Optimization and Machine Learning SI151

Lu Sun

School of Information Science and Technology

ShanghaiTech University

March 11, 2021

Today:

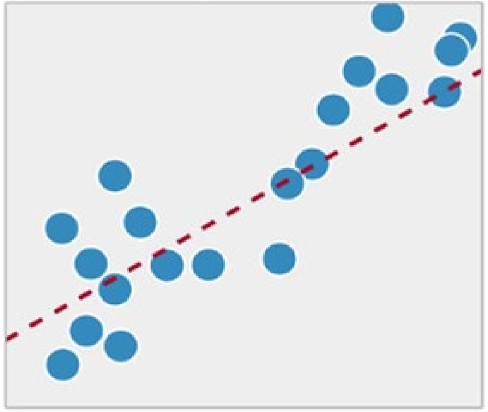
- Linear Methods for Classification I
  - Introduction
  - Linear regression of an indicator matrix
  - Linear discriminant analysis

Readings:

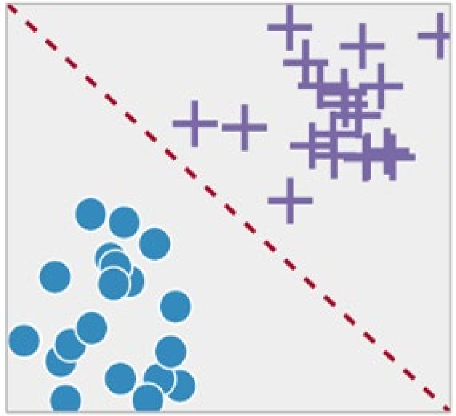
- The Elements of Statistical Learning (ESL), Chapters 4.1, 4.2 and 4.3

Linear Methods for Classification I

- Introduction
- Linear regression of an indicator matrix
- Linear discriminant analysis



Regression



Classification

Introduction

Example

Handwritten digits recognition

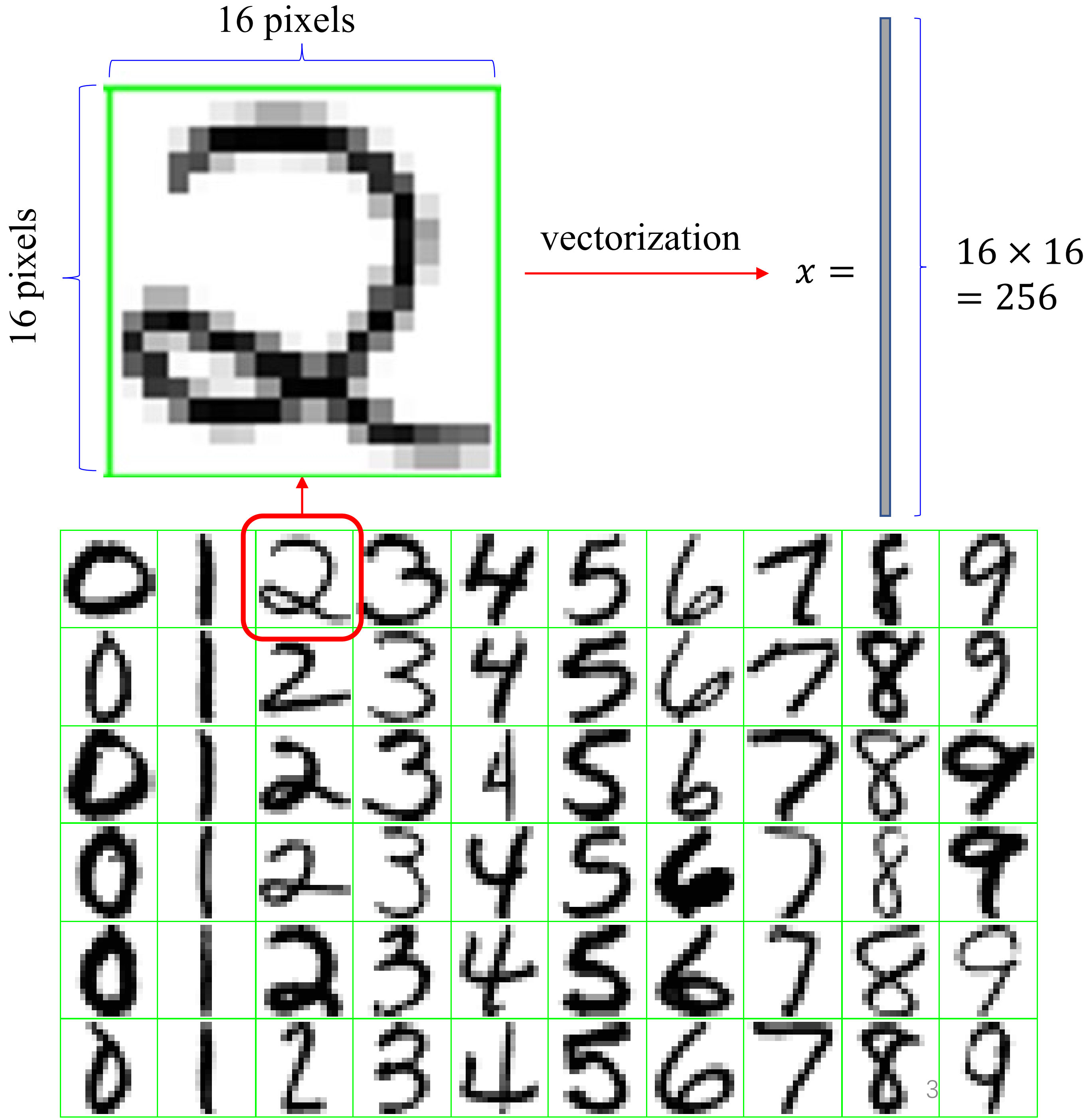
Input variables

$$X = (X_0, X_1, X_2, \dots, X_{256})^T$$

Categorical output variable G with values from

$$\mathcal{G} = \{0,1,2 \dots, 9\}$$

Non-binary (multi-class) classification



$\|a\|_2^2 = \langle a, a \rangle = a^T a$

$\|A\|_F^2 = \langle A, A \rangle = Tr(A^T A)$

$= Tr(A A^T)$

Introduction

Example

Handwritten digits recognition

qualitative: G → **classification**

(categorical or discrete)

coding $\xrightarrow{=2}$ a binary digit. {0,1} or {−1,1}

coding $\xrightarrow{>2}$ K binary digits. One is "on" at a time.

sample x_1^T

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

All-one vector for the intercept

Data matrix **X**

One-hot coding →

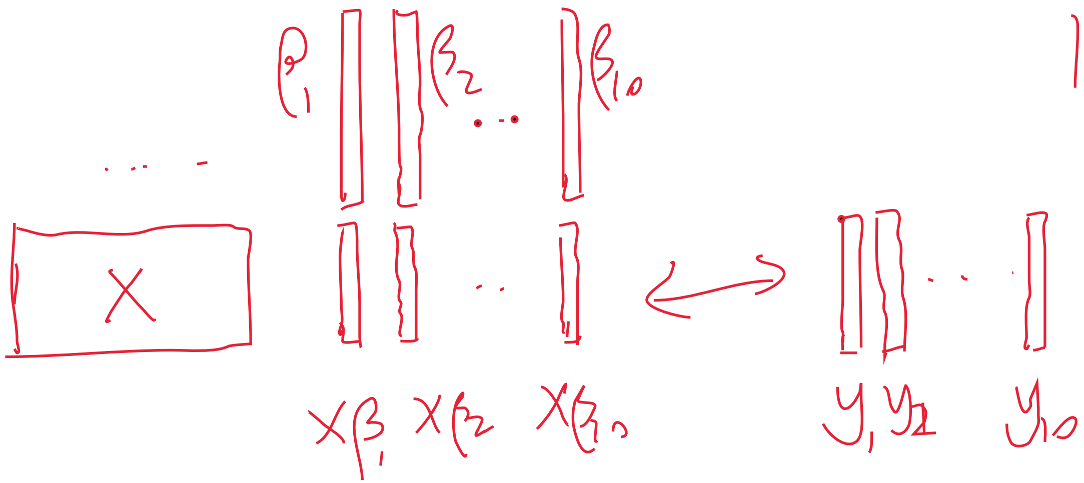
G
0
1
⋯
9

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

Coded output matrix **Y**

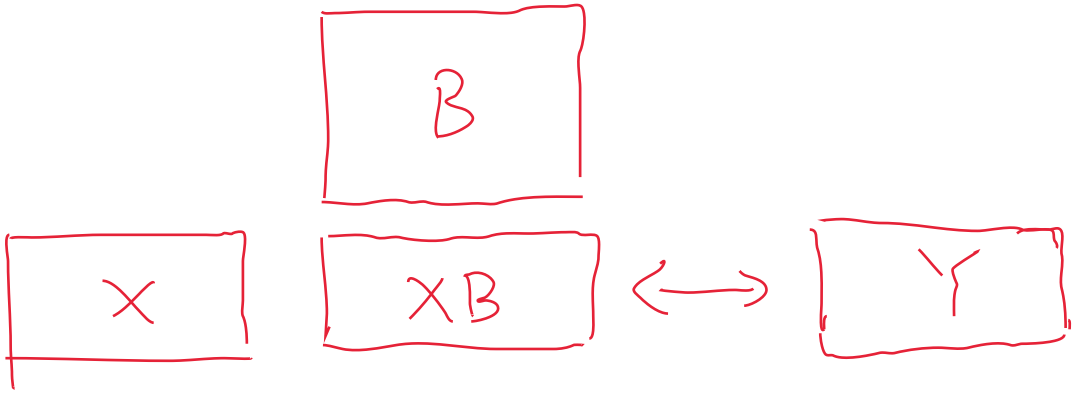
$$\min_{\mathbf{B}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 \longrightarrow \hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

1. Any problems?
2. Other methods?



$$\|y - X\beta\|_2^2 = \langle y - X\beta, y - X\beta \rangle$$

$$\|Y - XB\|_F^2 = \langle Y - XB, Y - XB \rangle$$



Introduction

Binary classification

- Linear regression

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

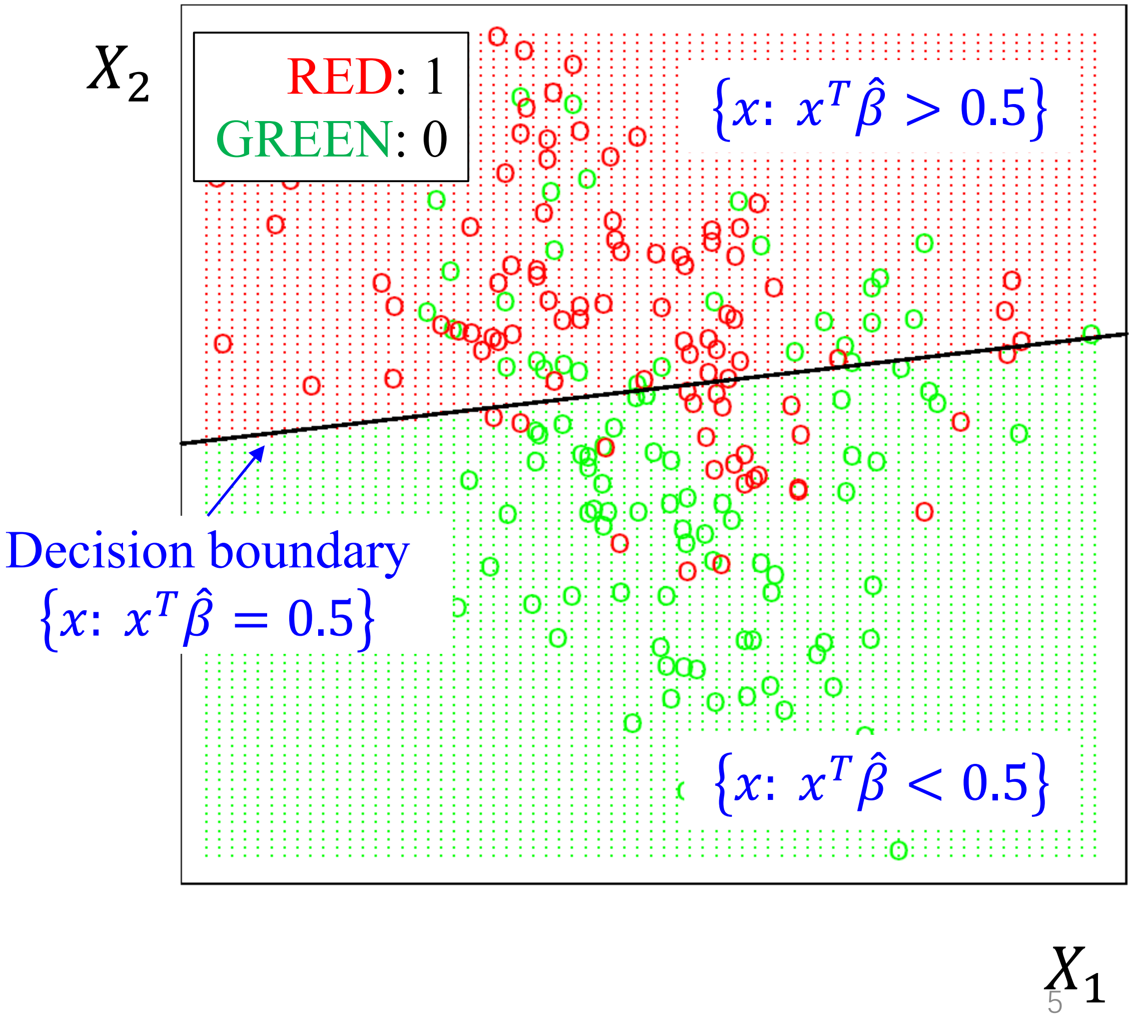
- Least squares solution

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

- Decision boundary

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

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



Introduction

Multi-class classification

- Linear regressions for K classes

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

- Decision boundary between classes k and ℓ:

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

For K classes, there are $\binom{K}{2} = \frac{K(K-1)}{2}$ decision boundaries

- That is an affine set or hyperplane:

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

Linear Methods for Classification I

- Introduction
- Linear regression of an indicator matrix
- Linear discriminant analysis

Linear Regression of an Indicator Matrix

- Indicator response matrix

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

N

G
0
1
...
9

coding

K

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

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

- Our problem:

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

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

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

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

$\mathcal{G} = \{1, 2, 3\}$

Linear Regression of an Indicator Matrix

$t_1 = (1, 0, 0) \rightarrow 1$

$t_2 = (0, 1, 0) \rightarrow 2$

$t_3 = (0, 0, 1) \rightarrow 3$

A new observation x is classified by

- Compute the fitted output

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

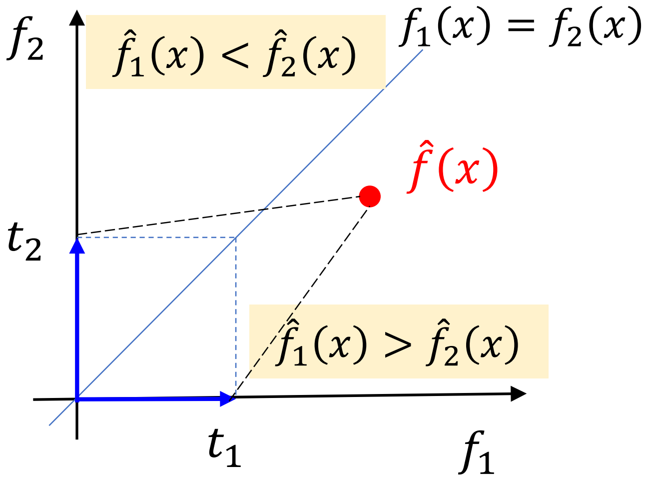
- Classify x according to

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

- Or equivalently,

$$\hat{G}(x) = \operatorname{argmin}_{k \in \mathcal{G}} \left\| \hat{f}(x) - t_k \right\|_2^2$$

where $t_k = (0, \dots, 0,1,0, \dots, 0)^T \in \mathbb{R}^K$ is a target with 1 being the k-th element



Linear Regression of an Indicator Matrix

Categorical output variable G with values from $\mathcal{G} = \{1, \ldots, K\}$.

- The zero-one loss function

$$L(k, \ell) = \begin{cases} 1, & k \neq \ell \\ 0, & k = \ell \end{cases}$$

- Expected prediction error (EPE) w.r.t. $\Pr(G, X)$

$$\mathrm{EPE} = \mathrm{E}\left[L\left(G, \hat{G}(X)\right)\right]$$

$\hat{G}(x) = \operatorname{arg\,max}_{k \in \mathcal{G}} \hat{f}_k(x)$

- Pointwise minimization leads to

$$\hat{G}(x) = \underset{k \in \mathcal{G}}{\operatorname{argmin}} \sum_{\ell=1}^{K} L(k, \ell) \Pr(G = \ell | X = x)$$

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

posterior

Linear Regression of an Indicator Matrix

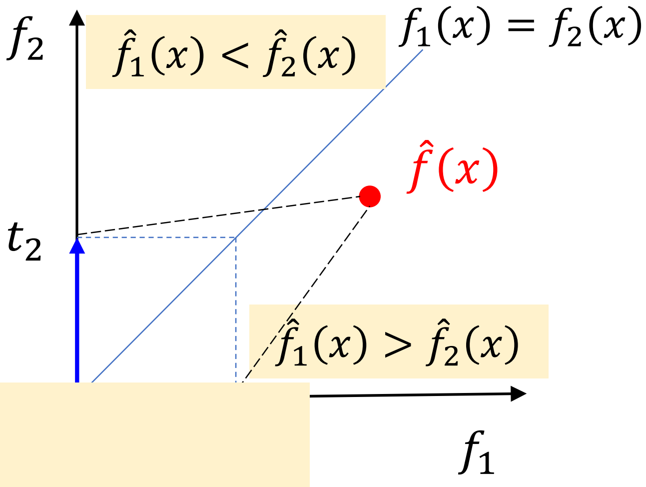
A new observation x is classified by

- Compute the fitted output

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

- Classify x according to

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



- Minimizing EPE w.r.t. the 0-1 loss gives rise to

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

- Our question:

  Are the $\hat{f}_k(x)$ reasonable estimates of the posterior $\Pr(G = k | X = x)$?

ment

Linear Regression of an Indicator Matrix

?

Linear classification:

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

$\hat{f}_k(x) = \hat{\beta}_{k0} + x^T \hat{\beta}_k$

Minimizing EPE:

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

Two defining properties of probability

1. $\sum P = 1$
2. $0 < P < 1$

- It can be verified that $\boxed{\sum_{k \in \mathcal{G}} \hat{f}_k(x) = 1}$
- However, it is possible that $\hat{f}_k(x) < 0$ or $\hat{f}_k(x) > 1$

Suppose that $\mathbf{X} \leftarrow (\mathbf{1}_N, \mathbf{X})$ and

$$\hat{\mathbf{Y}} = \hat{f}(\mathbf{X}) = \mathbf{X}\hat{\mathbf{B}} = \left(\hat{f}_1(\mathbf{X}), \dots, \hat{f}_K(\mathbf{X})\right)$$

We have the followings

Indicator matrix

$$\begin{aligned} \sum_{k=1}^{K} \hat{f}_K(\mathbf{X}) &= \hat{\mathbf{Y}} \cdot \mathbf{1}_K \\ &= \mathbf{X}\hat{\mathbf{B}} \cdot \mathbf{1}_K \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \boxed{\mathbf{Y} \cdot} \mathbf{1}_K \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot \mathbf{1}_N \\ &= \mathbf{H} \cdot \mathbf{1}_N \end{aligned}$$

$\mathbf{H} \cdot \mathbf{1}_N$ is a projection of $\mathbf{1}_N$ onto the column space of $\mathbf{X}$, thus $\mathbf{H} \cdot \mathbf{1}_N = \mathbf{1}_N$

Linear Regression of an Indicator Matrix

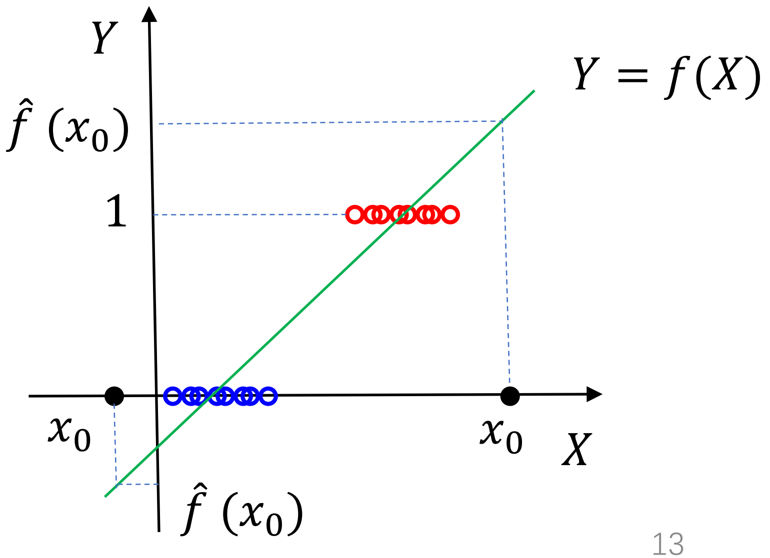
?

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

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

Two defining properties of probability
1. $\sum P = 1$
2. $0 < P < 1$

- It can be verified that $\sum_{k \in \mathcal{G}} \hat{f}_k(x) = 1$
- However, it is possible that $\boxed{\hat{f}_k(x) < 0 \text{ or } \hat{f}_k(x) > 1}$



It possibly suffers from the problem of masking
- a class may be masked by others, i.e., there is no region in the feature space that is labeled as this class

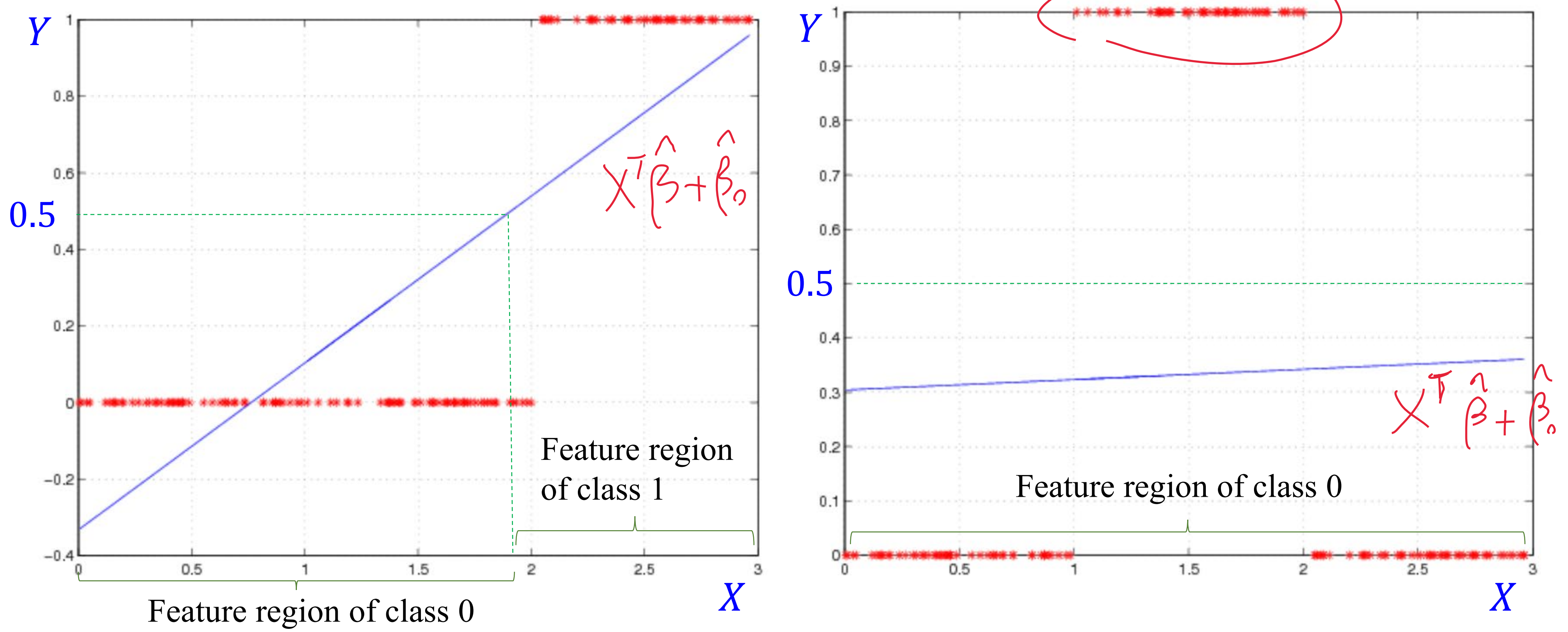
(K ≥ 3).

The Phenomenon of Masking

Decision boundary : $\{ x \mid x^T\hat{\beta}+\hat{\beta}_0 = 0.5 \}$

- A class may be masked by others, i.e., there is no region in the feature space that is labeled as this class
- The linear regression model is too rigid

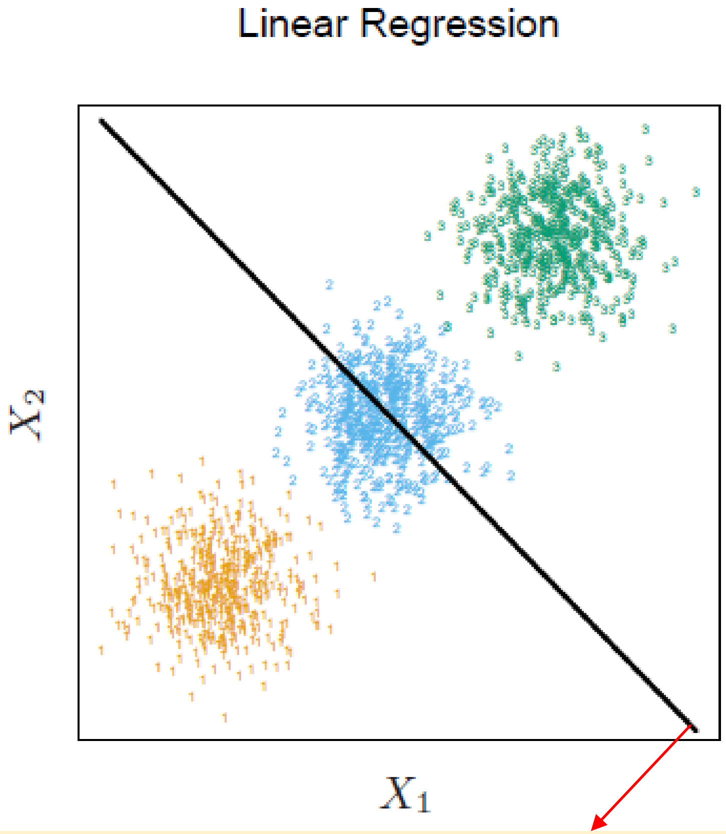
Class 1 is masked by class 0

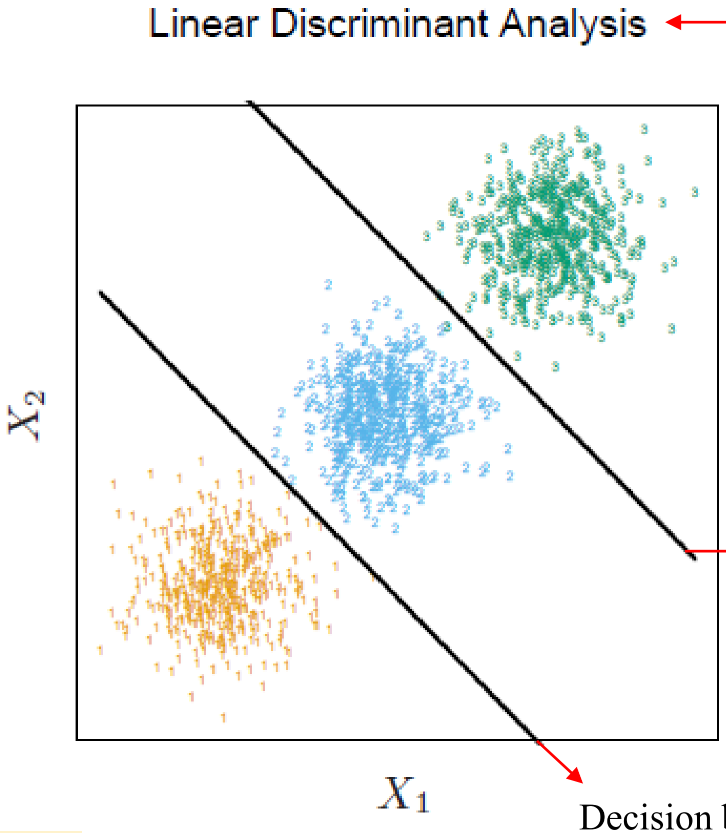


The Phenomenon of Masking

- 3-class classification

Yellow: class 1
Blue: class 2
Green: class 3





Ideal result

Decision boundary between classes 2 and 3

Decision boundary between classes 1 and 2

The decision boundaries between 1 and 2 and between 2 and 3 are the same, so we would never predict class 2.

The Phenomenon of Masking

- 3-class classification

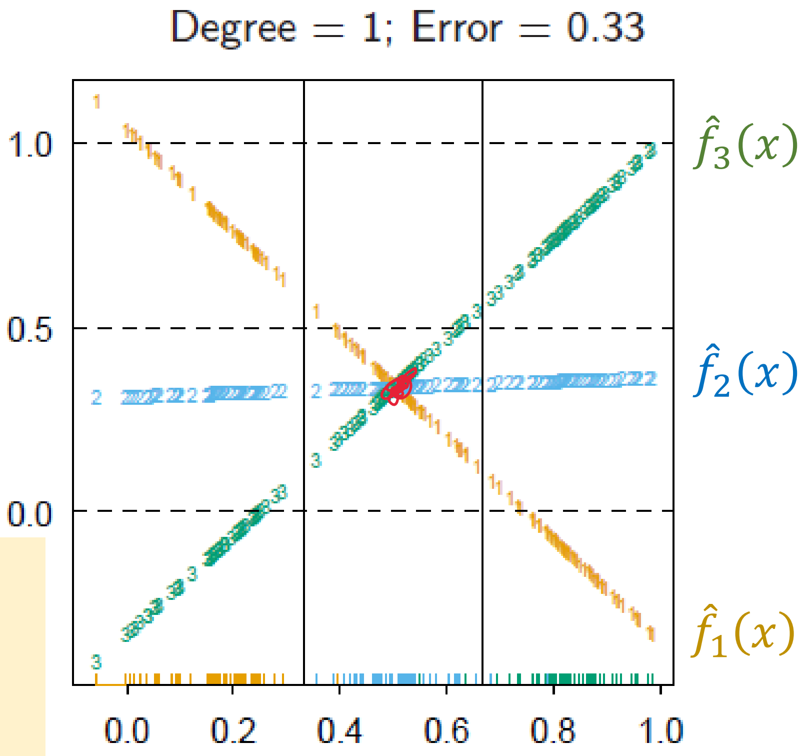
The indicator matrix

$$g = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \rightarrow \mathbf{Y} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

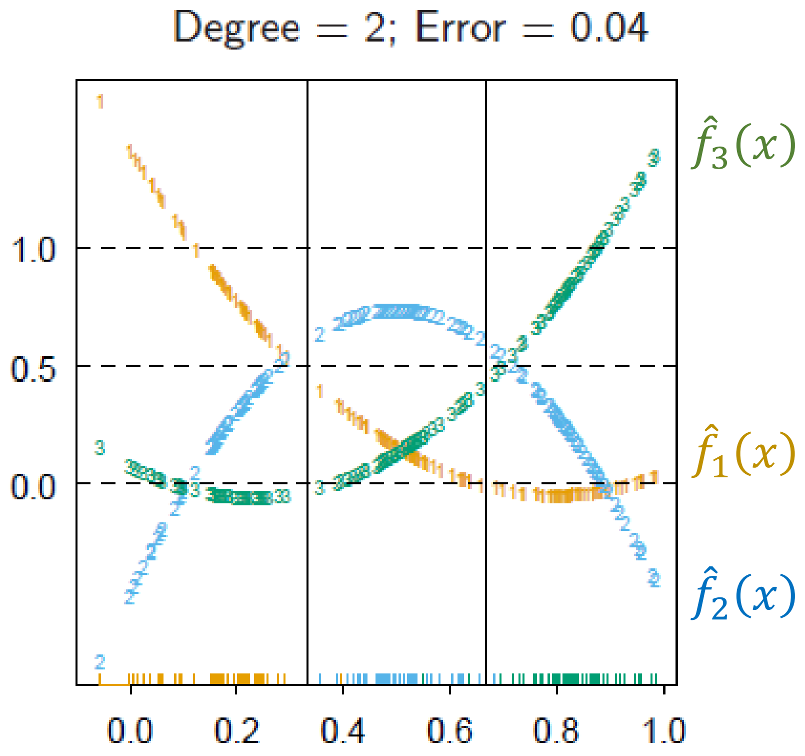
Yellow: class 1
Blue: class 2
Green: class 3

$$\widehat{\mathbf{B}} = \underset{\mathbf{B}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 ,$$
where $\mathbf{X} = (\mathbf{1}_N, \mathbf{x})$

$$\hat{f}(x) = \widehat{\mathbf{B}}^T \begin{pmatrix} 1 \\ x \end{pmatrix} = \begin{pmatrix} \hat{f}_1(x) \\ \hat{f}_2(x) \\ \hat{f}_3(x) \end{pmatrix}$$



Linear regression
$Y = \beta_0 + \beta X$



Quadratic regression
$Y = \beta_0 + \beta_1 X + \beta_2 X^2$

Linear Methods for Classification I

- Introduction
- Linear regression of an indicator matrix
- Linear discriminant analysis

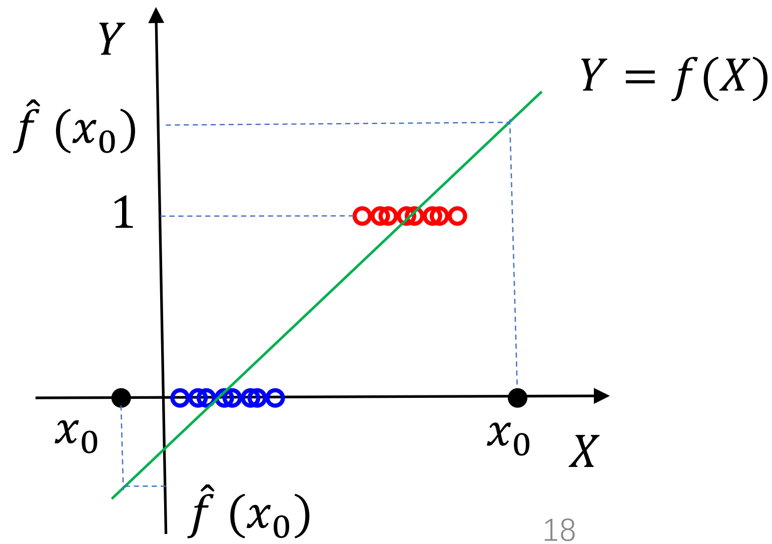
Linear Discriminant Analysis

- Recall our discussion on linear regression of an indicator matrix

Linear classification: $\hat{G}(x) = \underset{k \in \mathcal{G}}{\operatorname{argmax}} \boxed{\hat{f}_k(x)}$ $(-\infty, +\infty)$

Minimizing EPE: $\hat{G}(x) = \underset{k \in \mathcal{G}}{\operatorname{argmax}} \boxed{\Pr(G = k | X = x)}$ $[0, 1]$

- 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

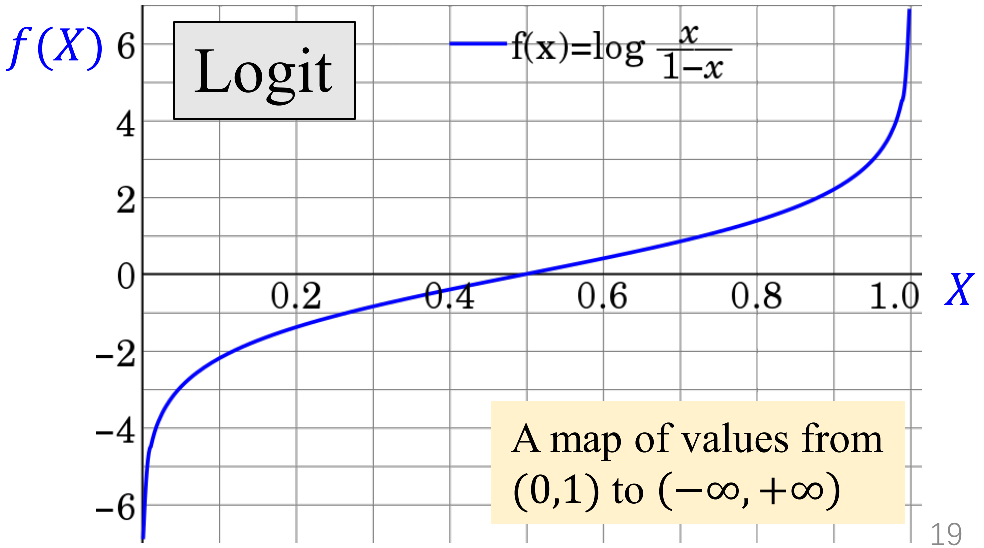
Odds (发生比)

- Logit transform

$$\text{logit}(\Pr(x)) = \log\left(\boxed{\frac{\Pr(x)}{1-\Pr(x)}}\right) = 0$$

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

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



A map of values from (0,1) to (−∞, +∞)

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

Q

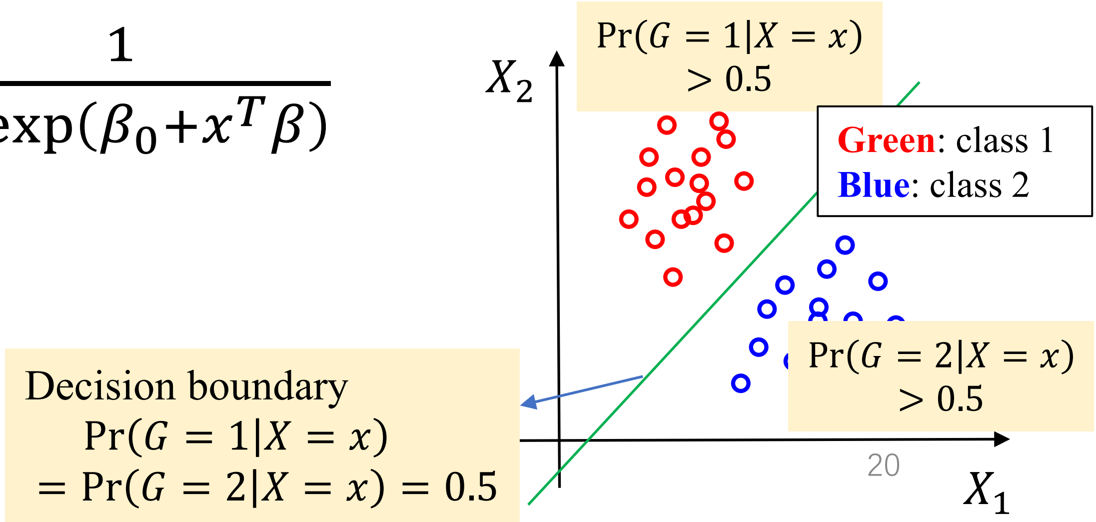
$$\Pr(G=1|X=x) = \frac{\boxed{\exp(\beta_0+x^T\beta)}}{1+\exp(\beta_0+x^T\beta)},$$

$\exp(x) = e^x$

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

- Decision boundary

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



Linear Discriminant Analysis

The Bayes theorem

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

- Idea:

  model the posterior $\Pr(G = k|X = x)$ based on the Bayes theorem

- Posterior

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

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

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

  - Class prior:

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

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

- It produces LDA, QDA (quadratic DA), MDA (mixture DA), kernel DA and naïve Bayes, under various assumptions on $f_k(x)$

Linear Discriminant Analysis

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

- Assumptions in LDA

  #paras: $= K \times p + K \times p \times p$

  1. Model each class density as multivariate Gaussian

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

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

- Compare two classes k and ℓ

Logit:

$$\log\frac{\Pr(G=k|X=x)}{\Pr(G=\ell|X=x)} = \log\frac{f_k(x)}{f_\ell(x)} + \log\frac{\pi_k}{\pi_\ell}$$
$$= \boxed{\log\frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k+\mu_\ell)^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_\ell) + x^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_\ell), \quad = 0}$$

Quadratic term vanished due to the common covariance

Decision boundary is linear w.r.t. X

Linear Discriminant Analysis

- Parameter estimation

$$\hat{\pi}_k = N_k/N, \text{ where } N_k \text{ is the number of class-}k \text{ observations;}$$

$$\hat{\mu}_k = \sum_{g_i=k} x_i/N_k;$$

$$\boxed{\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T/(N-K).}$$

Pooled covariance (合并方差)

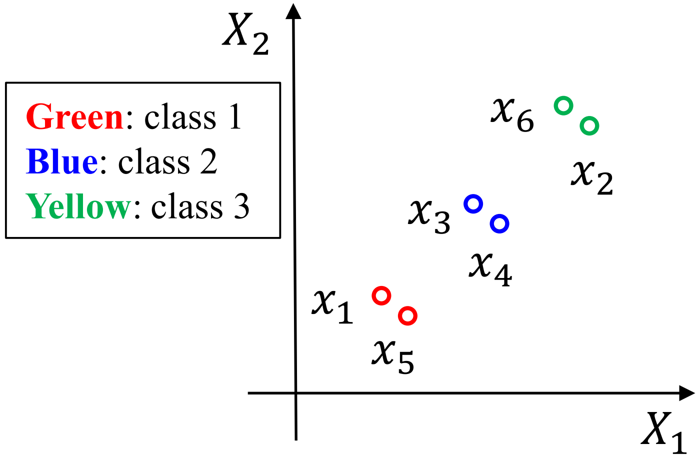
$$\hat{\Sigma} = \frac{(N_1-1)\hat{\Sigma}_1 + (N_2-1)\hat{\Sigma}_2 + \cdots + (N_K-1)\hat{\Sigma}_K}{\underline{(N_1-1) + (N_2-1) + \cdots + (N_K-1)}}, \text{where } \hat{\Sigma}_k = \frac{\sum_{g_i=k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T}{N_k-1}$$

Weighted average

Linear Discriminant Analysis

Data / Class

	X_1	X_2	G
x_1^T	0.2	0.3	1
x_2^T	0.8	0.7	3
x_3^T	0.4	0.6	2
x_4^T	0.6	0.4	2
x_5^T	0.3	0.2	1
x_6^T	0.7	0.8	3



- Class prior

$$\hat{\pi}_1 = \hat{\pi}_2 = \hat{\pi}_3 = \frac{1}{3}$$

- Class-specific sample mean

$$\hat{\mu}_1 = \frac{1}{2}(x_1 + x_5) = \frac{1}{2}\begin{pmatrix}0.2\\0.3\end{pmatrix} + \frac{1}{2}\begin{pmatrix}0.3\\0.2\end{pmatrix} = \begin{pmatrix}0.25\\0.25\end{pmatrix}$$

$$\hat{\mu}_2 = \frac{1}{2}(x_3 + x_4) = \frac{1}{2}\begin{pmatrix}0.4\\0.6\end{pmatrix} + \frac{1}{2}\begin{pmatrix}0.6\\0.4\end{pmatrix} = \begin{pmatrix}0.5\\0.5\end{pmatrix}$$

$$\hat{\mu}_3 = \frac{1}{2}(x_2 + x_6) = \frac{1}{2}\begin{pmatrix}0.8\\0.7\end{pmatrix} + \frac{1}{2}\begin{pmatrix}0.7\\0.8\end{pmatrix} = \begin{pmatrix}0.75\\0.75\end{pmatrix}$$

- Common covariance

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

$$\frac{\begin{pmatrix}0.005 & -0.005\\-0.005 & 0.005\end{pmatrix}+\begin{pmatrix}0.02 & -0.02\\-0.02 & 0.02\end{pmatrix}+\begin{pmatrix}0.005 & -0.005\\-0.005 & 0.005\end{pmatrix}}{6-3} = \begin{pmatrix}0.03 & -0.03\\-0.03 & 0.03\end{pmatrix}$$

Linear Discriminant Analysis

	Data X_1	Data X_2	Class G
x_1^T	0.2	0.3	1
x_2^T	0.8	0.7	3
x_3^T	0.4	0.6	2
x_4^T	0.6	0.4	2
x_5^T	0.3	0.2	1
x_6^T	0.7	0.8	3

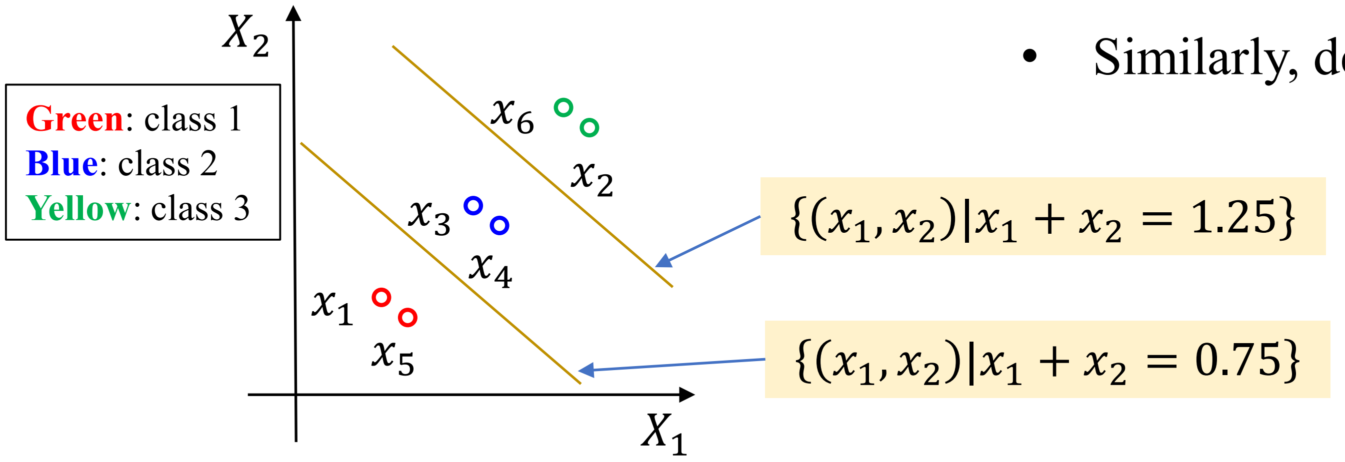
- For classes 1 and 2

$$\hat{\Sigma}_\lambda = \hat{\Sigma} + \lambda \mathbf{I} \leftarrow \lambda = 1$$

$$\log \frac{\Pr(G=1|X=x)}{\Pr(G=2|X=x)}$$
$$= \log \frac{\hat{\pi}_1}{\hat{\pi}_2} - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)^T \boxed{\hat{\Sigma}_\lambda^{-1}} (\hat{\mu}_1 - \hat{\mu}_2) + x^T \boxed{\hat{\Sigma}_\lambda^{-1}} (\hat{\mu}_1 - \hat{\mu}_2)$$
$$= \frac{1}{2}(0.75, 0.75)\begin{pmatrix} 0.972 & 0.028 \\ 0.028 & 0.972 \end{pmatrix}\begin{pmatrix} 0.25 \\ 0.25 \end{pmatrix} - (x_1, x_2)\begin{pmatrix} 0.972 & 0.028 \\ 0.028 & 0.972 \end{pmatrix}\begin{pmatrix} 0.25 \\ 0.25 \end{pmatrix}$$
$$= 0.1875 - (x_1, x_2)\begin{pmatrix} 0.25 \\ 0.25 \end{pmatrix} = 0$$

- Decision boundary 1-2: $\{(x_1, x_2) | x_1 + x_2 = 0.75\}$

- Similarly, decision boundary 2-3: $\{(x_1, x_2) | x_1 + x_2 = 1.25\}$



Linear Discriminant Analysis

- Suppose that $\log \frac{\Pr(G=k|X=x)}{\Pr(G=\ell|X=x)} = \delta_k(x) - \delta_\ell(x)$
  - $\delta_k(x) > \delta_\ell(x)$, class k
  - $\delta_k(x) < \delta_\ell(x)$, class ℓ
  - $\delta_k(x) = \delta_\ell(x)$, decision boundary

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

- Linear discriminant functions

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log \pi_k$$

Classify to class k that maximizes the discriminant function

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

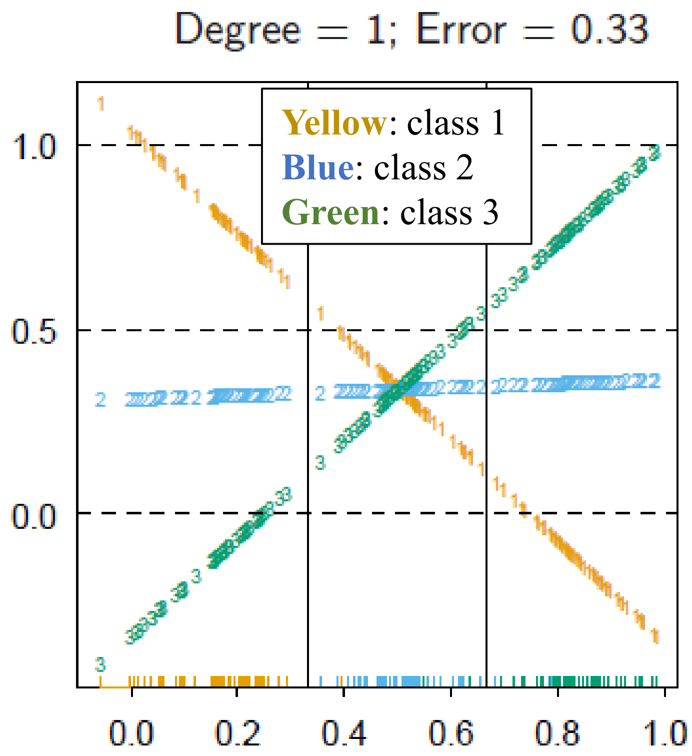
Any difference?

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

Linear Discriminant Analysis

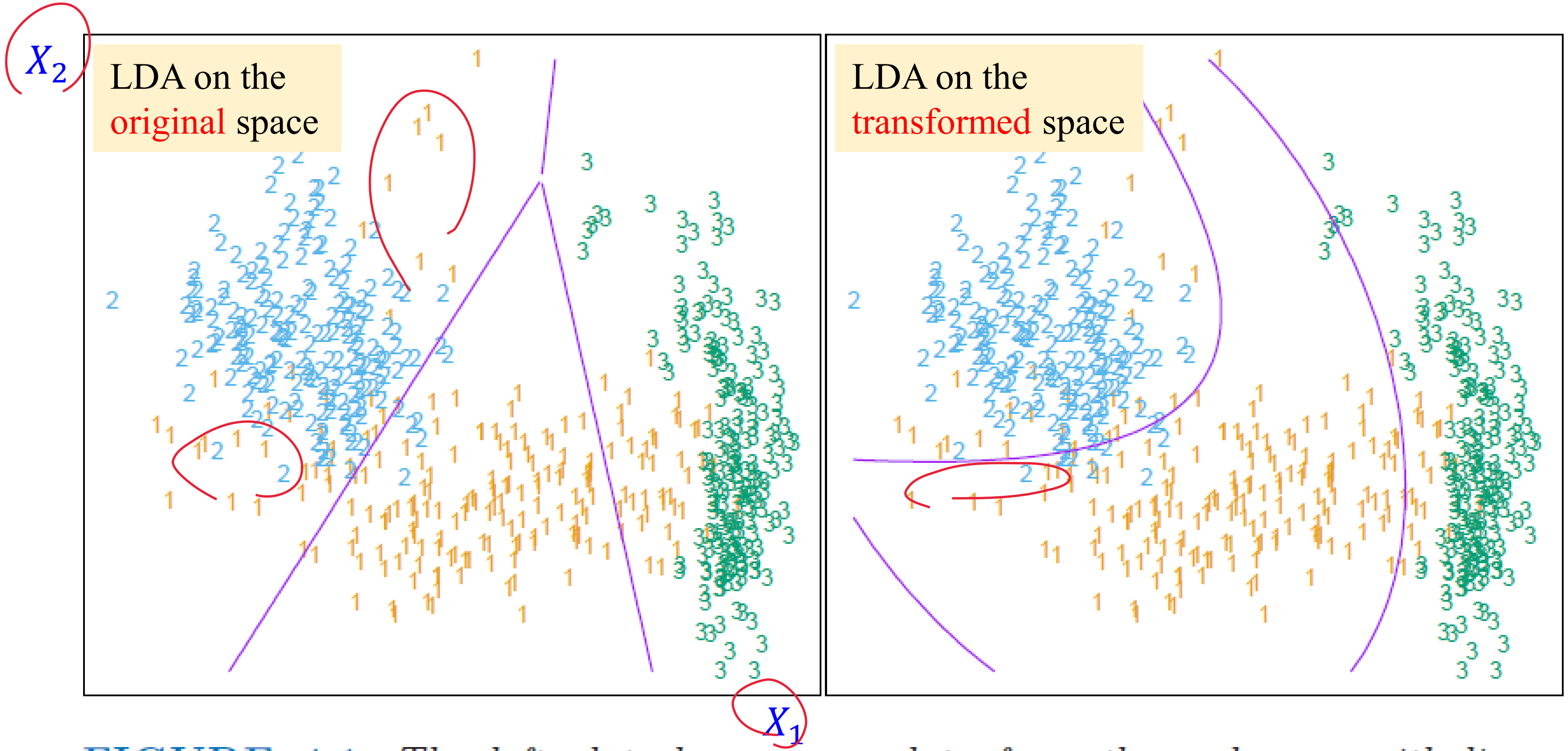
- Binary classification ($K = 2$)
  - Correspondence between LDA and linear classification

- Multi-class classification ($K \geq 3$)
  - LDA is different with linear classification
  - Avoid the masking problem



Class 2 is masked by classes 1 and 3

Linear Discriminant Analysis



$K\times p + p\times p$

FIGURE 4.1. *The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space* $\boxed{X_1, X_2, X_1X_2, X_1^2, X_2^2.}$ *Linear inequalities in this space are quadratic inequalities in the original space.*

Linear Discriminant Analysis

$P(G=y \mid X=x)$

$= p(G=b \mid X=x)$

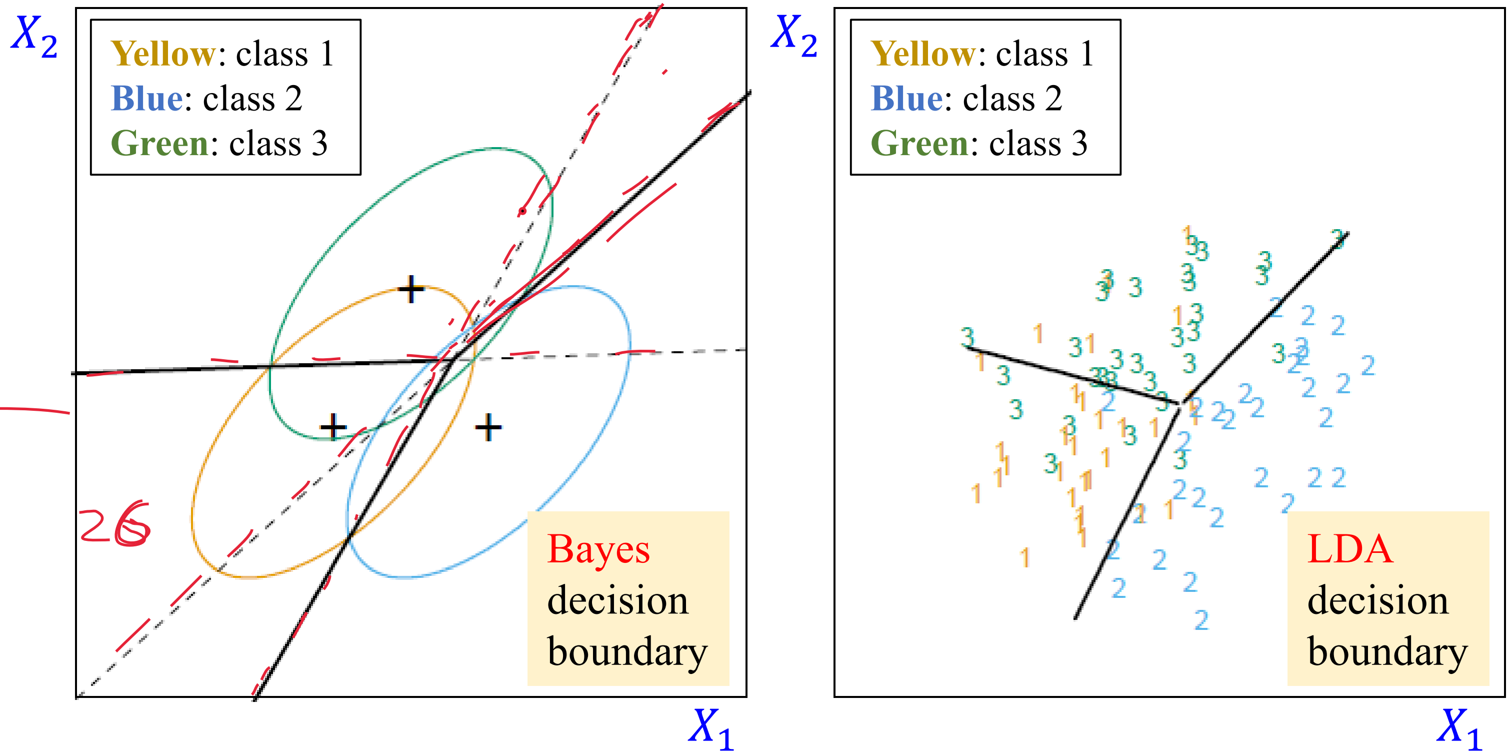


FIGURE 4.5. *The left panel shows three Gaussian distributions, with the same covariance and different means. Included are the contours of constant density enclosing 95% of the probability in each case. The Bayes decision boundaries between each pair of classes are shown (broken straight lines), and the Bayes decision boundaries separating all three classes are the thicker solid lines (a subset of the former). On the right we see a sample of 30 drawn from each Gaussian distribution, and the fitted LDA decision boundaries.*

Quadratic Discriminant Analysis

- Assumptions in LDA
  1. Model each class density as multivariate Gaussian
  $$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right)$$
  2. Assume that classes share a $\boxed{\text{common covariance } \Sigma_k = \Sigma, \forall k}$

- Assumption: Each class has a specific covariance Σ_k
- Quadratic discriminant functions

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) + \log \pi_k.$$

- The quadratic decision boundary between two classes k and ℓ

$$\{x: \delta_k(x) = \delta_\ell(x)\}$$

- Difference with LDA
  - Σ_k has to be estimated for each class
  - LDA need to estimate $\boxed{K \times p} + \boxed{p \times p}$ parameters
  - QDA need to estimate $\boxed{K \times p} + \boxed{K \times p \times p}$ parameters

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

Σ

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

Quadratic Discriminant Analysis

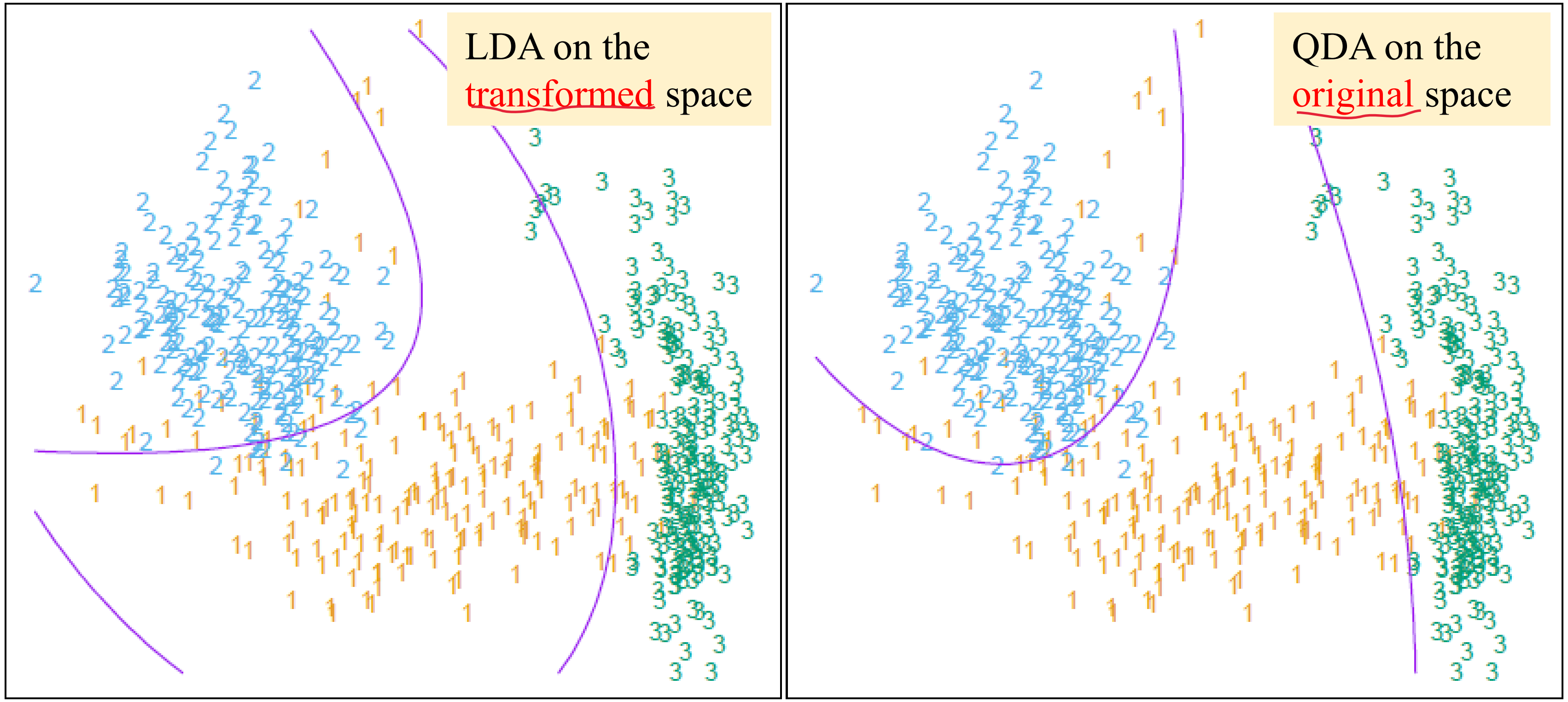


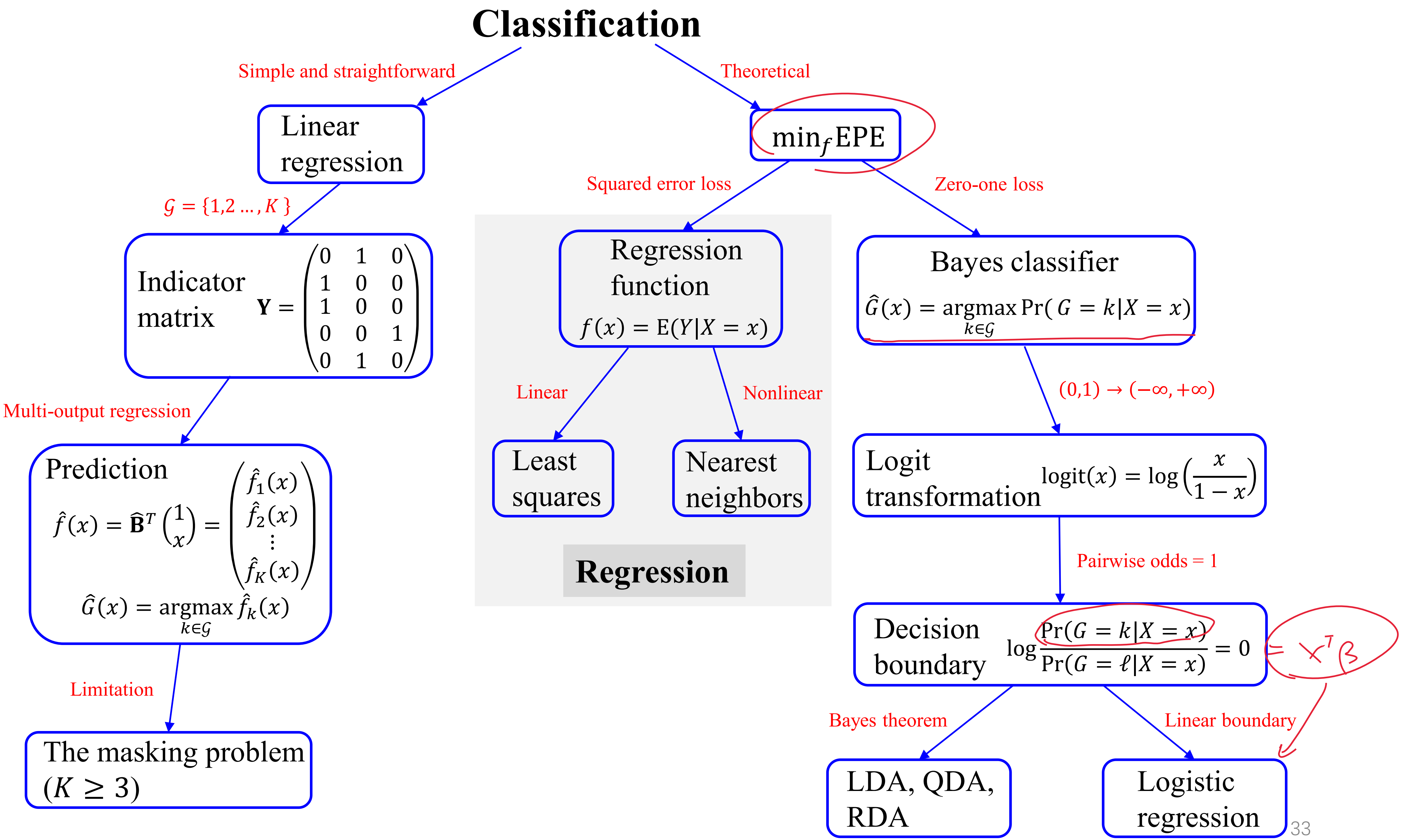
FIGURE 4.6. *Two methods for fitting quadratic boundaries. The left plot shows the quadratic decision boundaries for the data in Figure 4.1 (obtained using LDA in the five-dimensional space* $\boxed{X_1, X_2, X_1X_2, X_1^2, X_2^2}$*). The right plot shows the quadratic decision boundaries found by QDA. The differences are small, as is usually the case.*

Summary

- Linear regression of an indicator matrix
  - The indicator matrix one-hot
  - Prediction is conducted by $\hat{G}(x) = \operatorname{argmax}_k \hat{f}_k(x)$
  - Suffer from the masking problem

- Linear discriminant analysis
  - Logit transformation: $\operatorname{logit}(\Pr(x)) = \log\left(\frac{\Pr(x)}{1-\Pr(x)}\right)$
  - Model the posterior $\Pr(G = k|X = x)$
  - Assumptions on $\Pr(X = x|G = k)$
  - Discriminant functions $\delta_k(x)$

- Quadratic discriminant analysis
  - Difference with LDA


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
$f(x) = \mathrm{E}(Y|X = x)$
Bayes classifier
$\hat{G}(x) = \operatorname{argmax}_{k \in \mathcal{G}} \Pr(G = k|X = x)$
Linear
Nonlinear
$(0,1) \rightarrow (-\infty, +\infty)$
Multi-output regression
Least squares
Nearest neighbors
Logit transformation
$\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$
Prediction
$\hat{f}(x) = \hat{\mathbf{B}}^T \binom{1}{x} = \begin{pmatrix} \hat{f}_1(x) \\ \hat{f}_2(x) \\ \vdots \\ \hat{f}_K(x) \end{pmatrix}$
$\hat{G}(x) = \operatorname{argmax}_{k \in \mathcal{G}} \hat{f}_k(x)$
Regression
Pairwise odds = 1
Decision boundary
$\log \frac{\Pr(G = k|X = x)}{\Pr(G = \ell|X = x)} = 0$
$= x^T\beta$
Limitation
Bayes theorem
Linear boundary
The masking problem ($K \geq 3$)
LDA, QDA, RDA
Logistic regression