Introduction

Binary classification

- Linear regression

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

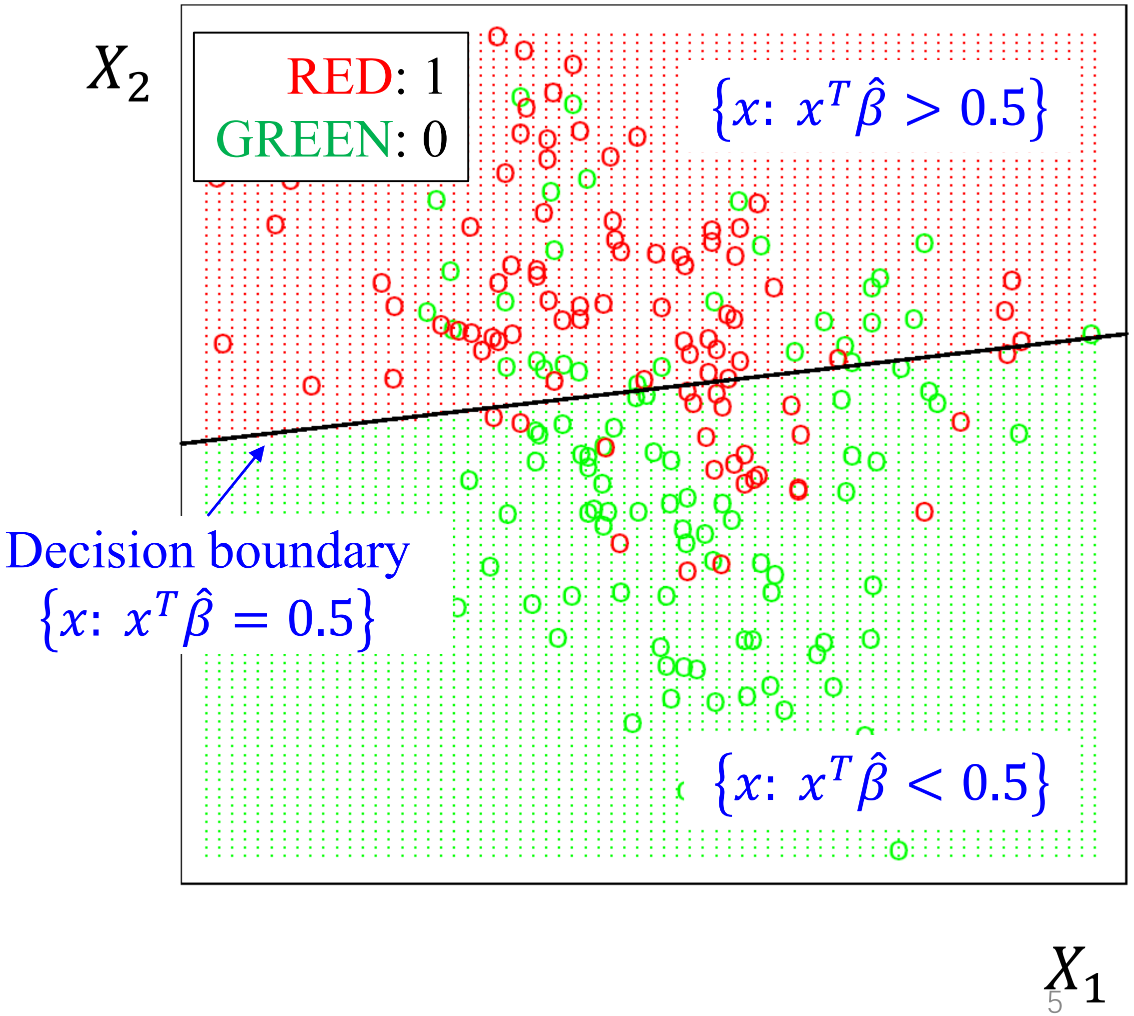
- Least squares solution

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

- 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

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

- 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 Regression of an Indicator Matrix

- Indicator response matrix

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

G
0
1
…
9

(N rows) coding →

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

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

- Our problem:

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

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

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

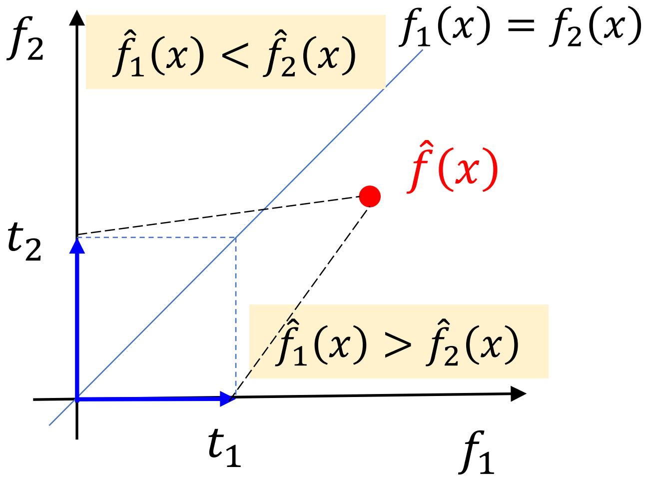
$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

Linear Regression of an Indicator Matrix

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

$X^T X$

?

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

K

N

$K \times 1$

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

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

I

$\mathbf{H} \cdot \mathbf{1}_N$ 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:

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

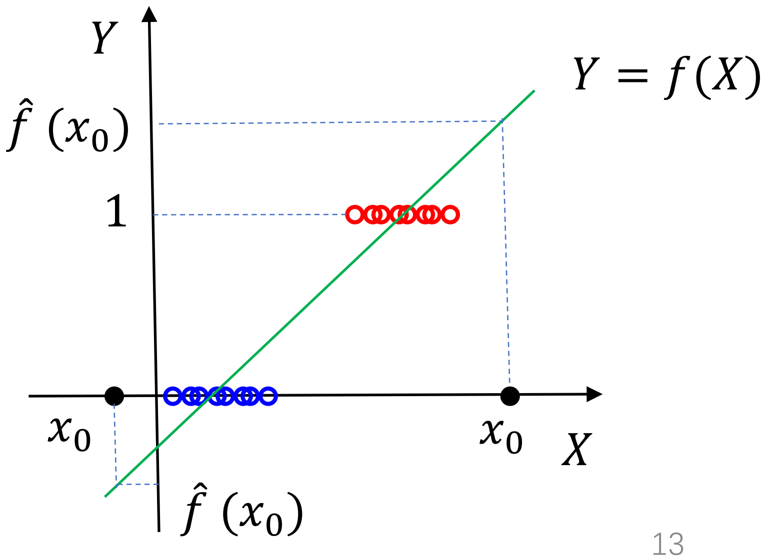
Minimizing EPE:

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

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 $\boxed{\hat{f}_k(x) < 0 \text{ or } \hat{f}_k(x) > 1}$

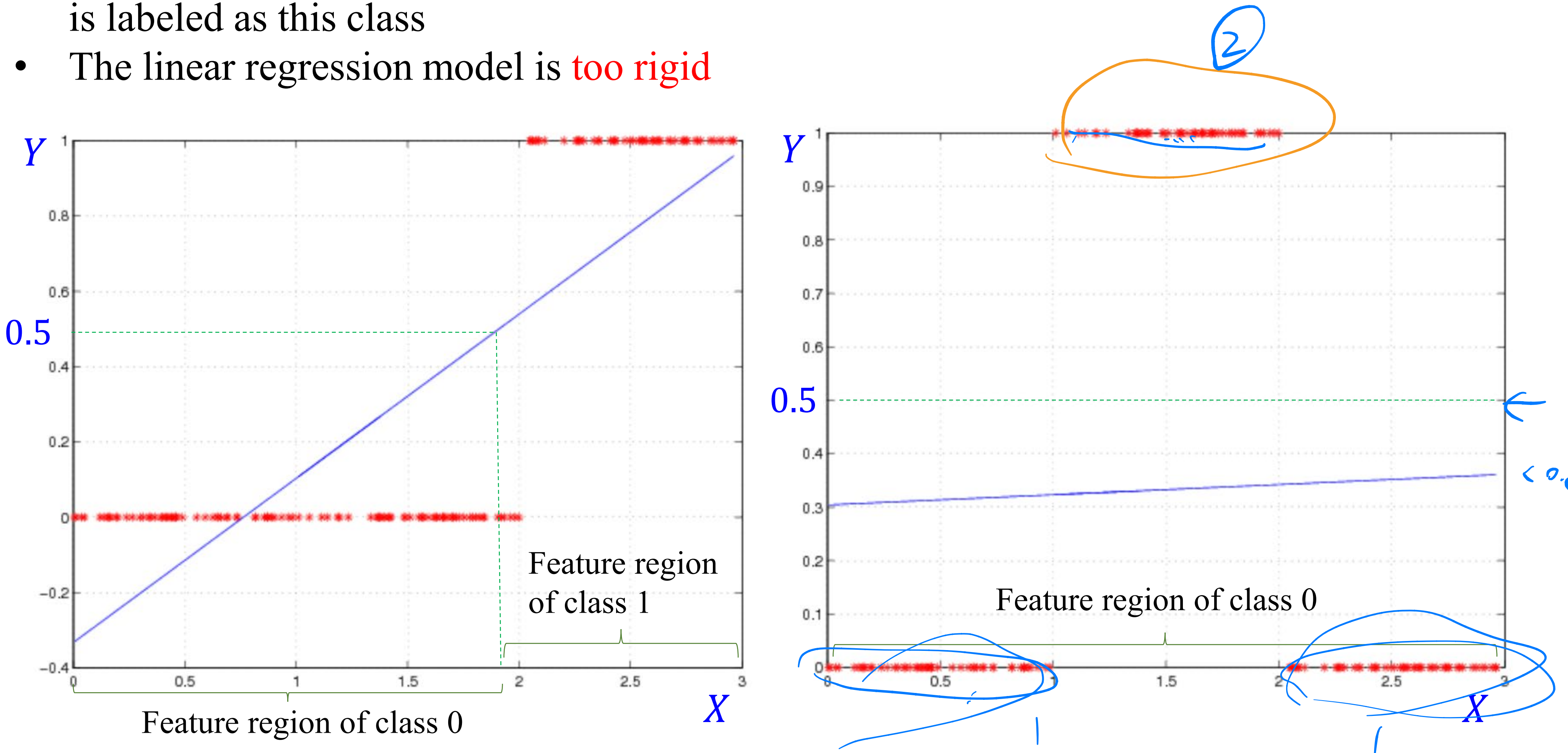


It possibly suffers from the problem of masking

- a class may be masked by others, i.e., there is no region in the feature space that is labeled as this class

The Phenomenon of Masking

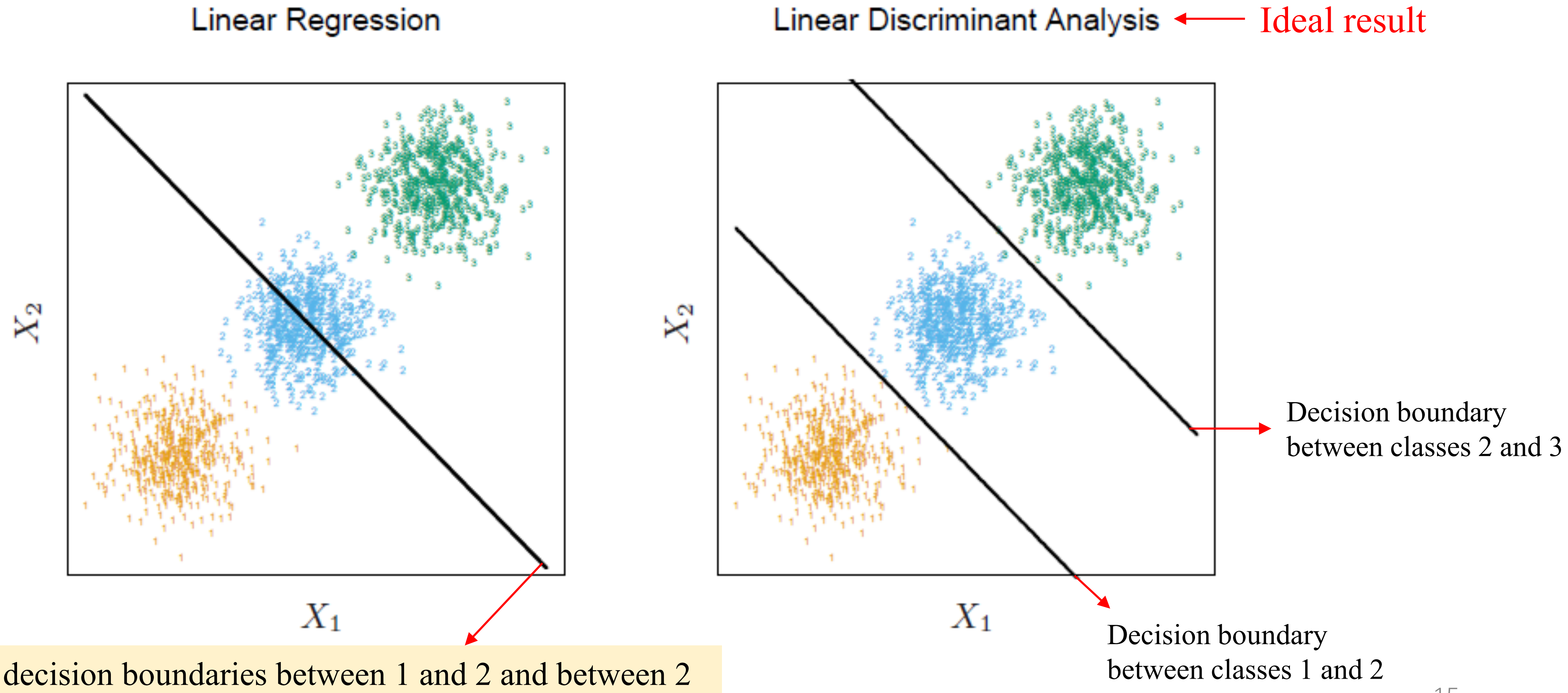
- A class may be masked by others, i.e., there is no region in the feature space that is labeled as this class
- The linear regression model is too rigid



The Phenomenon of Masking

- 3-class classification

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



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

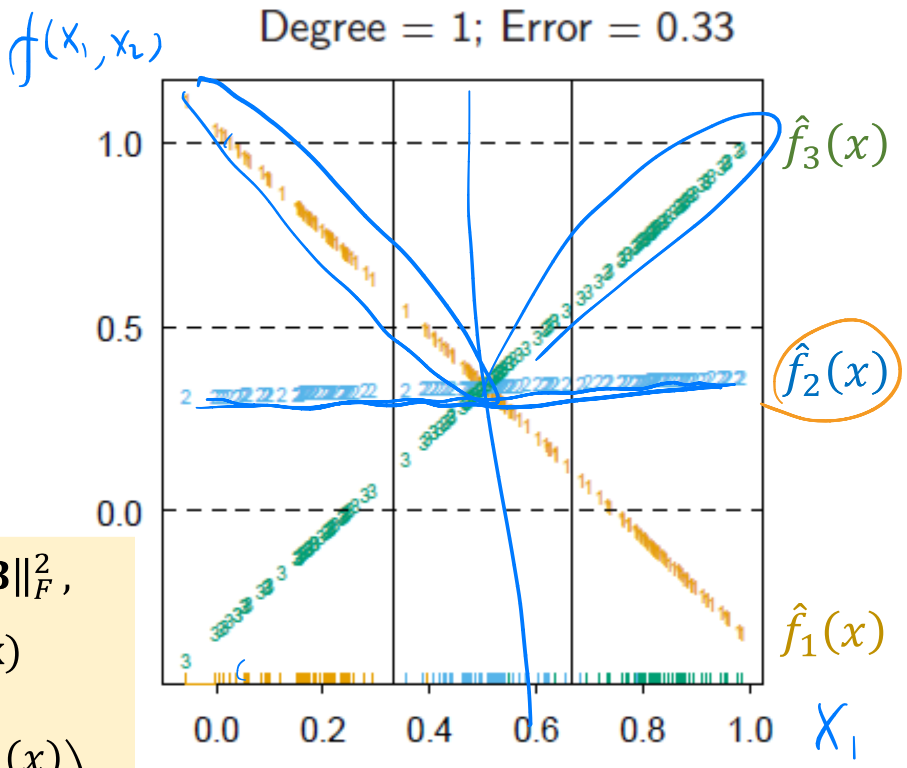
- 3-class classification

The indicator matrix

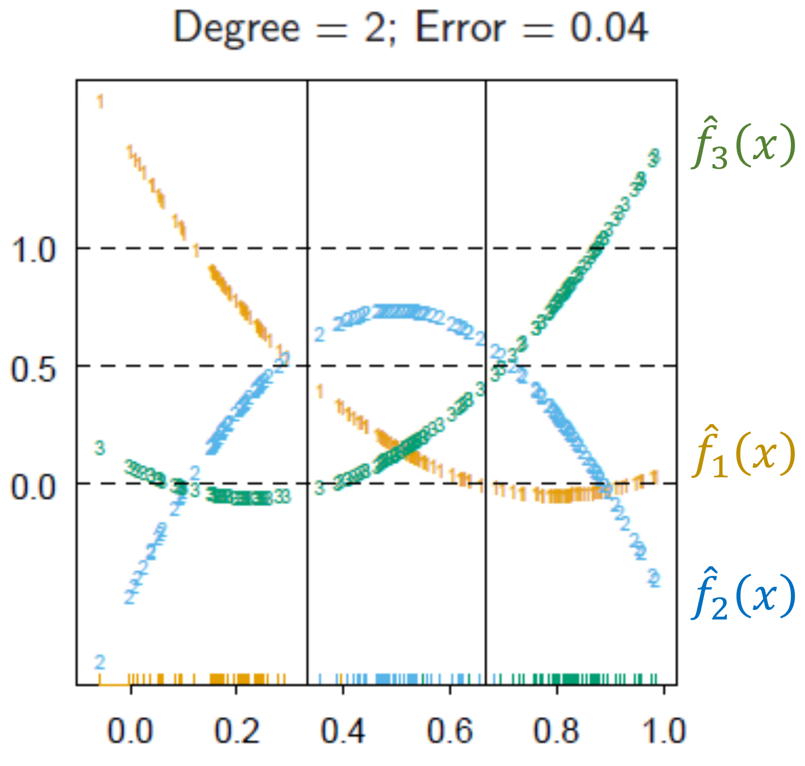
$$g = \begin{pmatrix}1\\2\\3\end{pmatrix} \rightarrow \mathbf{Y} = \begin{pmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{pmatrix}$$

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

$1 \leftarrow \quad \rightarrow 3$



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



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

$$\widehat{\mathbf{B}} = \underset{\mathbf{B}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2,$$
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 Discriminant Analysis

- Recall our discussion on linear regression of an indicator matrix

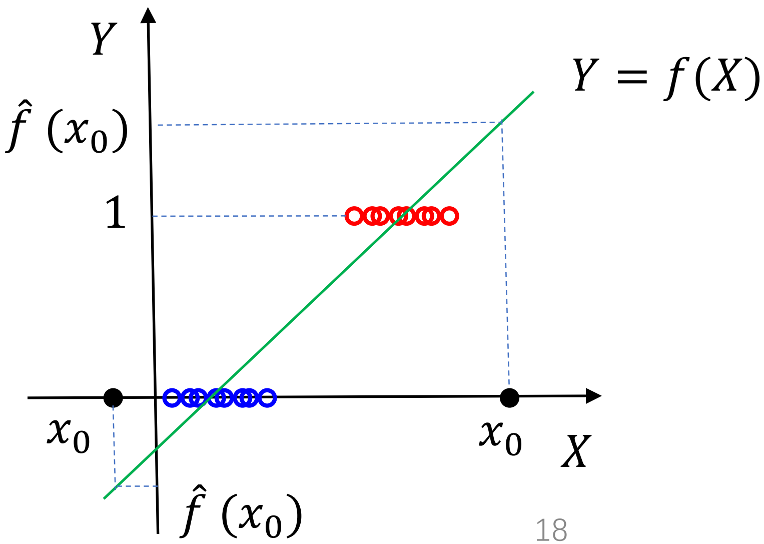
Linear classification:

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

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

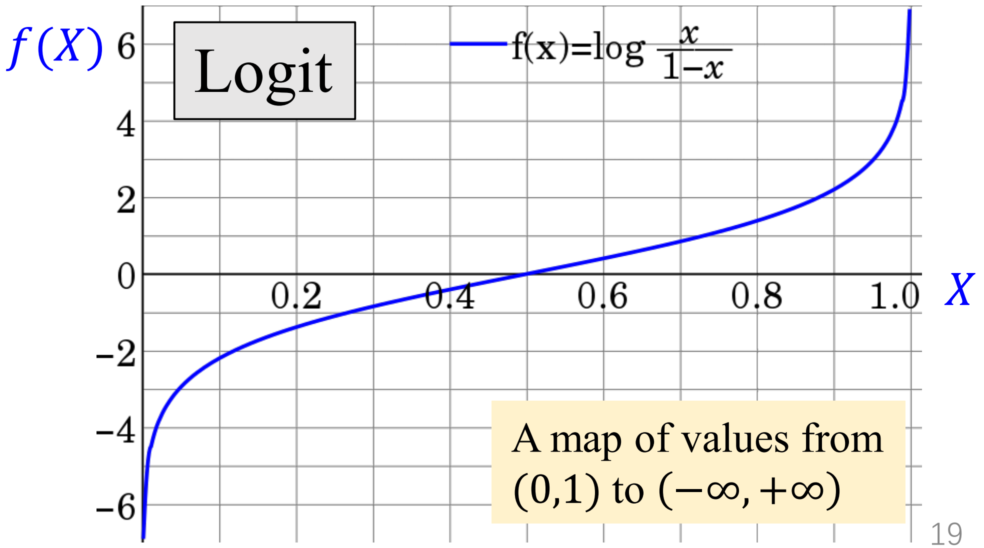
Odds (发生比)

- Logit transform

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

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

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



Linear Discriminant Analysis

- Example: binary (two class) classification =0

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

1 − Pr(G=1|X=x) exp

- The posterior probability

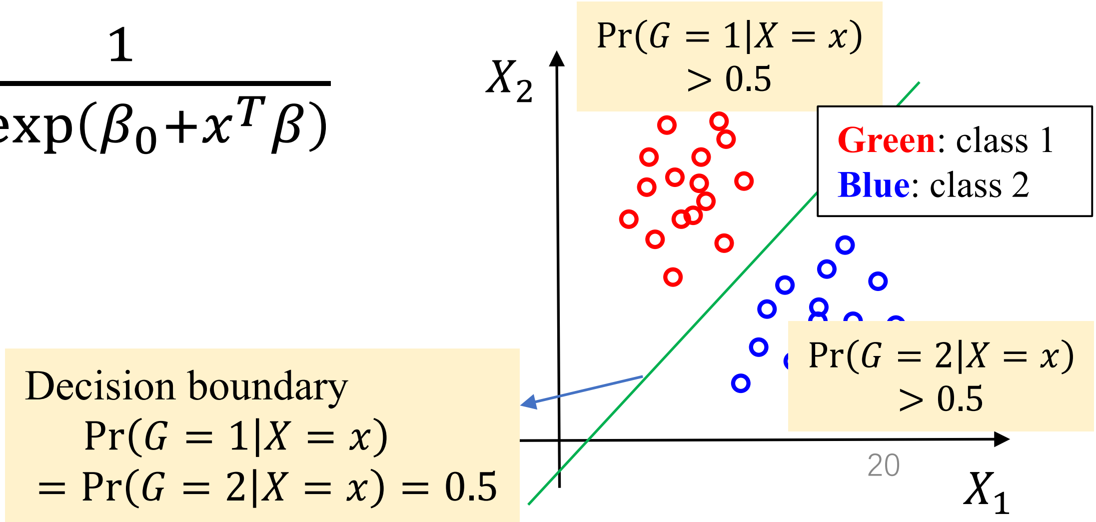
Q

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

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

- Decision boundary

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



Linear Discriminant Analysis

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

- Assumptions in LDA
  1. Model each class density as multivariate Gaussian

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

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

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

- Compare two classes k and ℓ

$Pr(X=x|G=k)\cdot Pr(G=k)/Pr(X=x)$

Logit:

$$\log\frac{\Pr(G=k|X=x)}{\Pr(G=\ell|X=x)} = \log\frac{f_k(x)}{f_\ell(x)} + \log\frac{\pi_k}{\pi_\ell}$$

const

$= 0$

$-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k) + \frac{1}{2}(x-\mu_\ell)^T\Sigma^{-1}(x-\mu_\ell)$

$=$

$$= \boxed{\log\frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k+\mu_\ell)^T\Sigma^{-1}(\mu_k-\mu_\ell) + x^T\Sigma^{-1}(\mu_k-\mu_\ell),}$$

Quadratic term vanished due to the common covariance

Decision boundary is linear w.r.t. X

Linear Discriminant Analysis

- Parameter estimation

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

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

Weighted average

Linear Discriminant Analysis

- Suppose that $\log \frac{\Pr(G=k|X=x)}{\Pr(G=\ell|X=x)} = \boxed{\boxed{\delta_k(x)} - \boxed{\delta_\ell(x)}} = 0$
  - $\delta_k(x) > \delta_\ell(x)$, class k
  - $\delta_k(x) < \delta_\ell(x)$, class ℓ
  - $\delta_k(x) = \delta_\ell(x)$, decision boundary

- 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

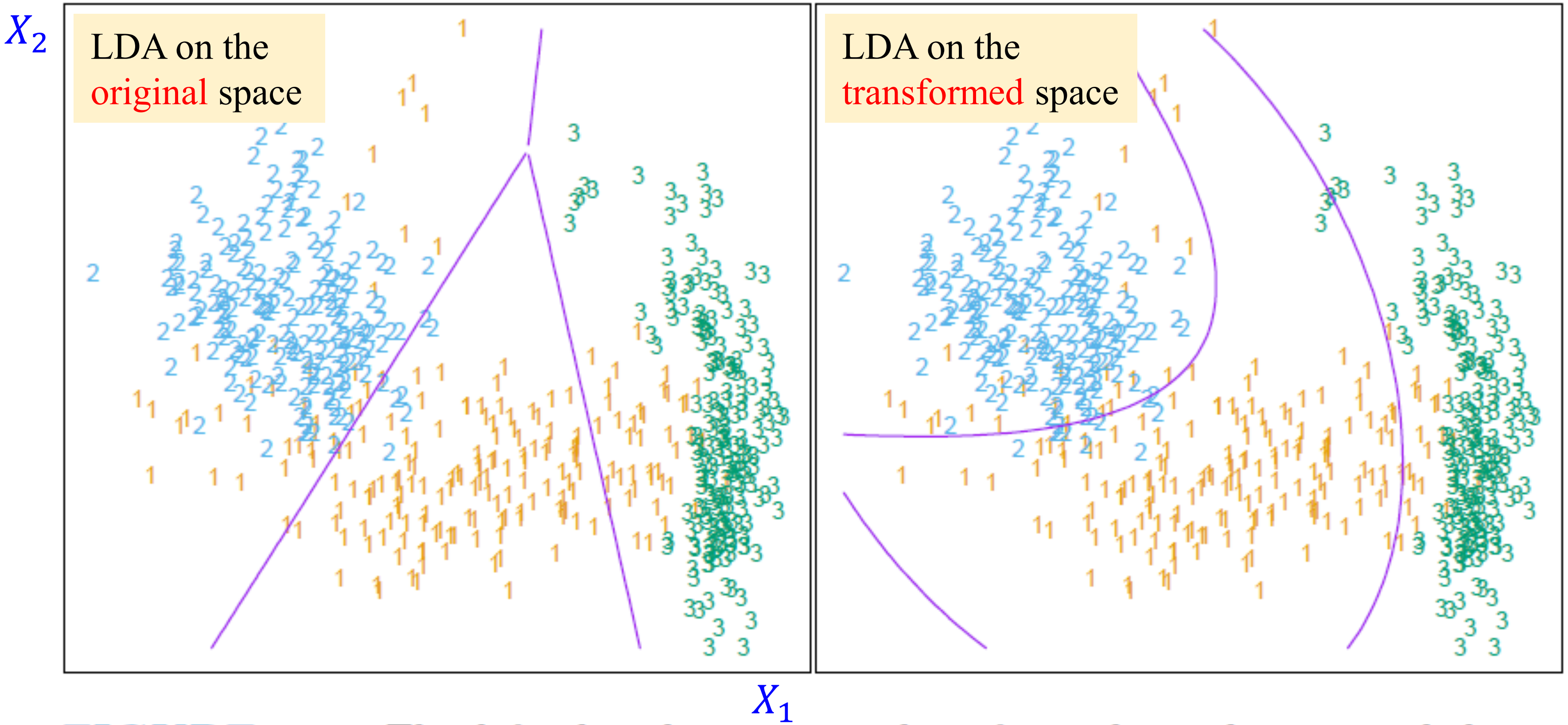


FIGURE 4.1. *The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space* $\boxed{X_1, X_2, X_1X_2, X_1^2, X_2^2.}$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

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

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

- 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

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

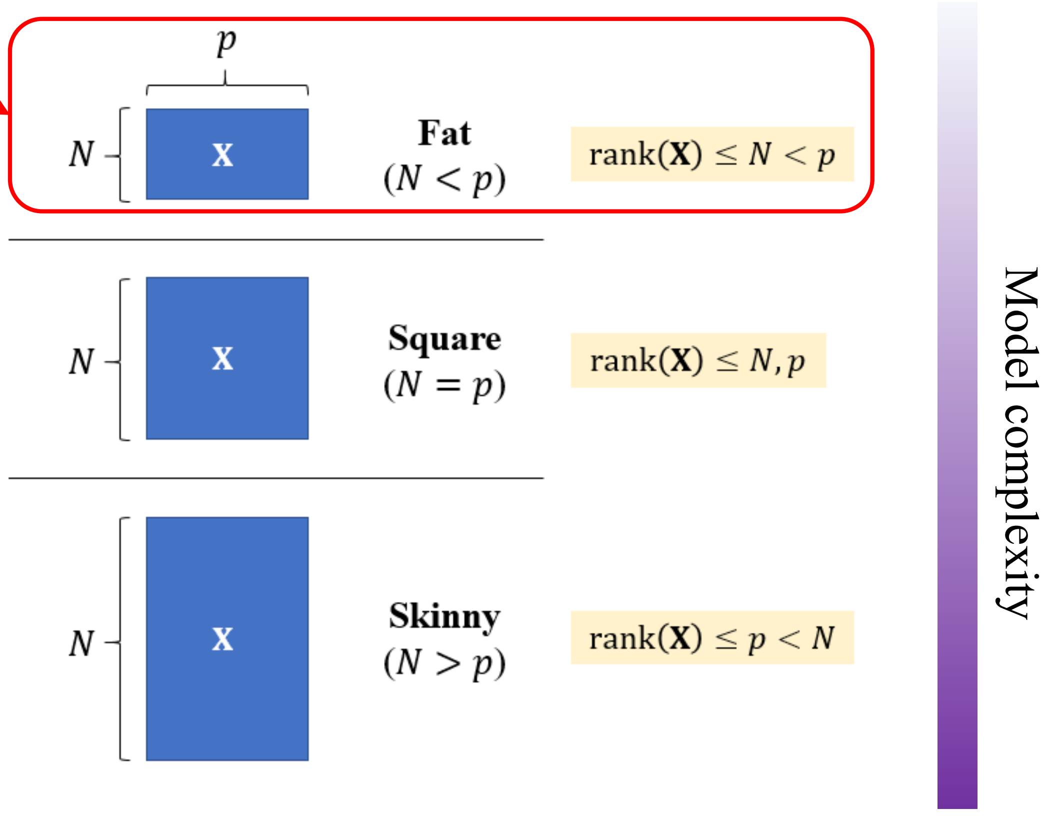
Σ

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

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
      #paras: $K \times p \times p \to K \times p$



Regularized Discriminant Analysis

Regularized LDA (RLDA)

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

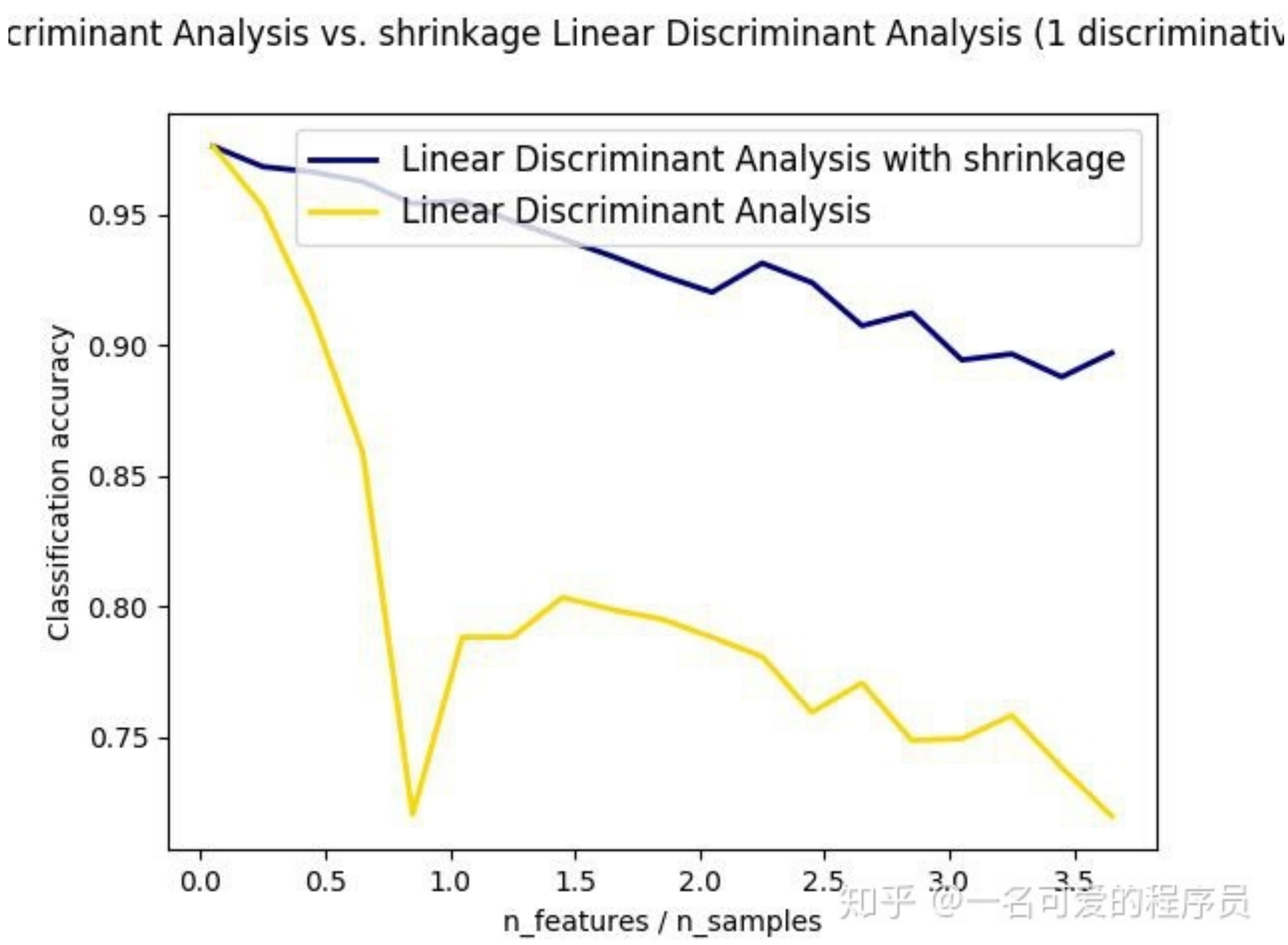
$$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\mathrm{diag}(\hat{\Sigma}), \gamma \in [0,1]$$

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

Diagonal LDA

- Independent assumption on feature dependencies P 独立

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



criminant Analysis vs. shrinkage Linear Discriminant Analysis (1 discriminativ
Linear Discriminant Analysis with shrinkage
Linear Discriminant Analysis
Classification accuracy
0.95
0.90
0.85
0.80
0.75
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
n_features / n_samples
知乎 @一名可爱的程序员

Regularized Discriminant Analysis

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

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

Regularized Discriminant Analysis

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

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

High bias
Low variance

强假设（p个 feature 互相独立）

Low bias
High variance