Introduction to Machine Learning CS182

Lu Sun

School of Information Science and Technology

ShanghaiTech University

October 19, 2023

Today:

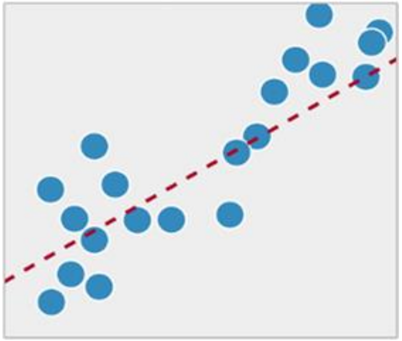
- Linear Methods for Classification I
  - Introduction
  - Linear regression of an indicator matrix
  - Linear discriminant analysis

Readings:

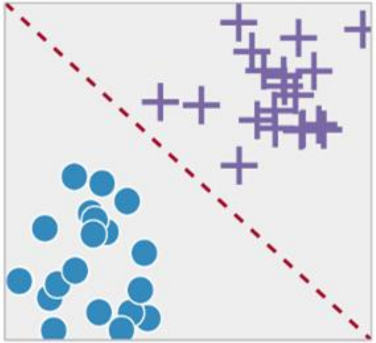
- The Elements of Statistical Learning (ESL), Chapters 4.1, 4.2 and 4.3

Linear Methods for Classification I

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



Regression



Classification

Introduction

Example

Handwritten digits recognition

Input variables

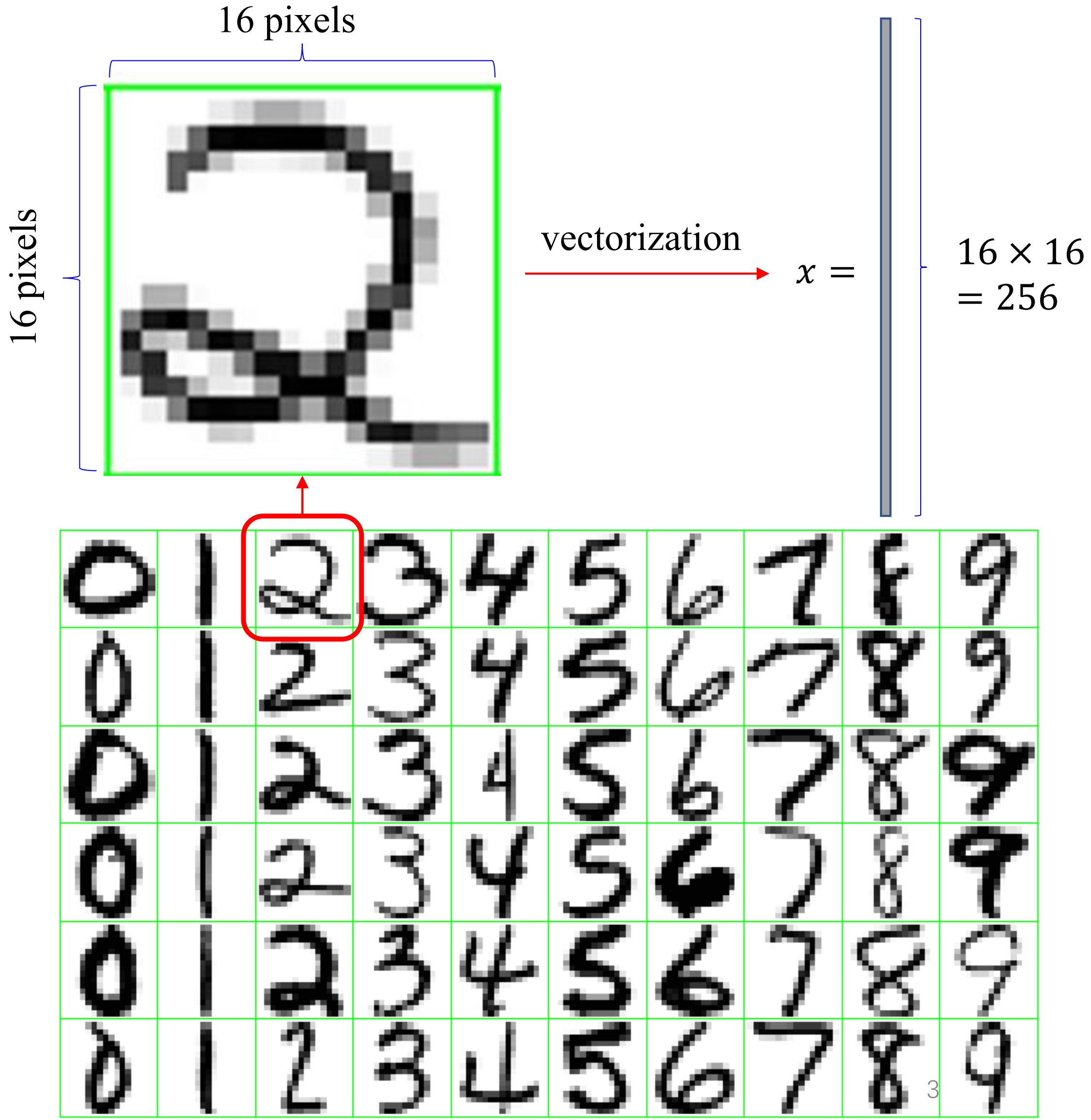
$$X = (X_0, X_1, X_2, \dots, X_{256})^T$$

Categorical output variable $\mathcal{G}$ with values from

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

多类别) 多类

Non-binary (multi-class) classification



Introduction

qualitative: G (categorical or discrete) → **classification**

coding $= 2$ → a binary digit. $\{0,1\}$ or $\{-1,1\}$

coding > 2 → K binary digits. One is “on” at a time.

Example

Handwritten digits recognition

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**

G
0
1
…
9

One-hot coding →

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?

Introduction

Binary classification

- Linear regression

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

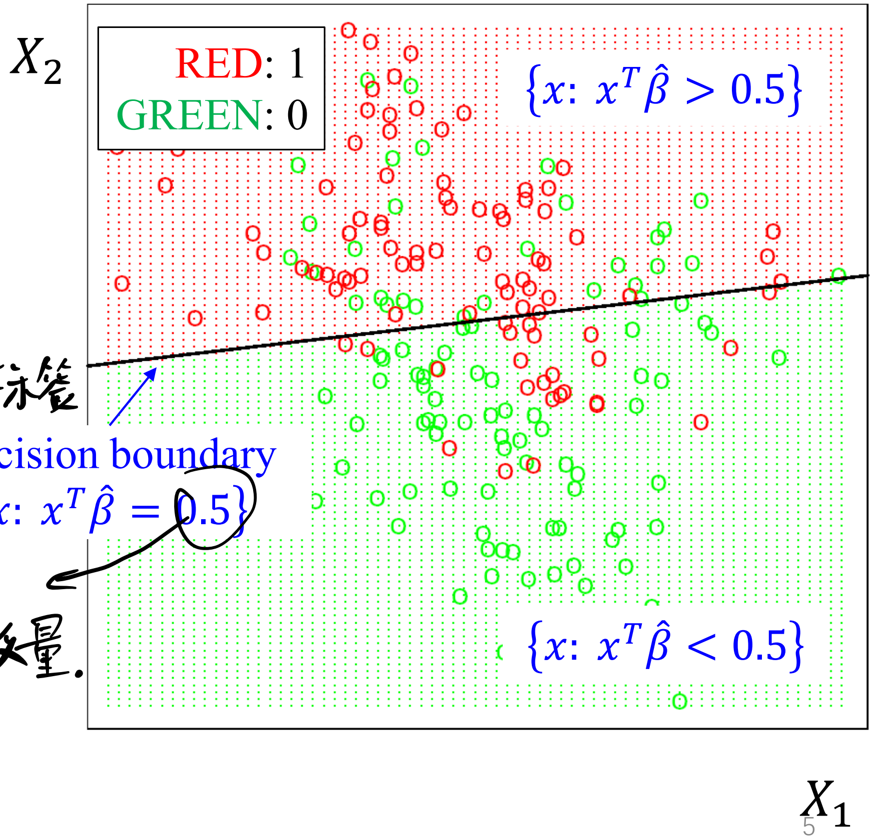
- Least squares solution

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

- Decision boundary

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

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



重标签

可改.

根据数量.

Introduction

Multi-class classification

- Linear regressions for K classes

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

有很多种情况(n²级)

- 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

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

G
0
1
…
9

N

coding →

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

K

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

- Our problem:

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

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

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

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

Linear Regression of an Indicator Matrix

A new observation x is classified by

- Compute the fitted output

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

向量⇒对每个类别输出
⇒找最大值对应的类别

- Classify x according to

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

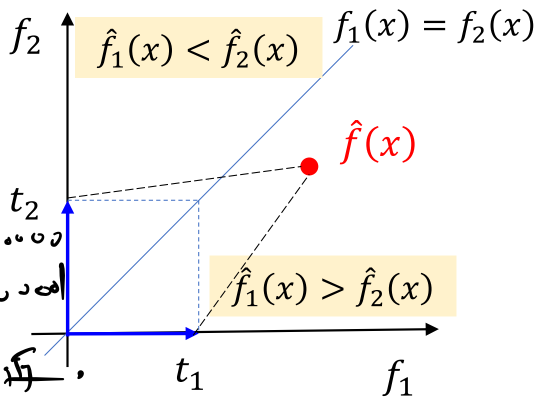
定义编码如 $\{ t_1 = (1,0\cdots 0)$ ⋮ $t_K : \cdots 0\,0\,1$

- Or equivalently,

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

找最近

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 $\mathcal{G} = \{1, \dots, K\}$.

- 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. $\Pr(G, X)$

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

- Pointwise minimization leads to

$$\begin{aligned} \hat{G}(x) &= \underset{k \in \mathcal{G}}{\operatorname{argmin}} \sum_{\ell=1}^{K} L(k, \ell) \Pr(G = \ell | X = x) \\ &= \underset{k \in \mathcal{G}}{\operatorname{argmax}} \boxed{\Pr(G = k | X = x)} \longleftarrow \text{posterior} \end{aligned}$$

Linear Regression of an Indicator Matrix

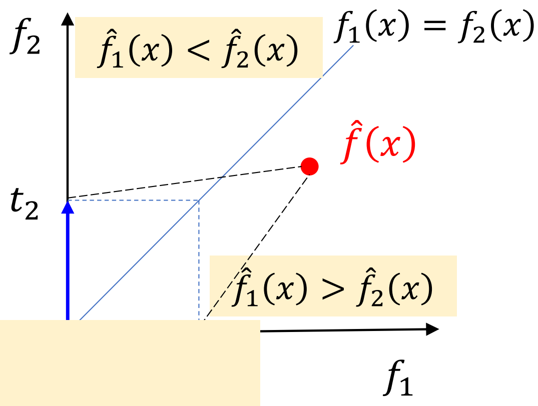
A new observation x is classified by

- 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 $\Pr(G = k | X = x)$?

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

可取负或+∞

将截距吸入

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

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

We have the followings 维度为k的列向量

Indicator matrix

标签矩阵

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

span (col(X))

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

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

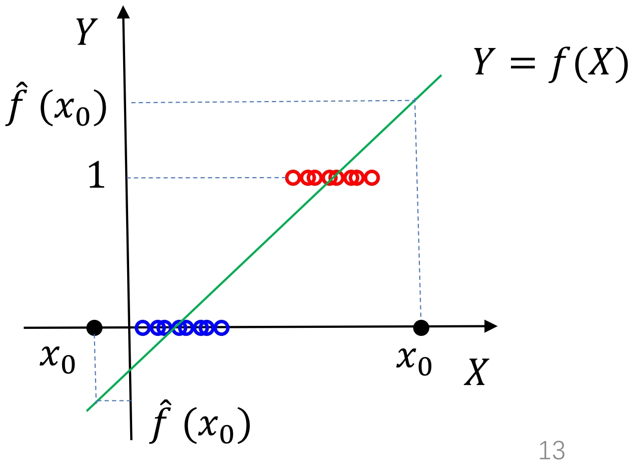
Minimizing EPE:

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

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

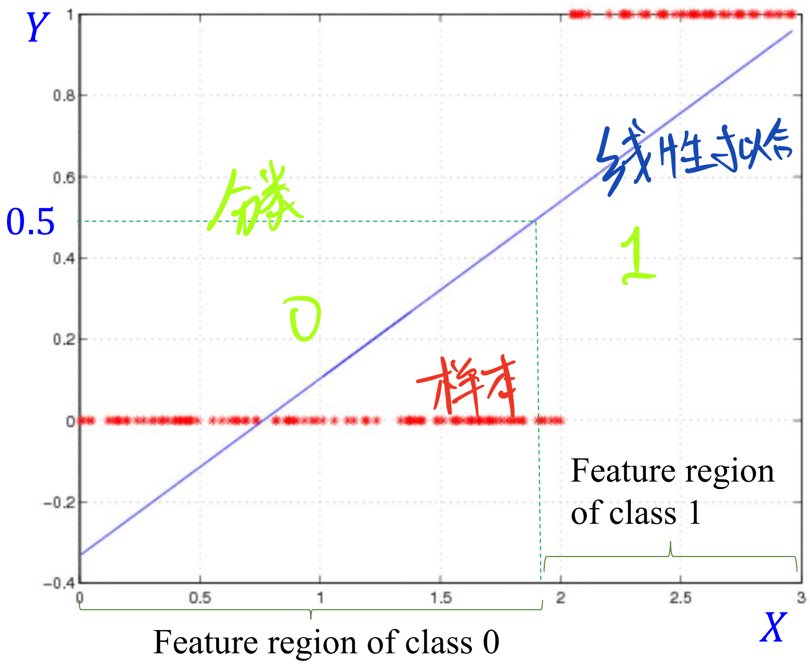


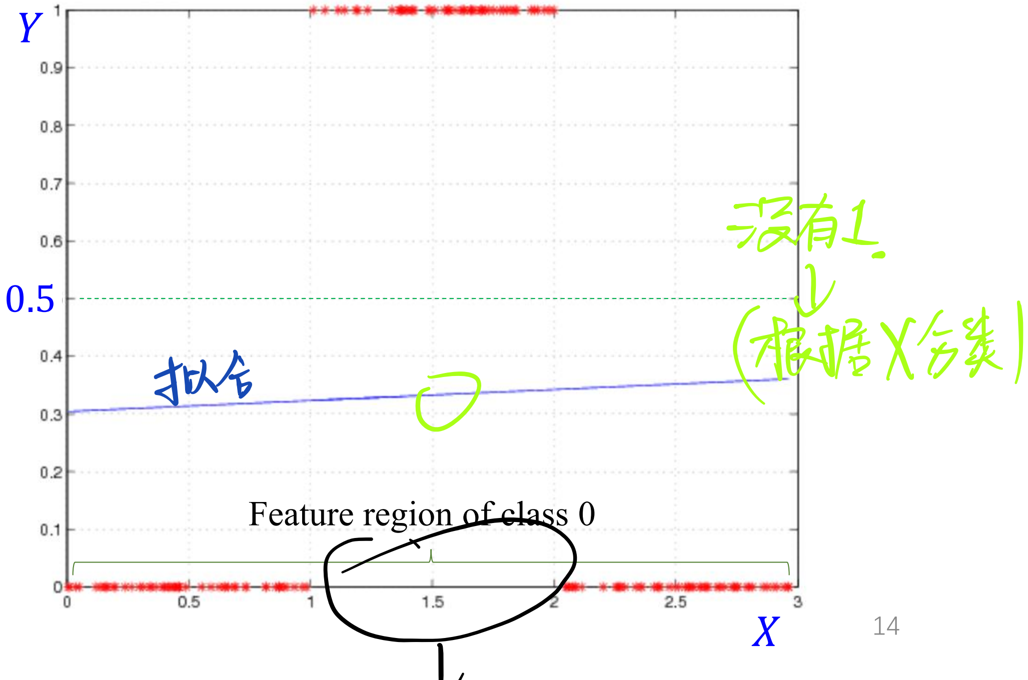
It possibly suffers from the problem of masking

- 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



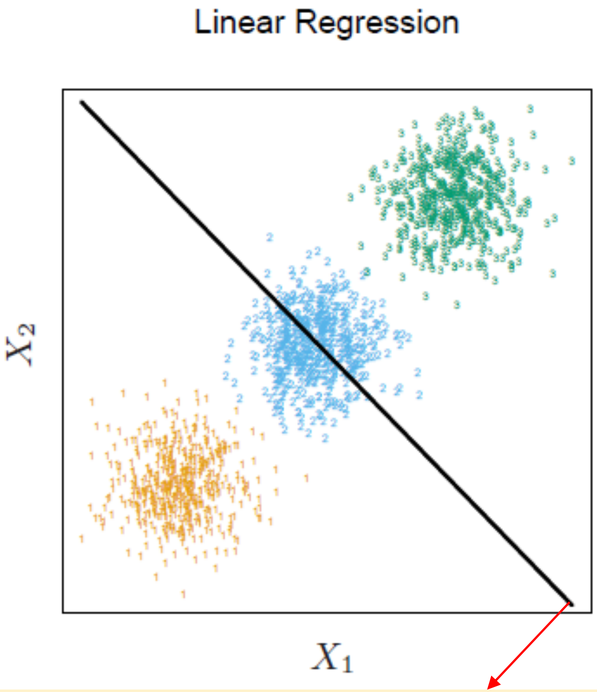


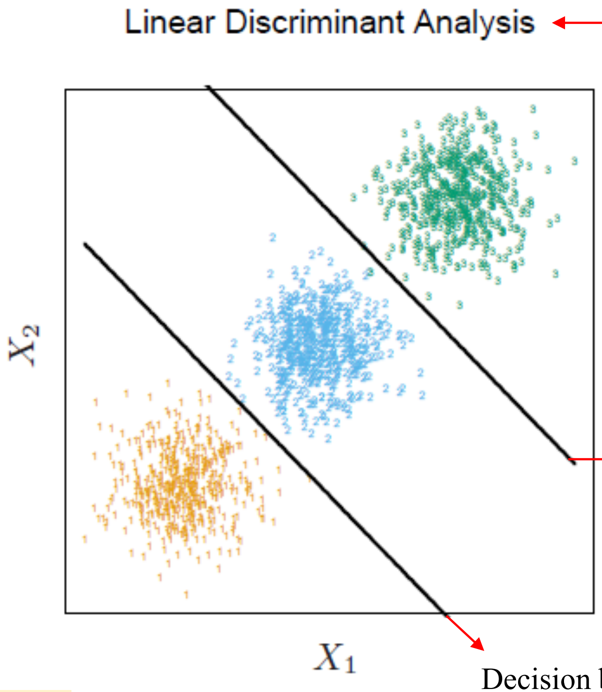
这里会被遮掉.
变成全为0的分类

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 → 找哪个值大就是哪一类

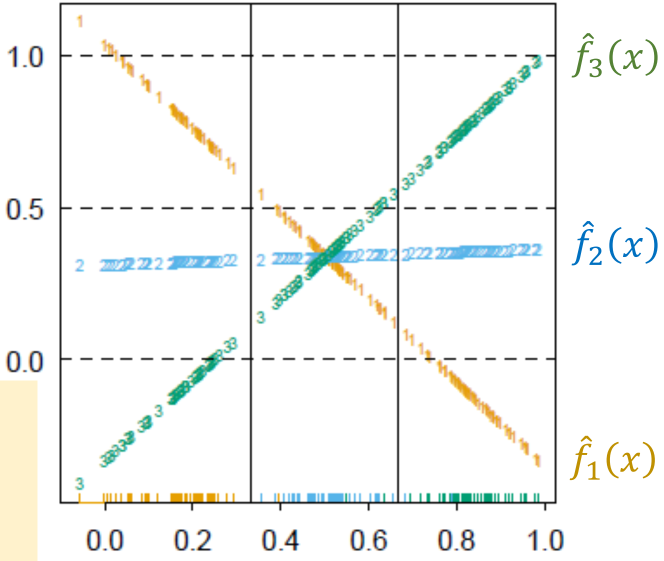
⇒ 不会有蓝色类.

The indicator matrix

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

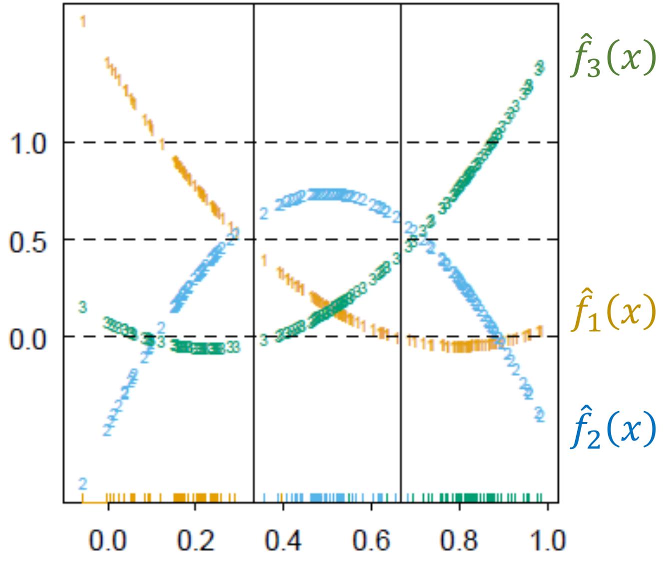
Yellow: class 1
Blue: class 2
Green: class 3

Degree = 1; Error = 0.33



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

Degree = 2; Error = 0.04



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

→ 拓展到非线性空间.

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

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

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.

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

  用gauss分布

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

  常数，可扔

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

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

  - Class prior:

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

$$\mathbb{1}_{x=k} = \begin{cases} 1 & x=k \\ 0 & x\neq k \end{cases}$$

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

伯努力 Bernoulli：$\pi^{x}(1-\pi)^{1-x}$　类别分布 Categorical：$\prod_{k=1}^{K}\pi_k^{\mathbb{1}_{x=k}}$

- It produces LDA (线性), QDA (二次) (quadratic DA), MDA (混合) (mixture DA), kernel DA (核) and naïve Bayes (朴素贝叶斯), under various assumptions on $f_k(x)$

边界：$\{x \mid P(G=k|X=x) = P(G=l|X=x)\}$ $\Leftrightarrow$ $\frac{\Pr(k|x)}{\ldots} = 1$ $\Leftrightarrow$ (贝叶斯) $\frac{\Pr(x|k)P(k)}{\ldots} = 1$

$\Leftrightarrow \ln \frac{P(x|k)P(k)}{P(x|l)P(l)} = 0 \Leftrightarrow$ LDA: $\beta^T X + \beta_0 = 0$

$P(X=x)$约掉了

Linear Discriminant Analysis

$p \in (0,1) \rightarrow$ LDA $\in (-\infty, +\infty)$

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

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

     p维向量

     高维 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 $\Sigma_k = \Sigma, \forall k$

     假设：任意k，Σ_k相互相等.

- Compare two classes k and ℓ

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

代入 $f_k(x)$

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

Quadratic term vanished due to the common covariance

另有一个x，是关于x的线性

Decision boundary is linear w.r.t. X

原有二次项，但因为有假设故消掉了

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)

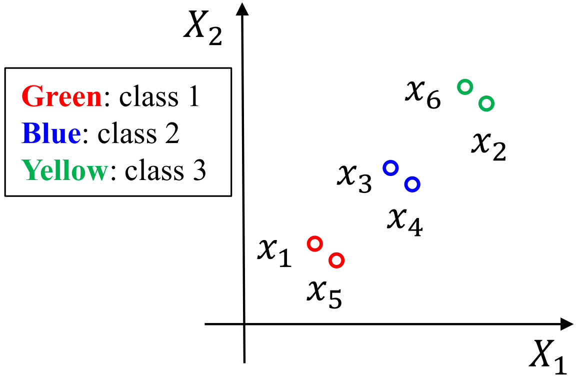
$$\hat{\Sigma} = \frac{(N_1-1)\hat{\Sigma}_1+(N_2-1)\hat{\Sigma}_2+\cdots+(N_K-1)\hat{\Sigma}_K}{(N_1-1)+(N_2-1)+\cdots+(N_K-1)}, \text{where } \hat{\Sigma}_k = \frac{\sum_{g_i=k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T}{N_k-1}$$

Weighted average

不是N_k，保证无偏估计，

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



- Class prior

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

- Class-specific sample mean

二维空间 ⟹ 二维向量

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

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

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

- Common covariance

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

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

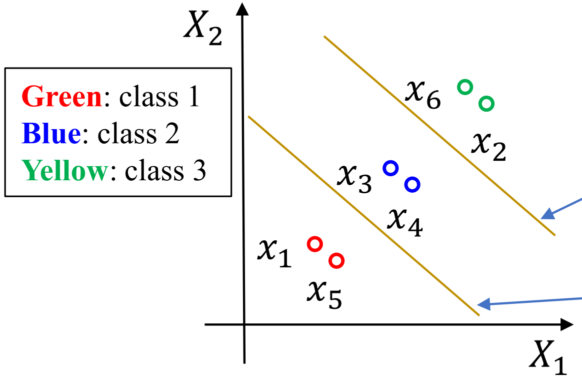
Linear Discriminant Analysis

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

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

- Decision boundary 1-2: $\{(x_1, x_2) | x_1 + x_2 = 0.75\}$
- Similarly, decision boundary 2-3: $\{(x_1, x_2) | x_1 + x_2 = 1.25\}$



$\{(x_1, x_2) | x_1 + x_2 = 1.25\}$

$\{(x_1, x_2) | x_1 + x_2 = 0.75\}$

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

线性判别函数：算每个类别的 δ_k 看哪个最大。

- Linear discriminant functions

Constant

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

Classify to class k that maximizes the discriminant function

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

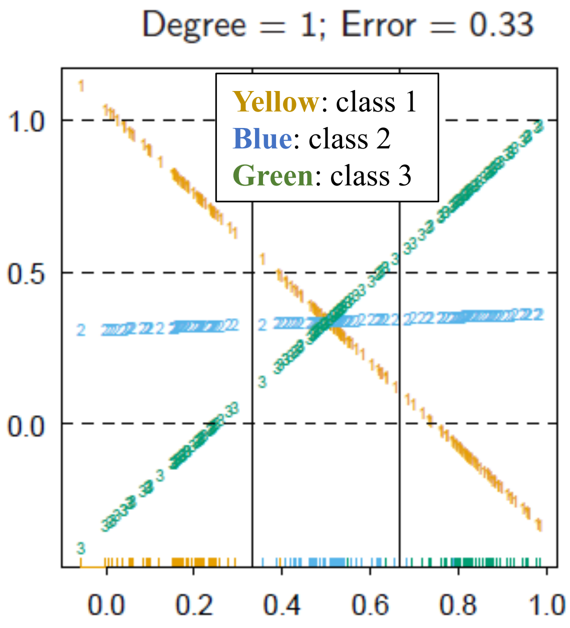
Any difference?

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

Linear Discriminant Analysis

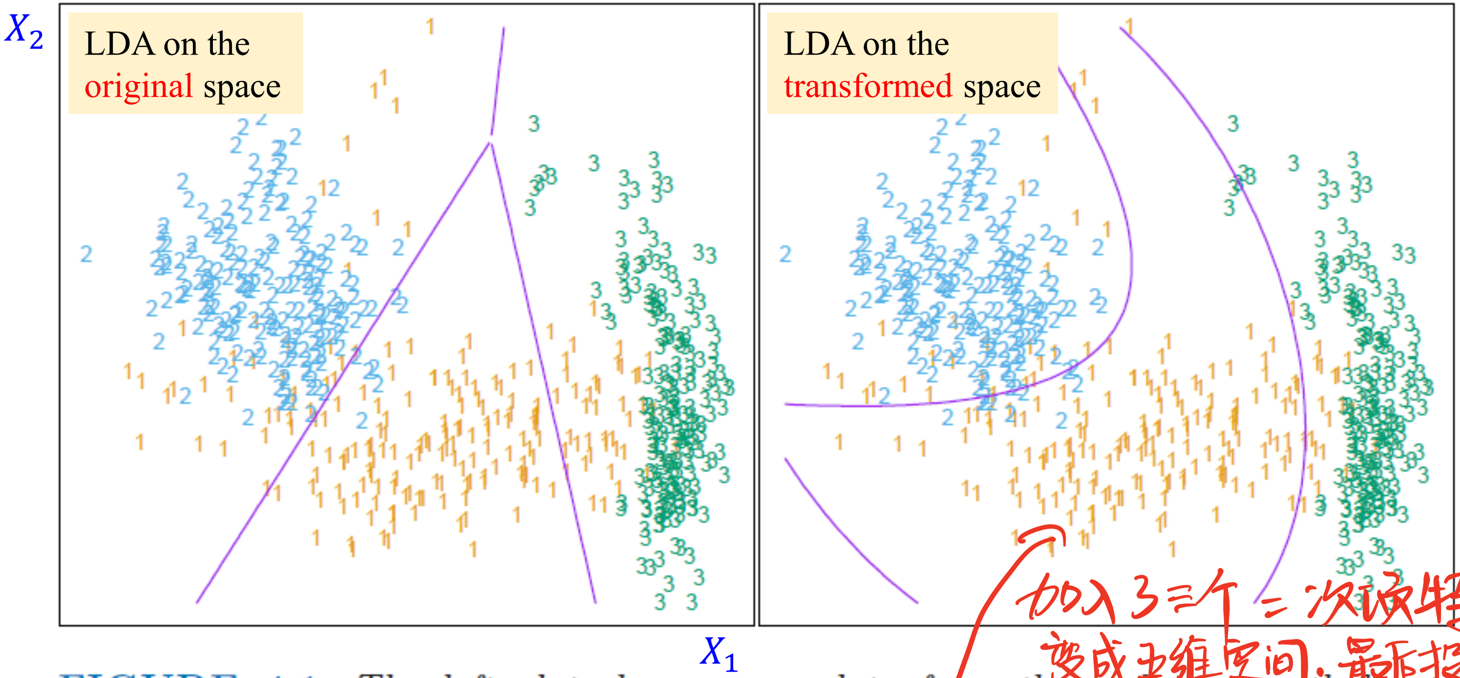
- Binary classification ($K = 2$)
  - Correspondence between LDA and linear classification

- 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



加入3三个二次项特征.
变成五维空间，最后投影回二维.

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* $X_1, X_2, X_1X_2, X_1^2, X_2^2$. *Linear inequalities in this space are quadratic inequalities in the original space.*

取哪些二次项是一个超参数.

Linear Discriminant Analysis

一定会交于一点.

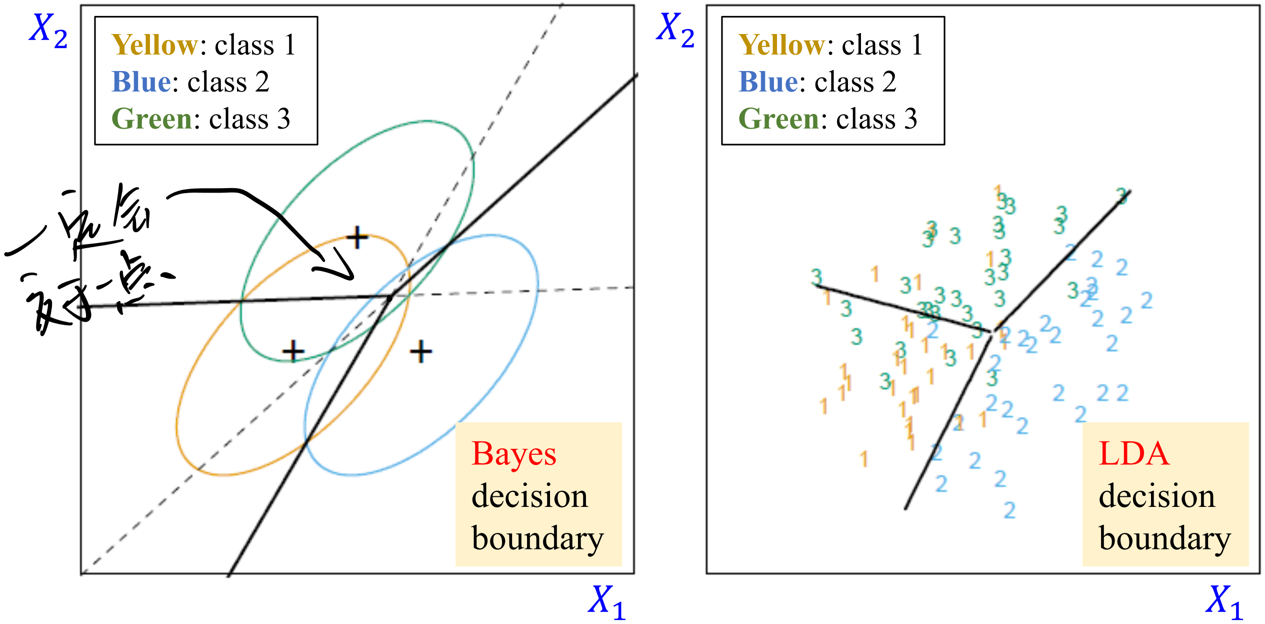


FIGURE 4.5. *The left panel shows three Gaussian distributions, with the same covariance and different means. Included are the contours of constant density enclosing 95% of the probability in each case. The Bayes decision boundaries between each pair of classes are shown (broken straight lines), and 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.*

特征表达更好但需要计算的更多.(相对LDA)

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

只估计 π, μ, Σ

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

需要K个Σ

Quadratic Discriminant Analysis

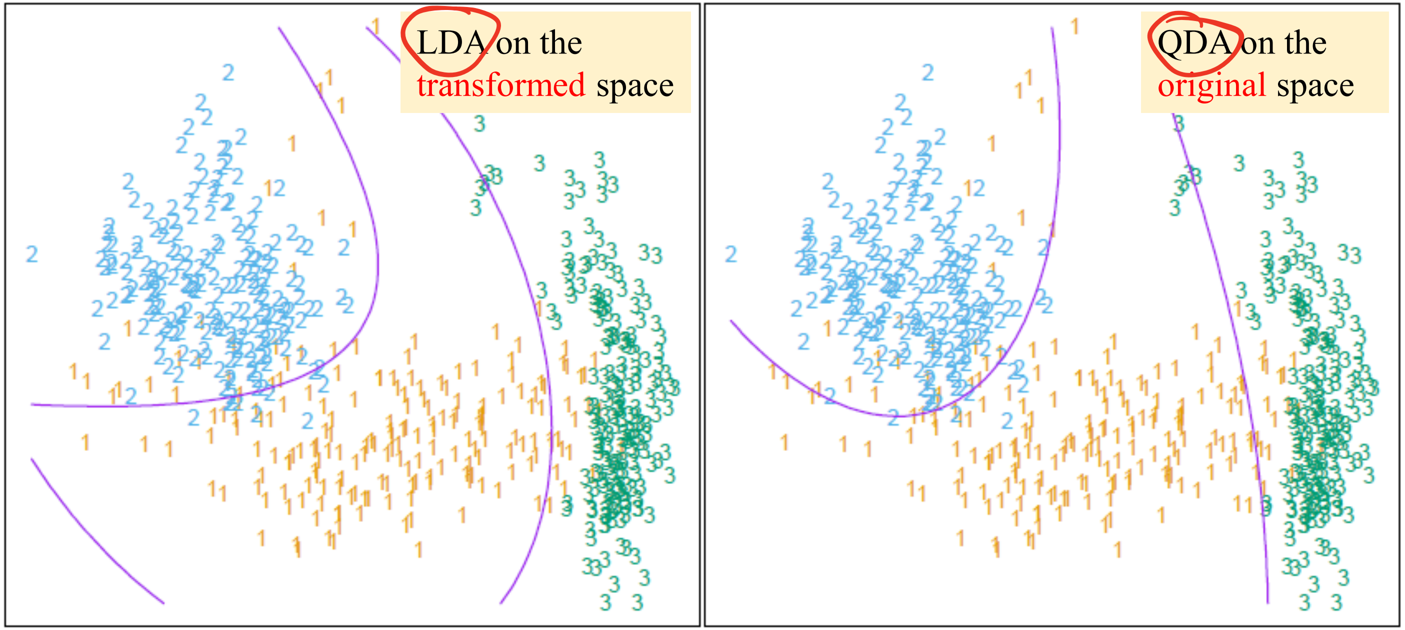


FIGURE 4.6. *Two methods for fitting quadratic boundaries. The left plot shows the quadratic decision boundaries for the data in Figure 4.1 (obtained using LDA in the five-dimensional space* $\boxed{X_1, X_2, X_1X_2, X_1^2, X_2^2}$*). The right plot shows the quadratic decision boundaries found by QDA.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

Summary

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

- Linear discriminant analysis
  - Logit transformation: $\text{logit}(\Pr(x)) = \log\left(\frac{\Pr(x)}{1-\Pr(x)}\right)$
  - 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