{\mu}_k$ and $\hat{\pi}_k$
2. Discriminant function

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

3. Classify to class k that maximizes the discriminant function

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

协方差矩阵

$$\Sigma = \frac{(x-\mu)(x-\mu)^T}{n-1}$$

求协

$$= \frac{\sum(\Sigma^{-\frac{1}{2}}x_i - \Sigma^{-\frac{1}{2}}\mu)(\Sigma^{-\frac{1}{2}}x_i - \Sigma^{-\frac{1}{2}}\mu)^T}{n-1}$$

LDA: Approach 2

1. Estimating $\hat{\boldsymbol{\Sigma}}$, $\hat{\mu}_k$ and $\hat{\pi}_k$
2. Eigen-decomposition: U:正交矩阵 D:对角

$$\hat{\boldsymbol{\Sigma}} = \mathbf{U}\mathbf{D}\mathbf{U}^T$$

3. Data sphering ($\hat{\Sigma}^* = \mathbf{I}$) 目的(球化)
   - $x^* = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T x = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}x$ 使协方差矩阵变成I
   - $\hat{\mu}_k^* = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T\hat{\mu}_k = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\hat{\mu}_k$
4. Classify to its closest class centroid in the transformed space

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

降下

$= \frac{\sum (x_i-\mu)(x_i-\mu)^T \Sigma^{-1}}{n-1} = I$

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. $\delta_k(x) \propto \log \Pr(G=k|X=x)$ ← $\Pr(G=k|X=x) = \frac{\Pr(X=x|G=k)\Pr(G=k)}{\Pr(X=x)}$

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

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

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

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

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

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

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 $\hat{\boldsymbol{\Sigma}}$, $\hat{\mu}_k$ and $\hat{\pi}_k$
2. Discriminant function

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

3. Classify to class k that maximizes the discriminant function

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

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

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

LDA: Approach 2

1. Estimating $\hat{\boldsymbol{\Sigma}}$, $\hat{\mu}_k$ and $\hat{\pi}_k$
2. Eigen-decomposition: $\hat{\boldsymbol{\Sigma}} = \mathbf{U}\mathbf{D}\mathbf{U}^T$
3. Data sphering ($\hat{\Sigma}^* = \mathbf{I}$)
   - $x^* = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T x = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} x$
   - $\hat{\mu}_k^* = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T \hat{\mu}_k = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \hat{\mu}_k$
4. Classify to its closest class centroid in the transformed space

$$\hat{G}(x) = \underset{k \in \mathcal{G}}{\operatorname{argmin}} \frac{1}{2} \|x^* - \hat{\mu}_k^*\|^2 - \ln \hat{\pi}_k$$

可以在这里保留一些维度

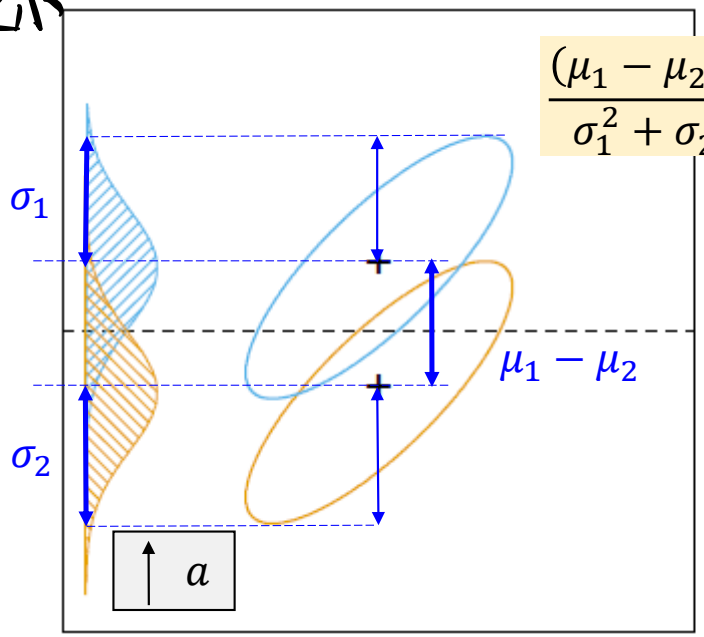
Fisher's Formulation of Discriminant Analysis

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

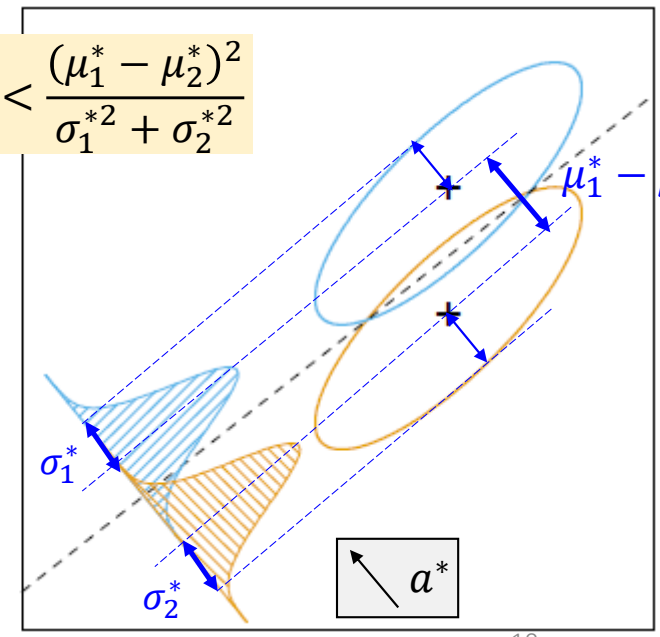
Our goal: $\max_a \dfrac{(\mu_1^* - \mu_2^*)^2}{\sigma_1^{*2} + \sigma_2^{*2}}$

Between class variance

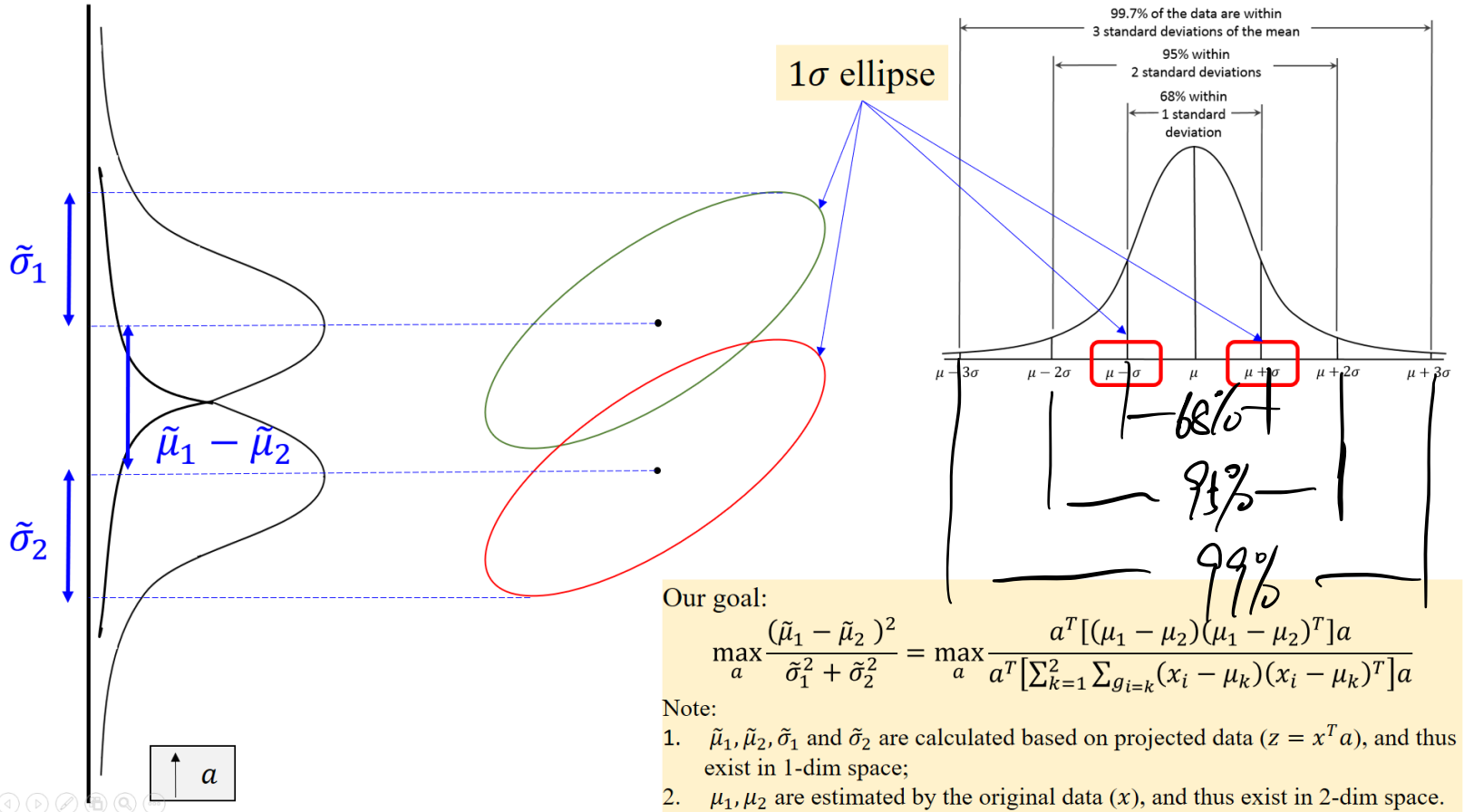
Within class variance



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



Fisher's Formulation of Discriminant Analysis



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.

$\tilde{\mu}, \tilde{\mu}$：投影后的是一个点。 $z_i = x_i' a$. $\hat{\mu} = \mu' a$

$\tilde{\mu}, \tilde{\mu} \in \mathbb{R}^1$ $\mu \in \mathbb{R}^p$ $\tilde{\sigma}^2 = a^T \Sigma a$

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

Fisher's Formulation of Discriminant Analysis

- Maximize the Rayleigh quotient:

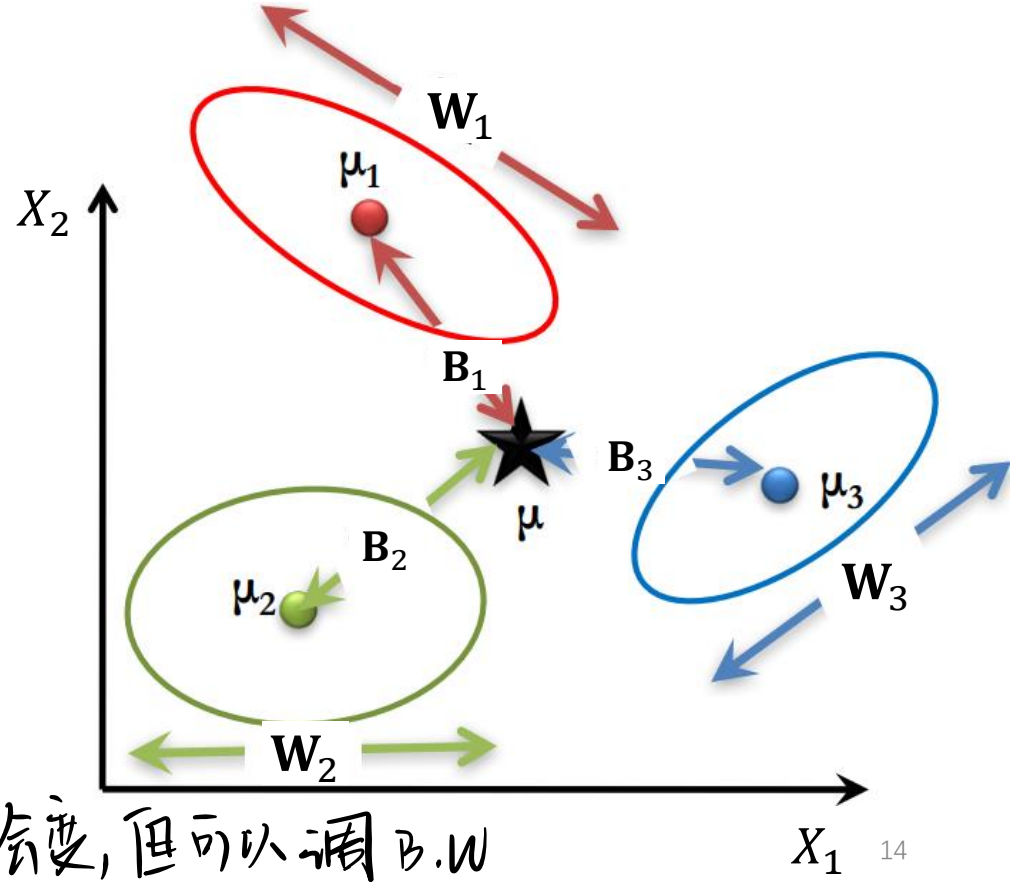
$$\max_a \frac{a^T \mathbf{B} a}{a^T \mathbf{W} a}$$

  - Between class variance

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

  - Within class variance

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



给定数据集后，T=B+W 不会变，但可以调 B.W

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,

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

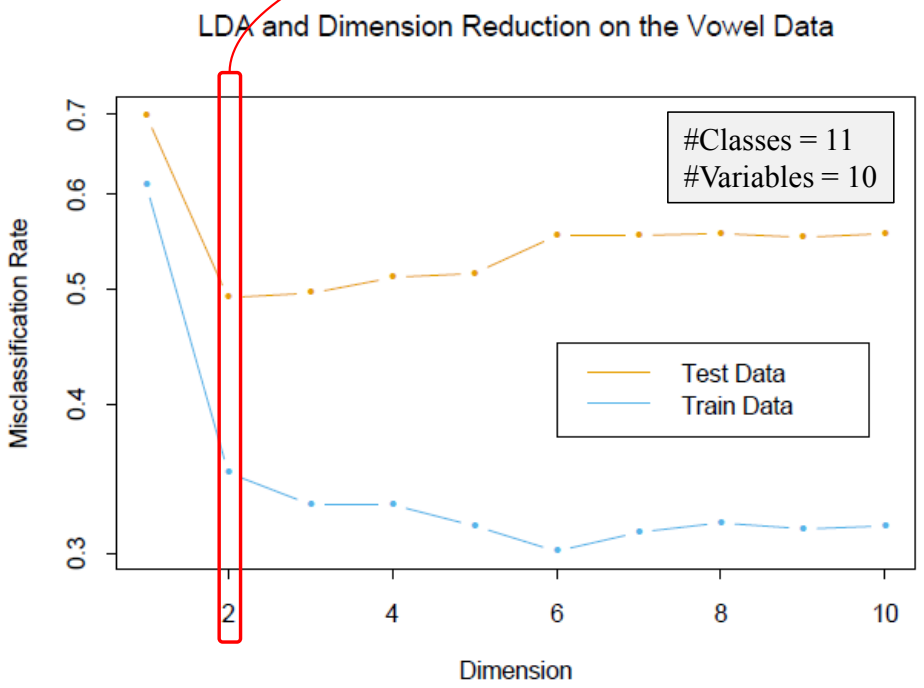
  - a is discriminant coordinates (canonical variates)
  - Generalized eigenvalue problem

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

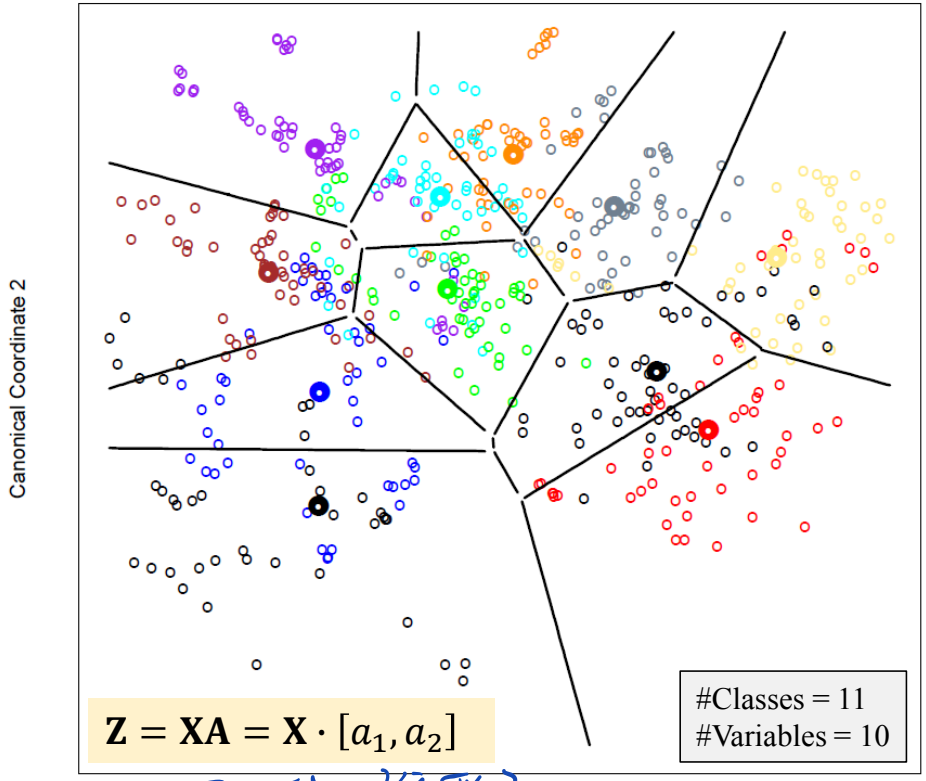
  which can be efficiently solved

Ex. 4.1.
Hint: Lagrangian multipliers

Fisher's Formulation of Discriminant Analysis



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



(有一些特征是无效的)

Linear Methods for Classification II

LDA

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

估计出后可以采样

$Pr(X,Y)$

也: LDA, NB

Native Bayes

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

生成式方法 Generative

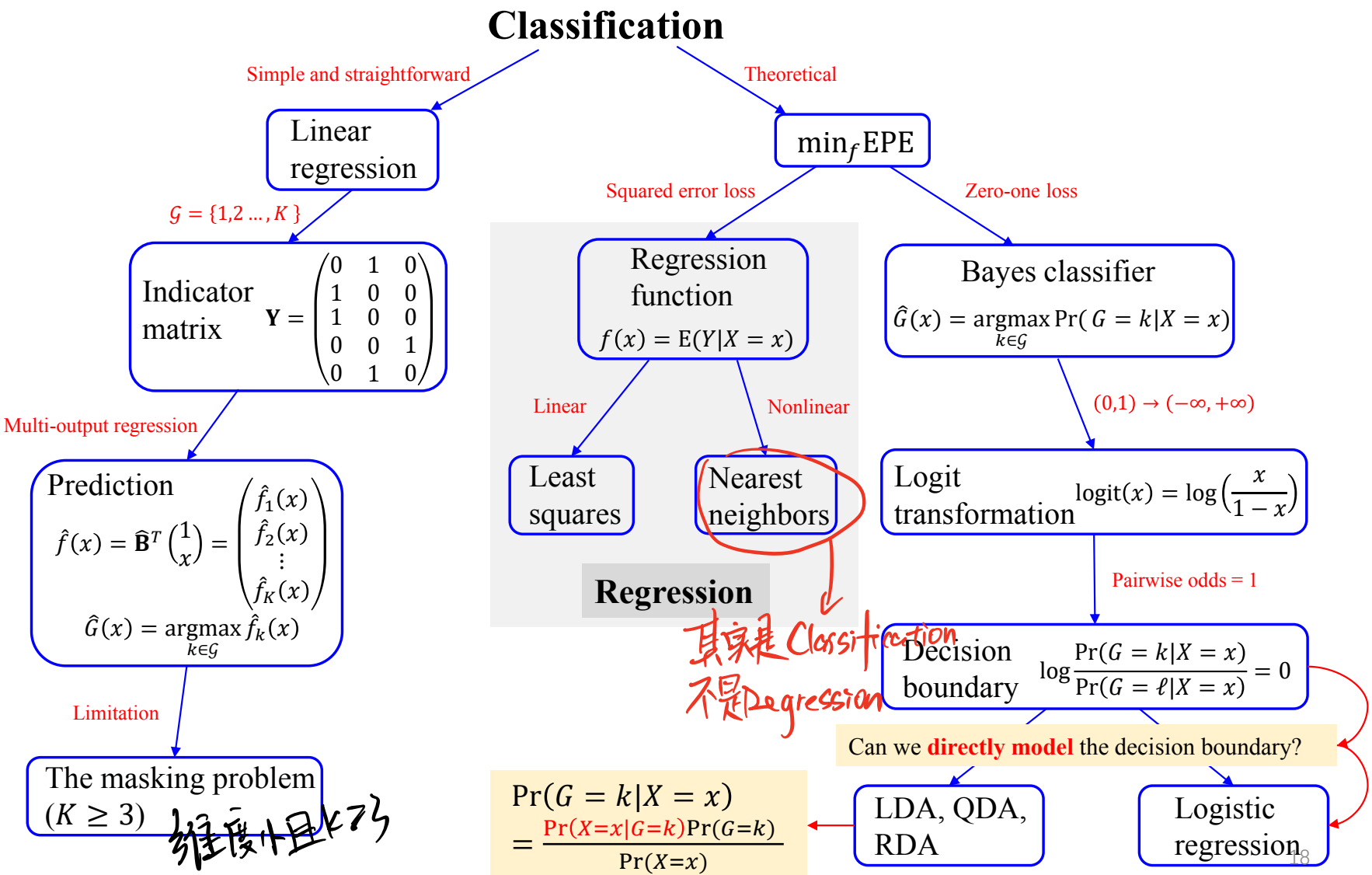
判别式方法: discrimative

$Pr(Y|X)$

Ridge, Logistic, LS

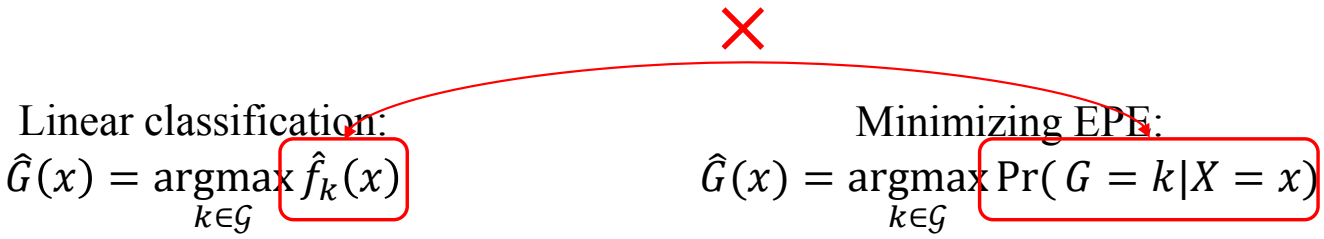
岭回归

Least Square.


Classification
Simple and straightforward
Theoretical
Linear regression
$\min_f \text{EPE}$
$\mathcal{G} = \{1,2 \ldots, K\}$
Squared error loss
Zero-one loss
Indicator matrix
$\mathbf{Y} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$
Regression function
$f(x) = \text{E}(Y|X = x)$
Bayes classifier
$\hat{G}(x) = \underset{k \in \mathcal{G}}{\text{argmax}} \Pr(G = k|X = x)$
Multi-output regression
Linear
Nonlinear
$(0,1) \to (-\infty, +\infty)$
Prediction
$\hat{f}(x) = \hat{\mathbf{B}}^T \begin{pmatrix} 1 \\ x \end{pmatrix} = \begin{pmatrix} \hat{f}_1(x) \\ \hat{f}_2(x) \\ \vdots \\ \hat{f}_K(x) \end{pmatrix}$
$\hat{G}(x) = \underset{k \in \mathcal{G}}{\text{argmax}} \hat{f}_k(x)$
Least squares
Nearest neighbors
Logit transformation
$\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$
Regression
Pairwise odds = 1
其实是Classification
不是Regression
Decision boundary
$\log \frac{\Pr(G = k|X = x)}{\Pr(G = \ell|X = x)} = 0$
Limitation
Can we directly model the decision boundary?
The masking problem ($K \geq 3$)
维度小且K多
$\Pr(G = k|X = x) = \frac{\Pr(X=x|G=k)\Pr(G=k)}{\Pr(X=x)}$
LDA, QDA, RDA
Logistic regression

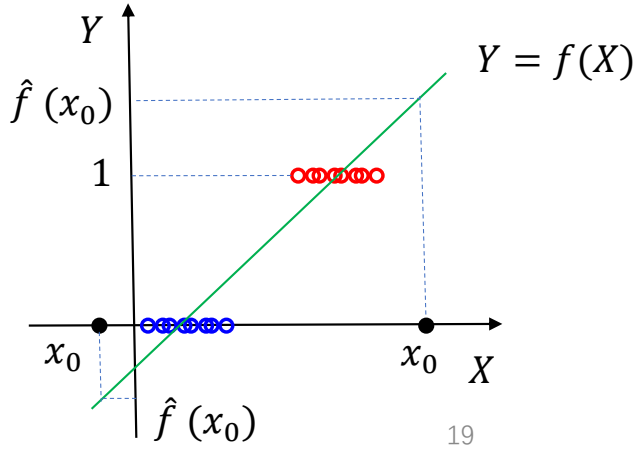
Linear Discriminant Analysis

- Recall our discussion on linear regression of an indicator matrix



- 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

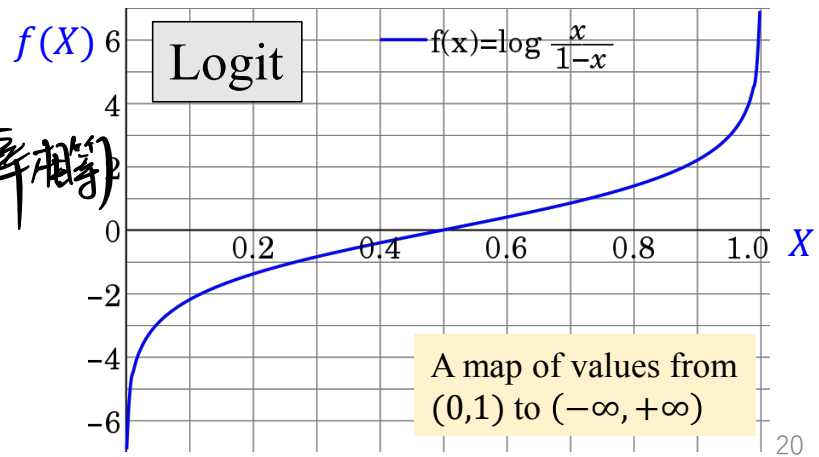
- Logit transform

Odds (发生比)

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

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

- Decision boundary
  - Odds equals to 1 → 发生 = 不发生，故在边界上（概率相等）
  - Or, logit equals to 0



Linear Discriminant Analysis

Sigmoid 函数 $\sigma(x) = \frac{1}{1+e^{-x}}$ 将 $(-\infty, +\infty)$ 映射到 $(0,1)$
(也是 Logistic 函数)
↳ 看作概率.

- 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

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

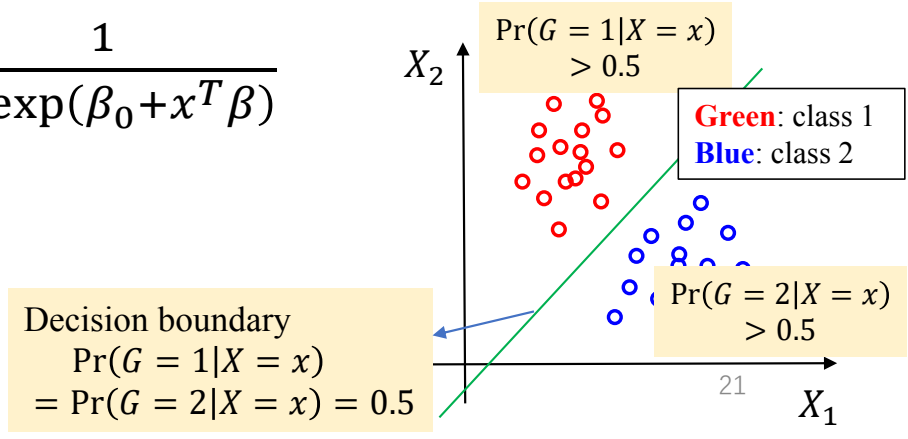
$\exp(x) = e^x$

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

- Decision boundary

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

$\sigma(-x) = 1-\sigma(x)$ $\quad \sigma'(x) = \sigma(x)(1-\sigma(x))$ 但导数太小



Linear Logistic Regression

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

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

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

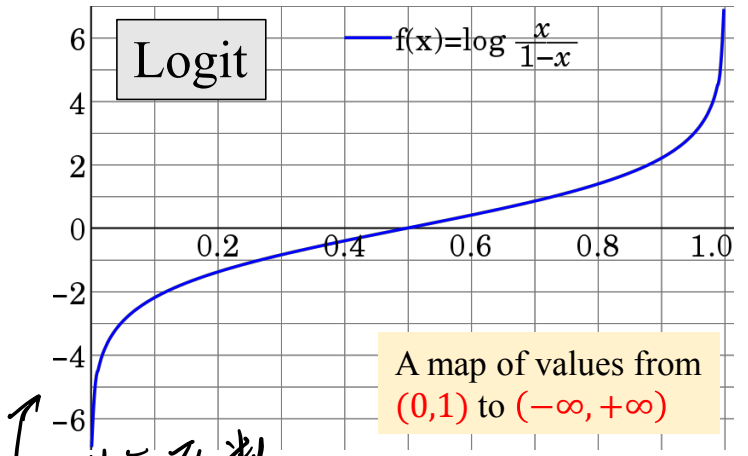
$$\vdots$$

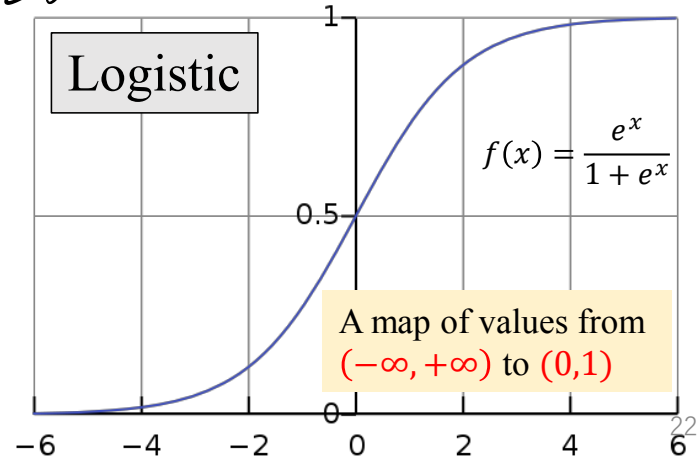
$$\log\frac{\Pr(G=K-1|X=x)}{\Pr(G=K|X=x)} = \beta_{(K-1)0} + x^T\beta_{K-1}$$

- $K-1$ 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

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

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

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

- 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

- Estimating parameter set $\theta = \{\beta_{10}, \beta_1, \ldots, \beta_{(K-1)0}, \beta_{K-1}\}$
  - 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	$p(x; \theta)$	$1 - p(x; \theta)$

熵：$H(p) = -\int p(x) \log p(x)\, dx$　交叉熵：$H(p,q) = -\int p(x) \log q(x)\, dx$.

Linear Logistic Regression

- Two classes

这里取 Cross entropy，$p(x)$指观测的 y，$q(x)$为自己算得的

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

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

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

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

$$x_i \leftarrow \begin{pmatrix} 1 \\ x_i \end{pmatrix} \qquad \beta \leftarrow \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix}$$

没有闭式解 ⇒ 只能逼近

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

过程：① $\ln \frac{Pr(G=1|X=x)}{Pr(G=2|X=x)} = 0 = \beta^T X$

② $Pr(G=1|X=x) = \frac{1}{1+e^{-\beta^T X}} = \sigma(\beta_1^T X)$

$Pr(G=2|X=x) = \sigma(\beta_2^T X) = 1 - \sigma(\beta_1^T X) = \sigma(-\beta_1^T X)$

③ $Pr(Y=y|X=x) = \sigma^y(\beta^T X)(1-\sigma(\beta^T X))^{1-y}$

④ MLE $\max_\beta \ell(\beta) = \sum_{i=1}^n \ln Pr(Y_i|X_i)$

若 $y \in \{1, -1\}$ 则可简化：③ $Pr(Y=y|X=x) = \sigma(y\beta^T X)$

④ MLE：$\ell(\beta) = \sum_{i=1}^n \ln \frac{1}{1+e^{-y\beta^T X}} = -\sum_{i=1}^n \ln(1+e^{-y\beta^T X})$

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 $p(x_i)(1 - p(x_i))$

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

- The Newton-Raphson step:

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

- Given the response

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

- it is represented as a weighted least squares problem:

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

Linear Logistic Regression *

迭代加权最小二乘法.

- Iteratively reweighted least squares (IRLS) algorithm

随机或用 Least Square 简单计算一次.

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

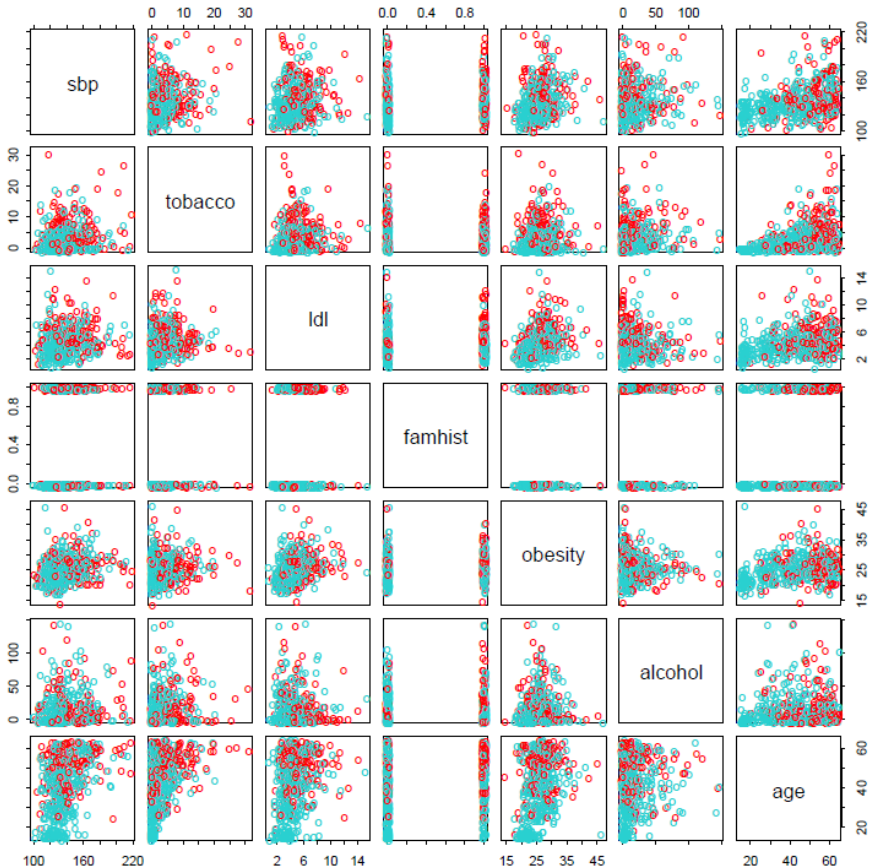
$$z_i = x_i^T \beta + \frac{y_i - p_i}{p_i(1-p_i)}$$

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

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

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

超参数.

↳ 无 β_0

- 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?

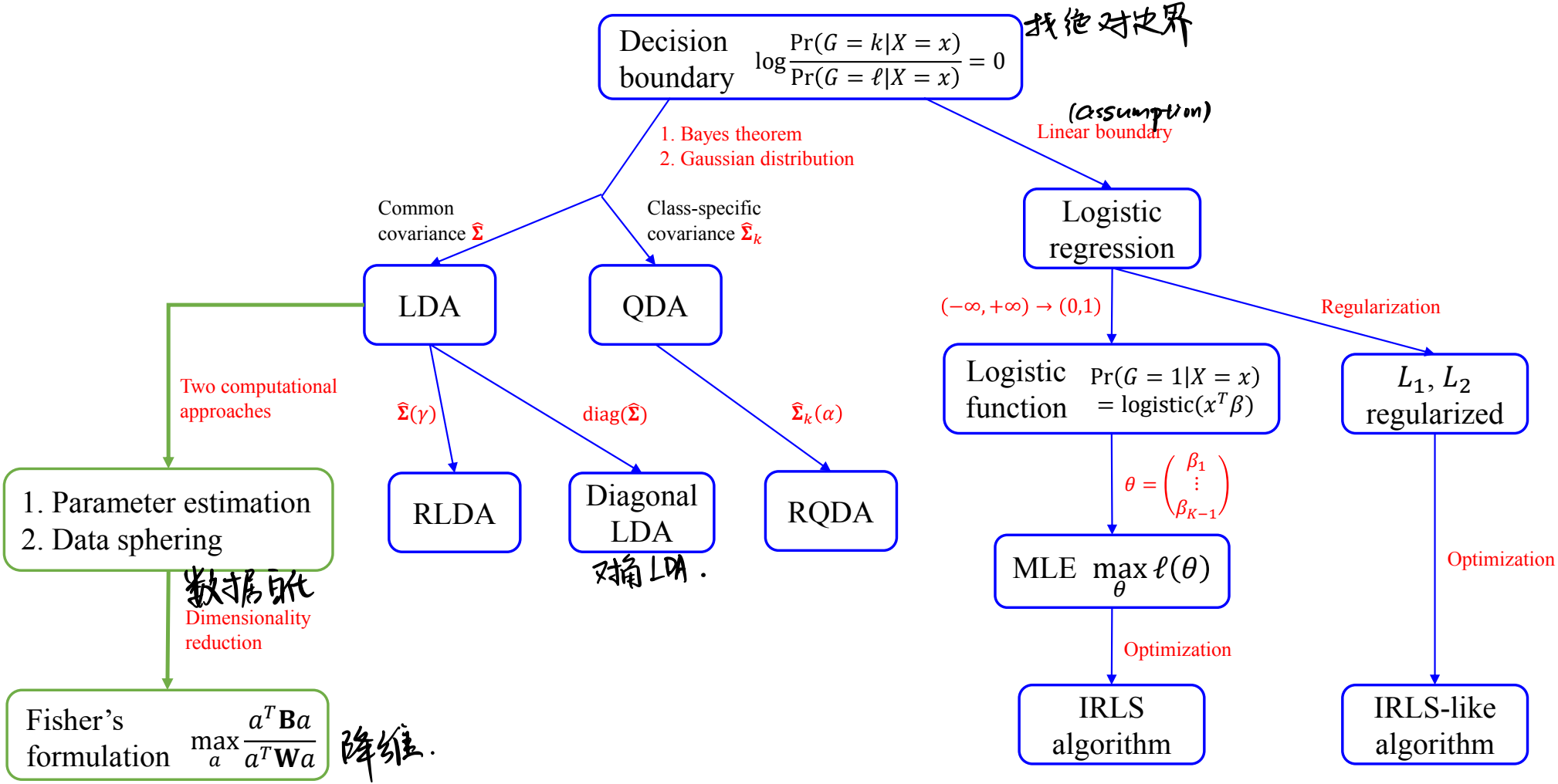
↳ 加权岭回归

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 $\log \frac{\Pr(G=k|X=x)}{\Pr(G=\ell|X=x)} = 0$
找他对边界
1. Bayes theorem
2. Gaussian distribution
Linear boundary
(assumption)
Common covariance $\hat{\Sigma}$
Class-specific covariance $\hat{\Sigma}_k$
LDA
QDA
Logistic regression
Two computational approaches
$\hat{\Sigma}(\gamma)$
diag($\hat{\Sigma}$)
$\hat{\Sigma}_k(\alpha)$
$(-\infty, +\infty) \to (0,1)$
Regularization
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
对角LDA.
RQDA
$\theta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{K-1} \end{pmatrix}$
MLE $\max_\theta \ell(\theta)$
Optimization
数据白化
Dimensionality reduction
Optimization
Fisher's formulation $\max_a \frac{a^T \mathbf{B} a}{a^T \mathbf{W} a}$
降维.
IRLS algorithm
IRLS-like algorithm