

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

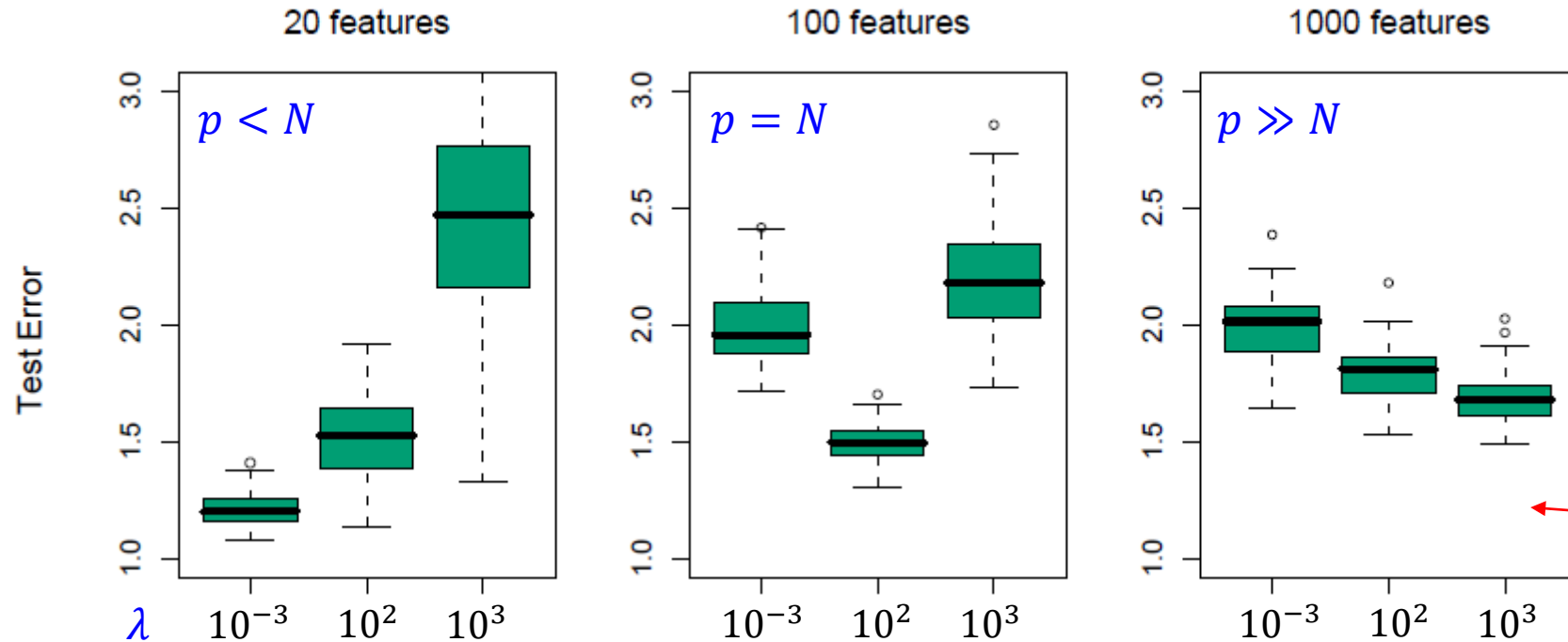
Regularized Discriminant Analysis

High dimensional problems ($p \gg N$)

- genomics problem, signal/image analysis
- **Less fitting** is better

Example

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

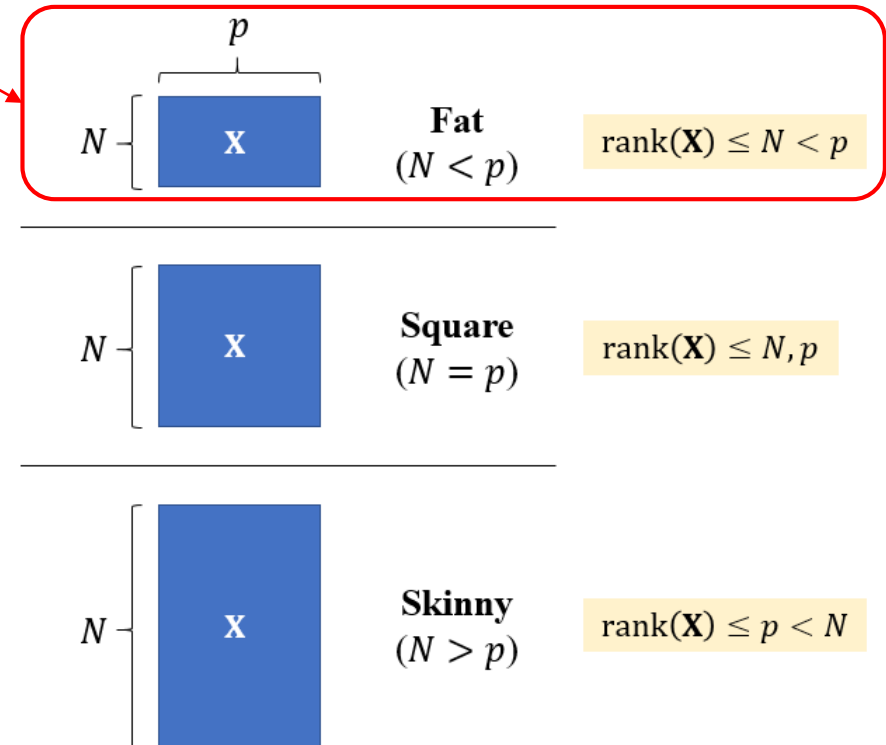


No enough information to estimate the high-dimensional covariance matrix

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 $\text{rank}(\Sigma) < N \ll p$
- Regularization is necessary
 - No enough data to estimate feature dependencies
 - E.g., independent assumption on features
 - Diagonal within-class covariance matrix



Model complexity

Regularized Discriminant Analysis

Regularized LDA (RLDA)

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

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

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

Diagonal LDA

- Independent assumption on feature dependencies

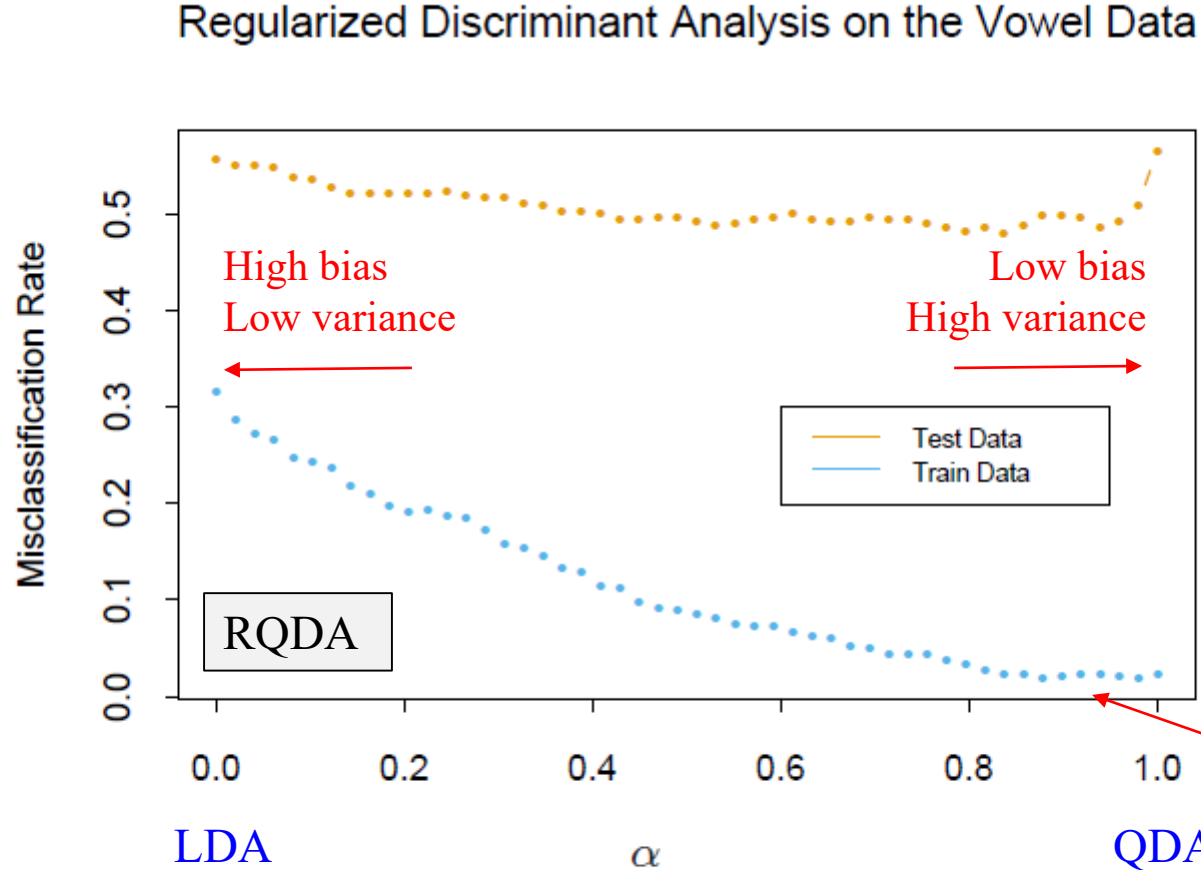
$$\hat{\Sigma} = \text{diag}(\hat{\Sigma})$$

Regularized Discriminant Analysis

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

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

Regularized Discriminant Analysis



RQDA:

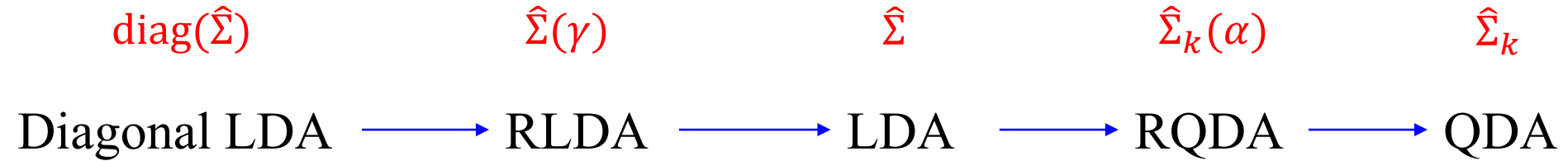
$$\hat{\Sigma}_k(\alpha) = \alpha \hat{\Sigma}_k + (1 - \alpha) \hat{\Sigma}$$

- $\alpha = 0$, 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



High bias
Low variance

Low bias
High variance

Fisher's Formulation of Discriminant Analysis

LDA: Approach 1

1. Estimating $\hat{\Sigma}$, $\hat{\mu}_k$ and $\hat{\pi}_k$
2. Discriminant function
$$\delta_k(x) = x^T \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log \hat{\pi}_k$$
3. Classify to class k that maximizes the discriminant function
$$\hat{G}(x) = \operatorname{argmax}_{k \in \mathcal{G}} \delta_k(x)$$

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

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

LDA: Approach 2

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

Fisher's Formulation of Discriminant Analysis

LDA: Approach 1

1. Estimating $\hat{\Sigma}$, $\hat{\mu}_k$ and $\hat{\pi}_k$
2. Discriminant function
$$\delta_k(x) = x^T \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log \hat{\pi}_k$$
3. Classify to class k that maximizes the discriminant function

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

LDA: Approach 2

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

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

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

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

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

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

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

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

Fisher's Formulation of Discriminant Analysis

LDA: Approach 1

1. Estimating $\hat{\Sigma}$, $\hat{\mu}_k$ and $\hat{\pi}_k$
2. Discriminant function
$$\delta_k(x) = x^T \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log \hat{\pi}_k$$
3. Classify to class k that maximizes the discriminant function
$$\hat{G}(x) = \operatorname{argmax}_{k \in \mathcal{G}} \delta_k(x)$$

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

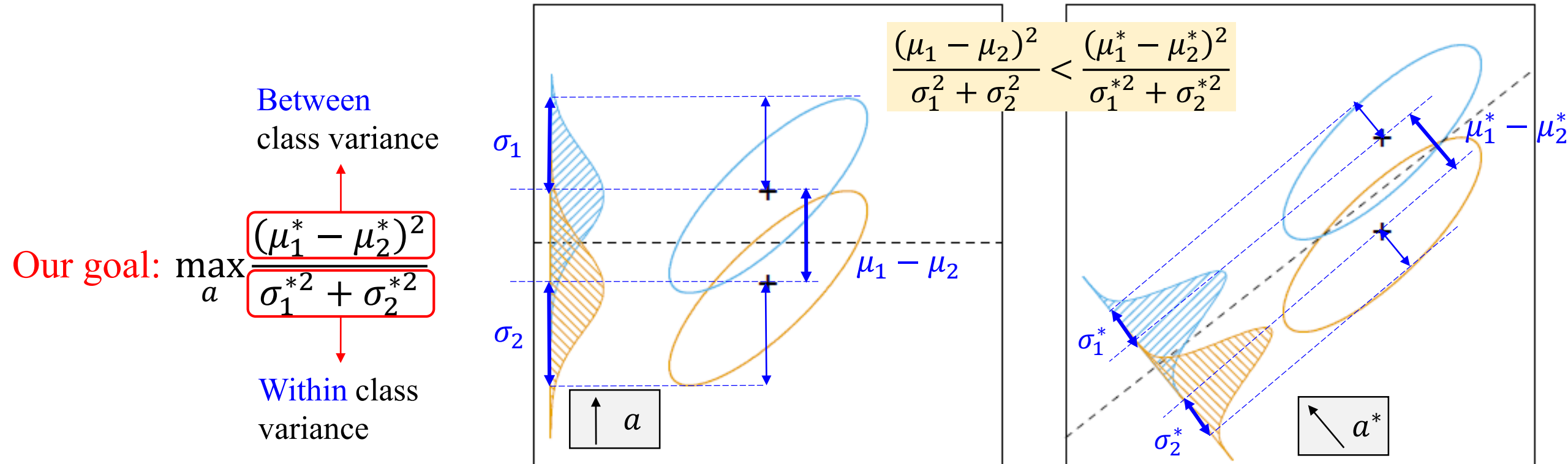
- Two approaches have almost the **same** time and storage complexity
- Approach 2 shows the potential of LDA for **dimension reduction**

LDA: Approach 2

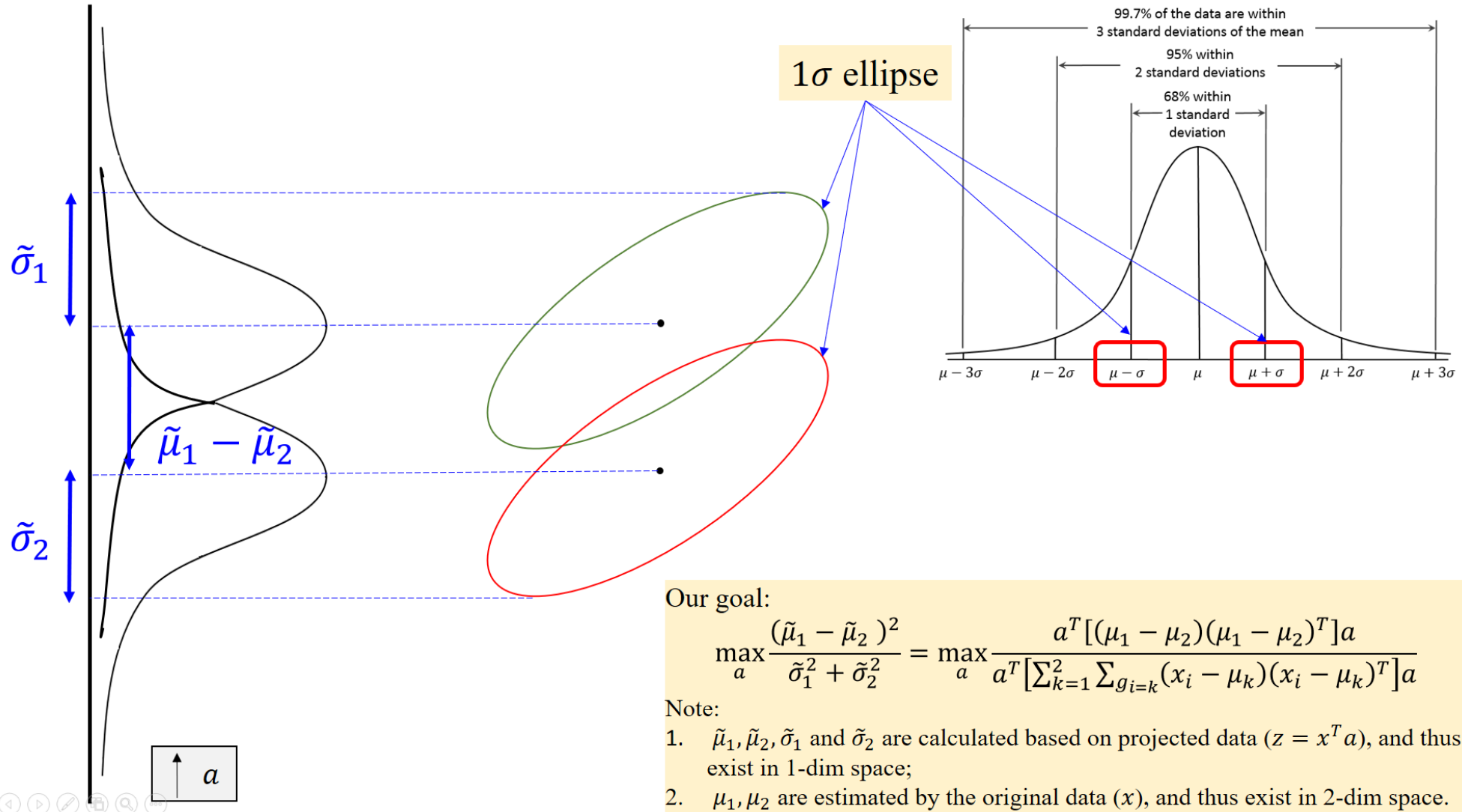
1. Estimating $\hat{\Sigma}$, $\hat{\mu}_k$ and $\hat{\pi}_k$
2. Eigen-decomposition:
$$\hat{\Sigma} = \mathbf{U} \mathbf{D} \mathbf{U}^T$$
3. Data sphering ($\hat{\Sigma}^* = \mathbf{I}$)
 - $x^* = \mathbf{D}^{-\frac{1}{2}} \mathbf{U}^T x = \hat{\Sigma}^{-\frac{1}{2}} x$
 - $\hat{\mu}_k^* = \mathbf{D}^{-\frac{1}{2}}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

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.



Fisher's Formulation of Discriminant Analysis



Fisher's Formulation of Discriminant Analysis

- Maximize the **Rayleigh quotient**:

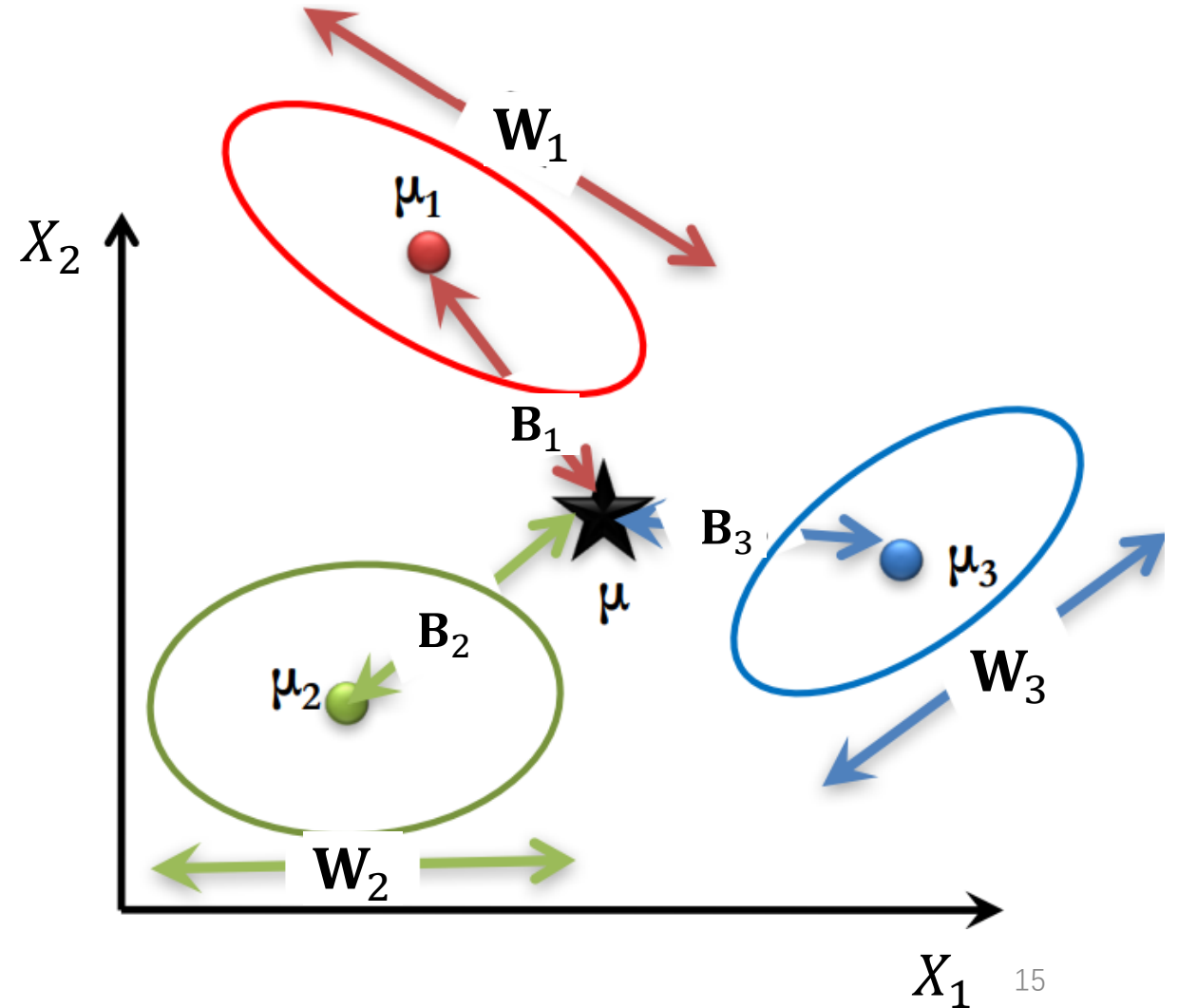
$$\max_a \frac{a^T \mathbf{B} a}{a^T \mathbf{W} a}$$

- 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,

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

- a is discriminant coordinates (canonical variates)

- **Generalized eigenvalue problem**

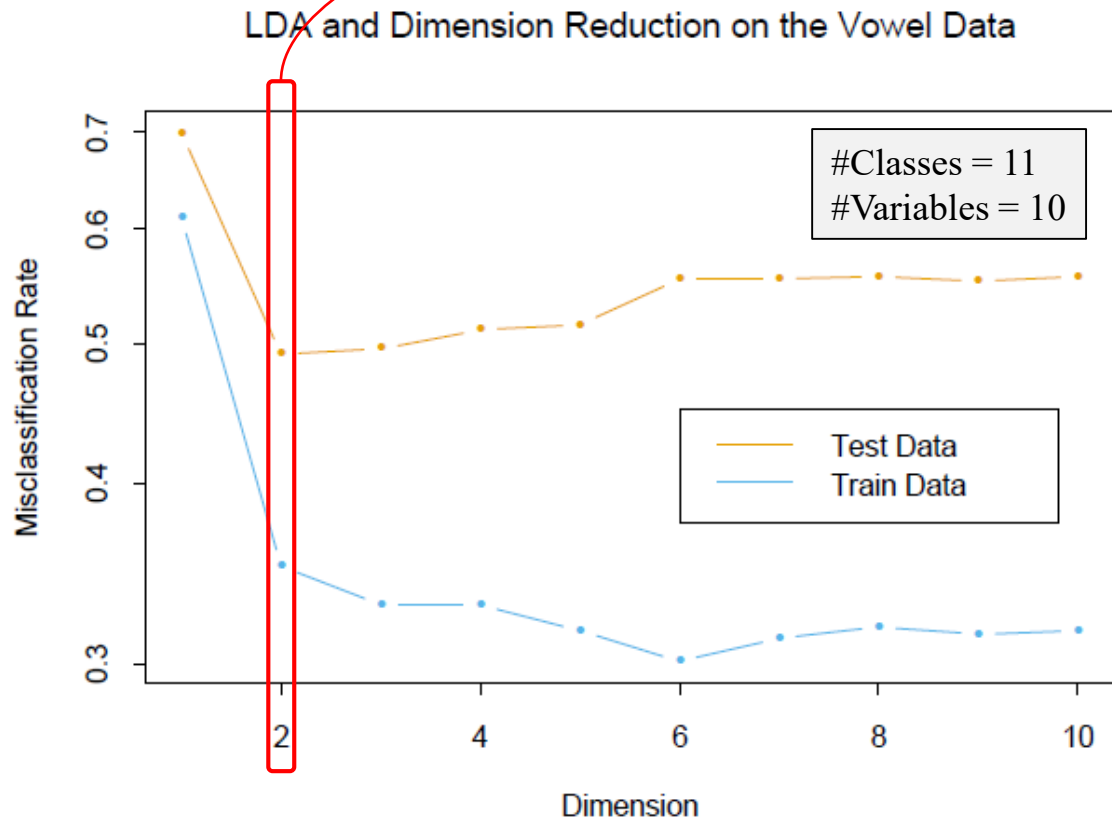
$$\mathbf{B} a = \lambda \mathbf{W} a$$

which can be efficiently solved

Ex. 4.1.

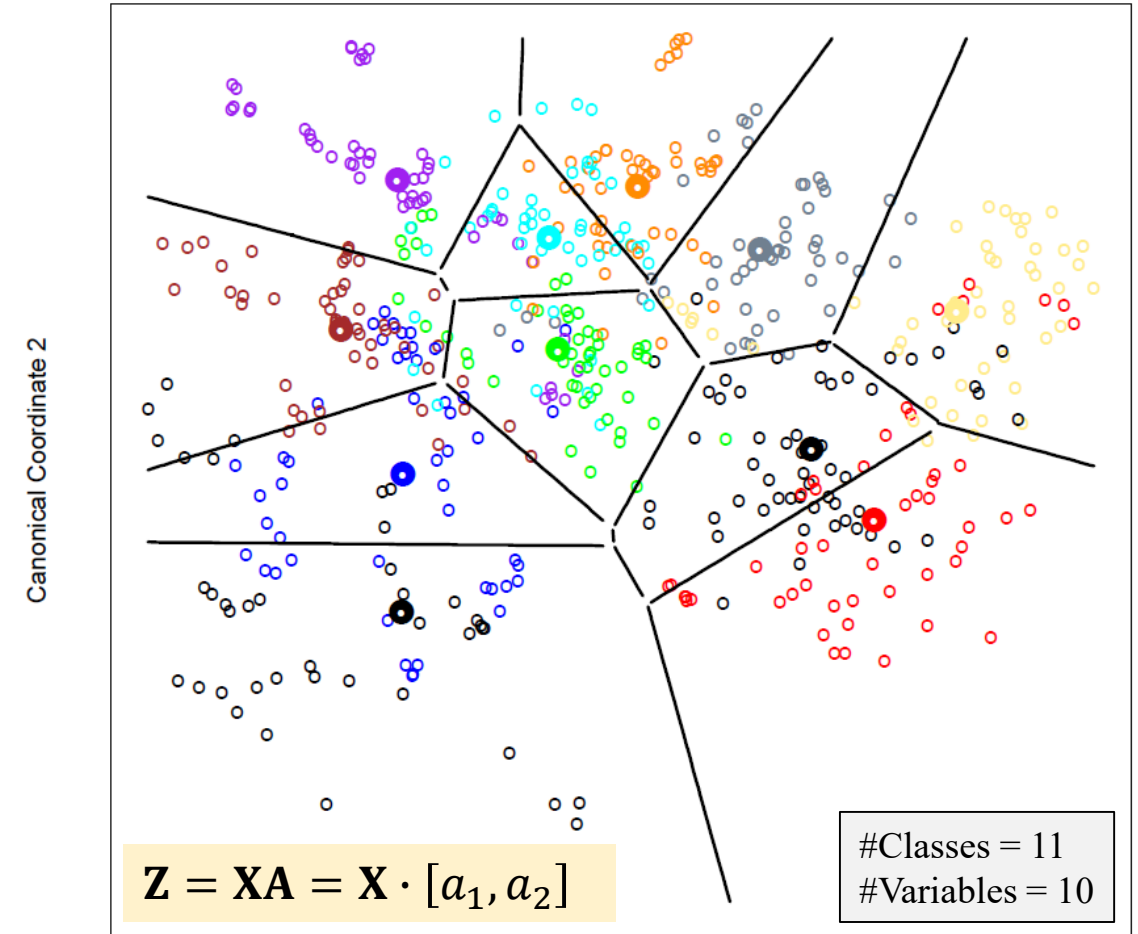
Hint: Lagrangian multipliers

Fisher's Formulation of Discriminant Analysis



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

Classification in Reduced Subspace



Fisher's Formulation of Discriminant Analysis

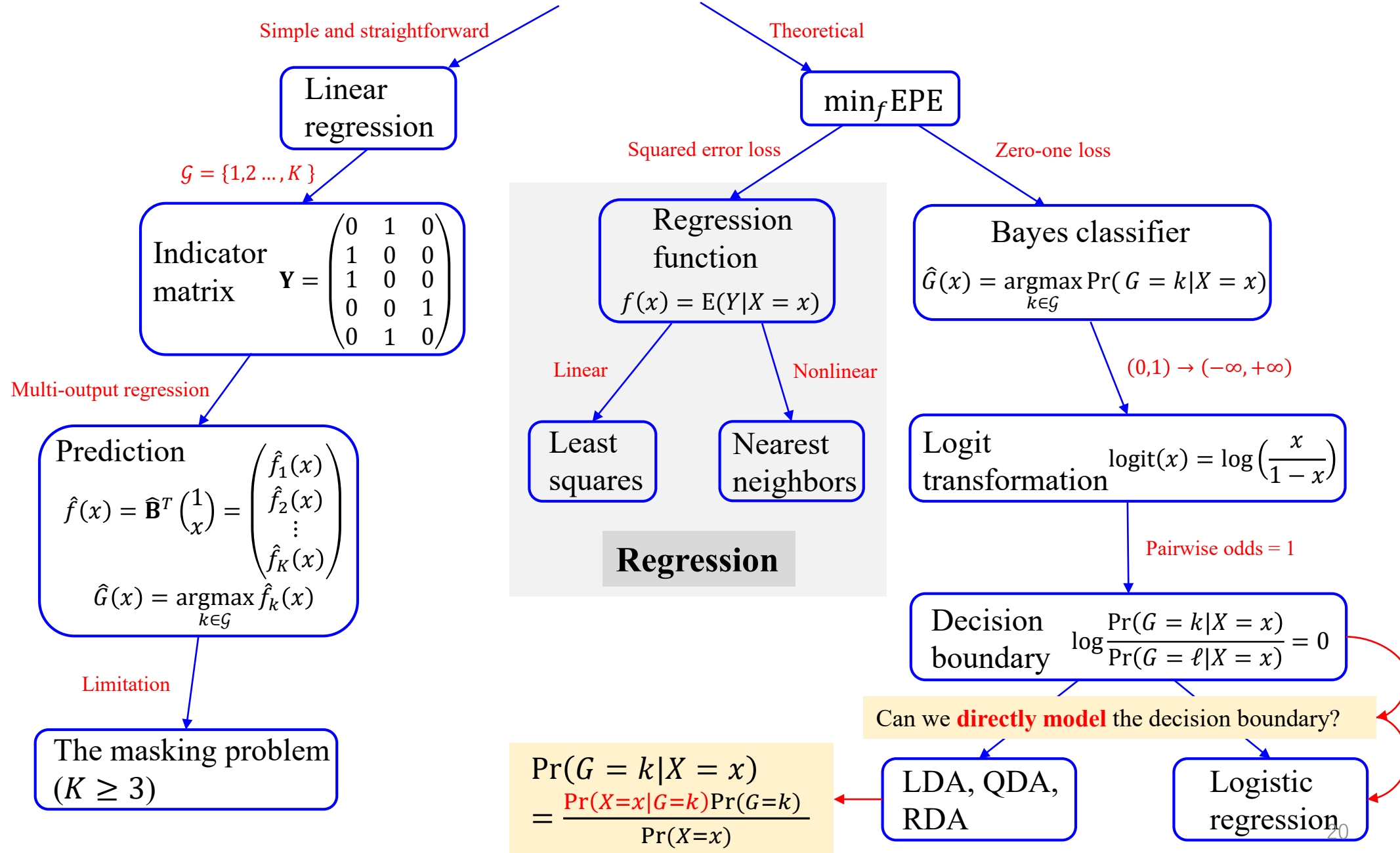
We will discuss an example in our **live lecture** according to:

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



Linear Logistic Regression

- **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

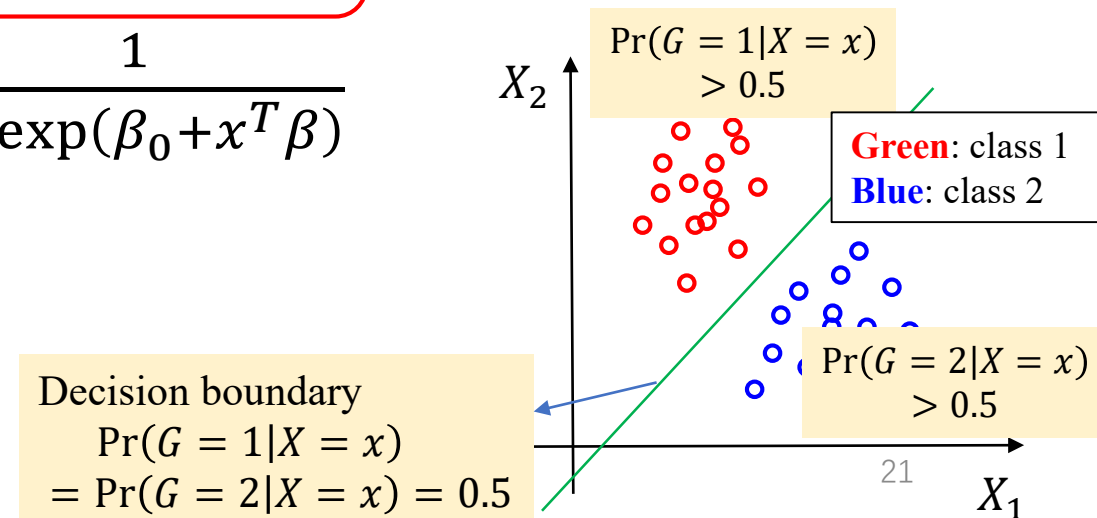
$$\Pr(G = 1|X = x) = \frac{\exp(\beta_0 + x^T \beta)}{1 + \exp(\beta_0 + x^T \beta)}$$

$\text{logistic}(x^T \beta)$
 $(-\infty, +\infty) \rightarrow (0,1)$

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

- Decision boundary

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



Linear Logistic Regression

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

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

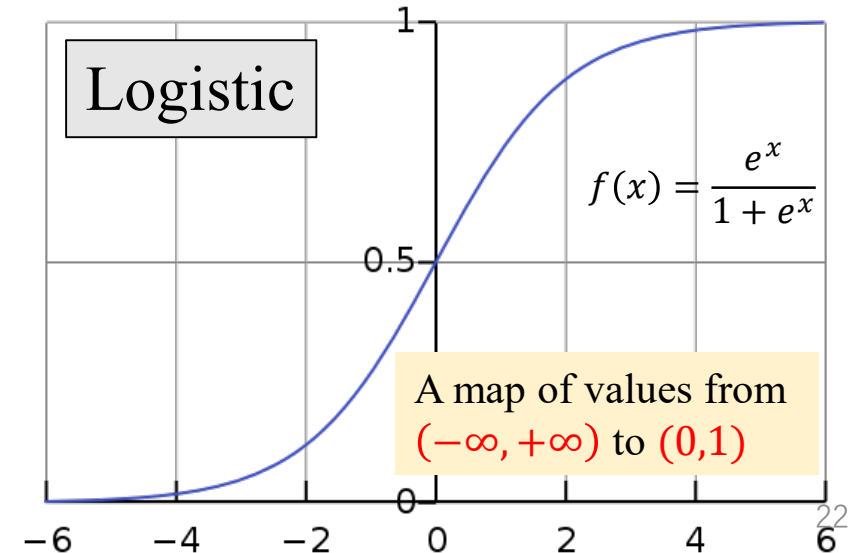
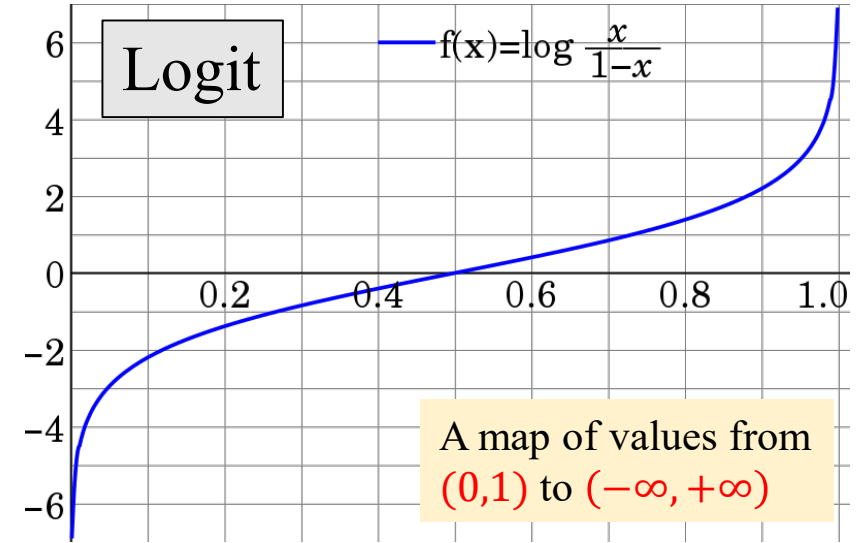
$$\log \frac{\Pr(G = 2|X = x)}{\Pr(G = K|X = x)} = \beta_{20} + x^T \beta_2$$

\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



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

?



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

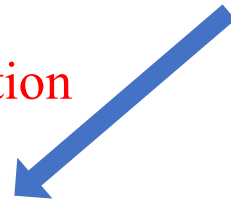
Linear Logistic Regression

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



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

summation



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

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



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



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

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

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

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^N \{y_i \log p(x_i; \theta) + (1 - y_i) \log(1 - p(x_i; \theta))\} \\ &= \sum_{i=1}^N \left\{ y_i \left[x_i^T \beta - \log(1 + e^{x_i^T \beta}) \right] - (1 - y_i) \log(1 + e^{x_i^T \beta}) \right\} \\ &= \sum_{i=1}^N \left\{ y_i x_i^T \beta - \log(1 + e^{x_i^T \beta}) \right\}\end{aligned}$$

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

Linear Logistic Regression

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

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

- The *second* derivative (**Hessian**)

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

- In matrix form

$$\begin{aligned}\frac{\partial \ell(\beta)}{\partial \beta} &= \mathbf{X}^T (\mathbf{y} - \mathbf{p}) \\ \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} &= -\mathbf{X}^T \mathbf{W} \mathbf{X}\end{aligned}$$

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:

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

- 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:

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

Linear Logistic Regression

- Iteratively reweighted least squares (IRLS) algorithm

1. Initialize β

2. *Repeat*

3. Form linearized responses

$$z_i = x_i^T \beta + \frac{y_i - p_i}{p_i(1 - p_i)} \quad \leftarrow \quad \mathbf{z} = \mathbf{X}\beta^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})$$

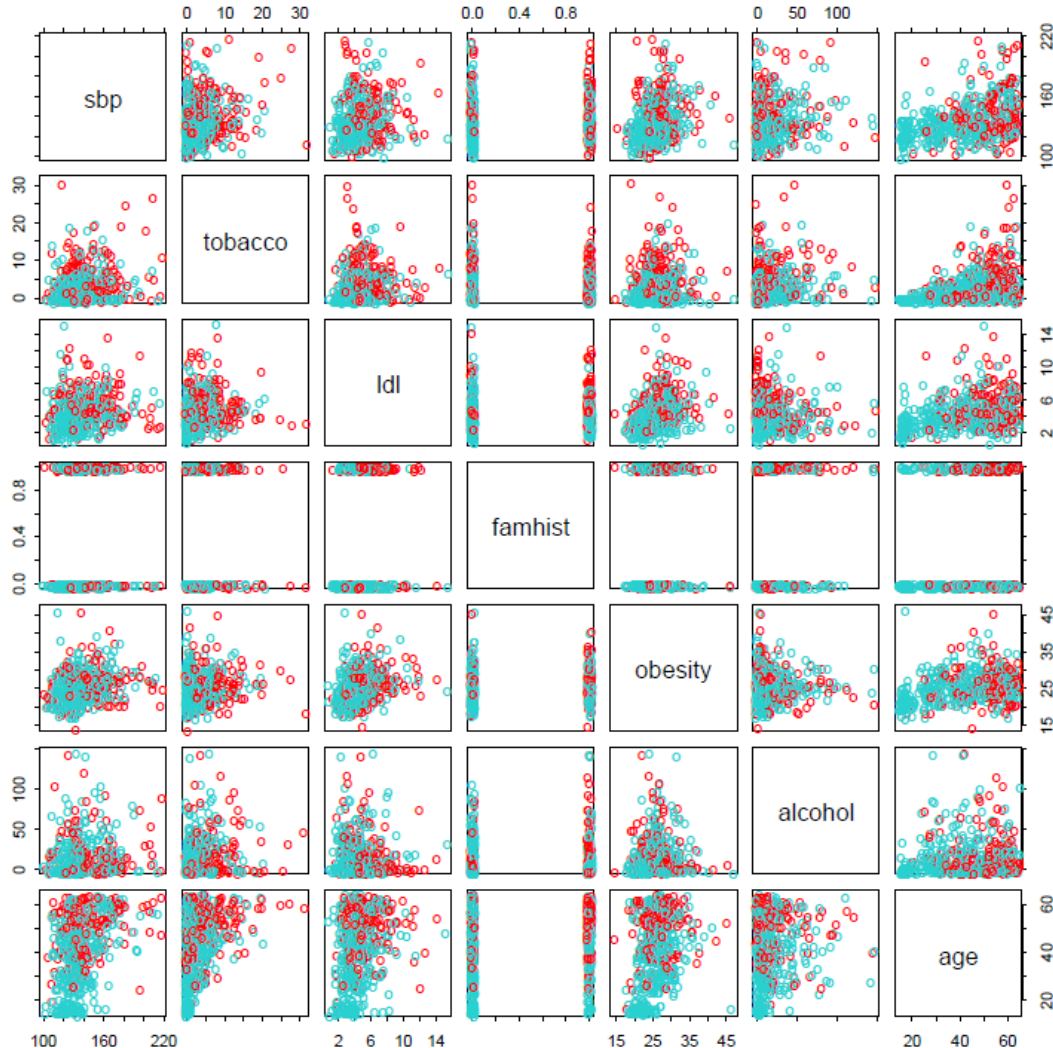
4. Form weights $w_i = p_i(1 - p_i)$

5. Update β by weighted least squares of z_i on x_i with w_i , $\forall i$

6. *Until convergence*

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

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
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
alcohol	0.001	0.004	0.136
age	0.043	0.010	4.184

收缩压

肥胖
饮酒

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

