Classification

Discriminant analysis

- 방법 : 각 범주의 분포를 개별적으로 모델링한 후 베이즈 정리 (Bayes theorem) 를 사용하여 확률을 구함
- 판별분석 (Discriminant analysis) : 각 범주의 분포로 정규분포를 사용
- 다른 분포 사용 가능

Bayes theorem for classification

- Bayes Theorem :

$$P(Y = k|X = x) = \frac{P(X = x|Y = k) \cdot P(Y = k)}{P(X = x)}$$

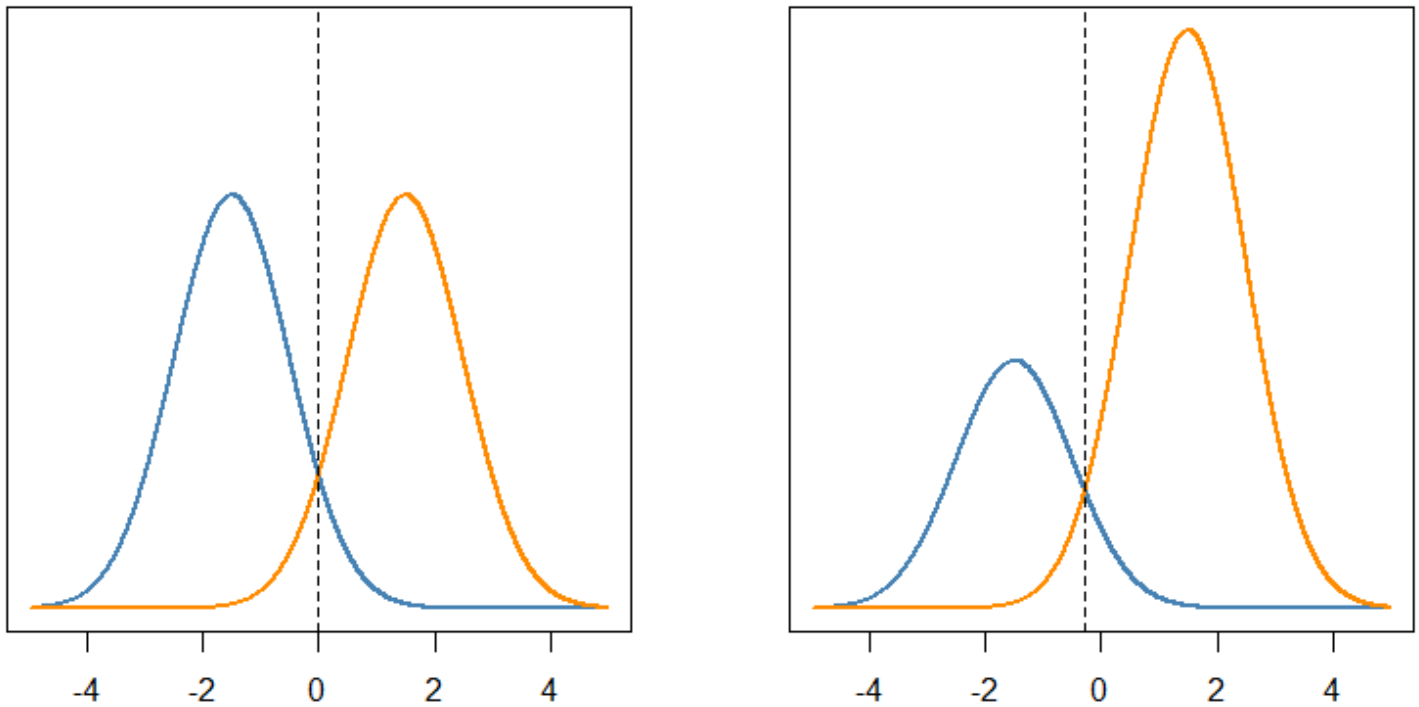
- Discriminant analysis form :

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

  - $f_k(x) = P(X = x|Y = k)$: k번째 범주에서의 X의 확률밀도함수(probability density function)
  - $\pi_k = P(Y = k)$: k번째 범주의 사전확률(prior probability)

Classify to the highest density

Figure: Left : $\pi_1 = 0.5, \pi_2 = 0.5$, Right : $\pi_1 = 0.3, \pi_2 = 0.7$



새로운 데이터에 대해 밀도함수값을 가장 크게 해주는 범주로 분류

Logistic Regression vs. Discriminant Analysis

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis does not suffer from this problem.
- If n is small and the distribution of the predictors X is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model.
- Linear discriminant analysis is popular when we have more than two response classes, because it also provides low-dimensional views of the data.

Linear Discriminant Analysis (LDA) ($p = 1$)

- The Gaussian density

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

  μ_k, σ_k^2 : k번째 범주의 평균과 분산

- 가정 : $\sigma_k = \sigma$, 모든 범주의 분산은 동일

- posterior probability :

$$p_k(x) = P(Y = k|X = x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}}$$

판별함수

- $X = x$일때, $p_k(x)$ 값을 가장 크게하는 범주 k로 분류
- 판별함수(discriminant function)

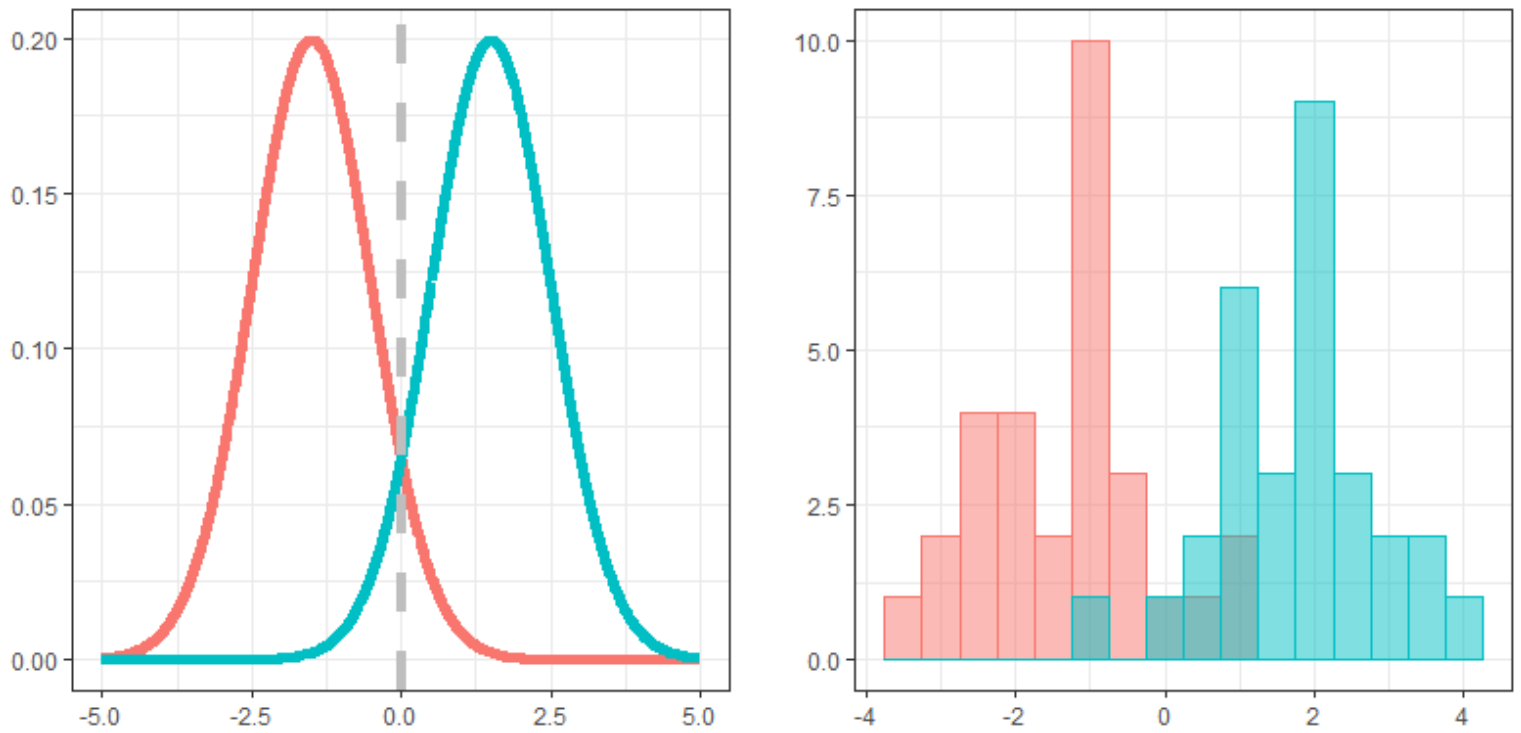
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

- 판별점수(discriminant score)를 가장 크게 하는 범주 k로 분류
- $\delta_k(x)$: x에 대한 선형함수
- $K = 2, \pi_1 = \pi_2 = 0.5$ 일 때, decision boundary

$$x = \frac{\mu_1 + \mu_2}{2}$$

Example

Figure: $\pi_1 = \pi_2 = 0.5, \mu_1 = -1.5, \mu_2 = 1.5, \sigma^2 = 1$

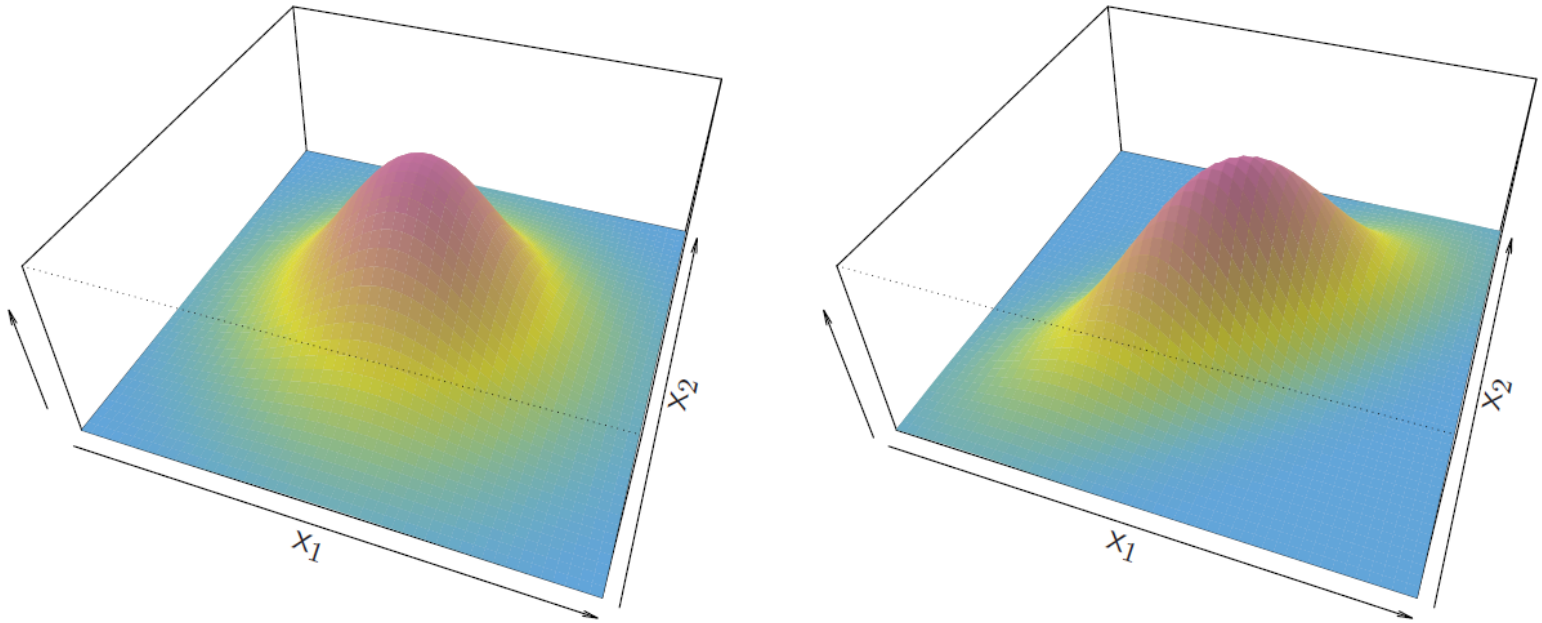


모수추정

$$
\begin{aligned}
\hat{\pi}_k &= \frac{n_k}{n} \\
\hat{\mu}_k &= \frac{1}{n_k} \sum_{i:y_i=k} x_k \\
\hat{\sigma}^2 &= \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2 \\
&= \sum_{k=1}^{K} \frac{n_k - 1}{n-K} \hat{\sigma}_k^2
\end{aligned}
$$

where $\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$.

LDA ($p > 1$)



- Density : $f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$
- Discriminant function :

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

Example : $p = 2, K = 3$

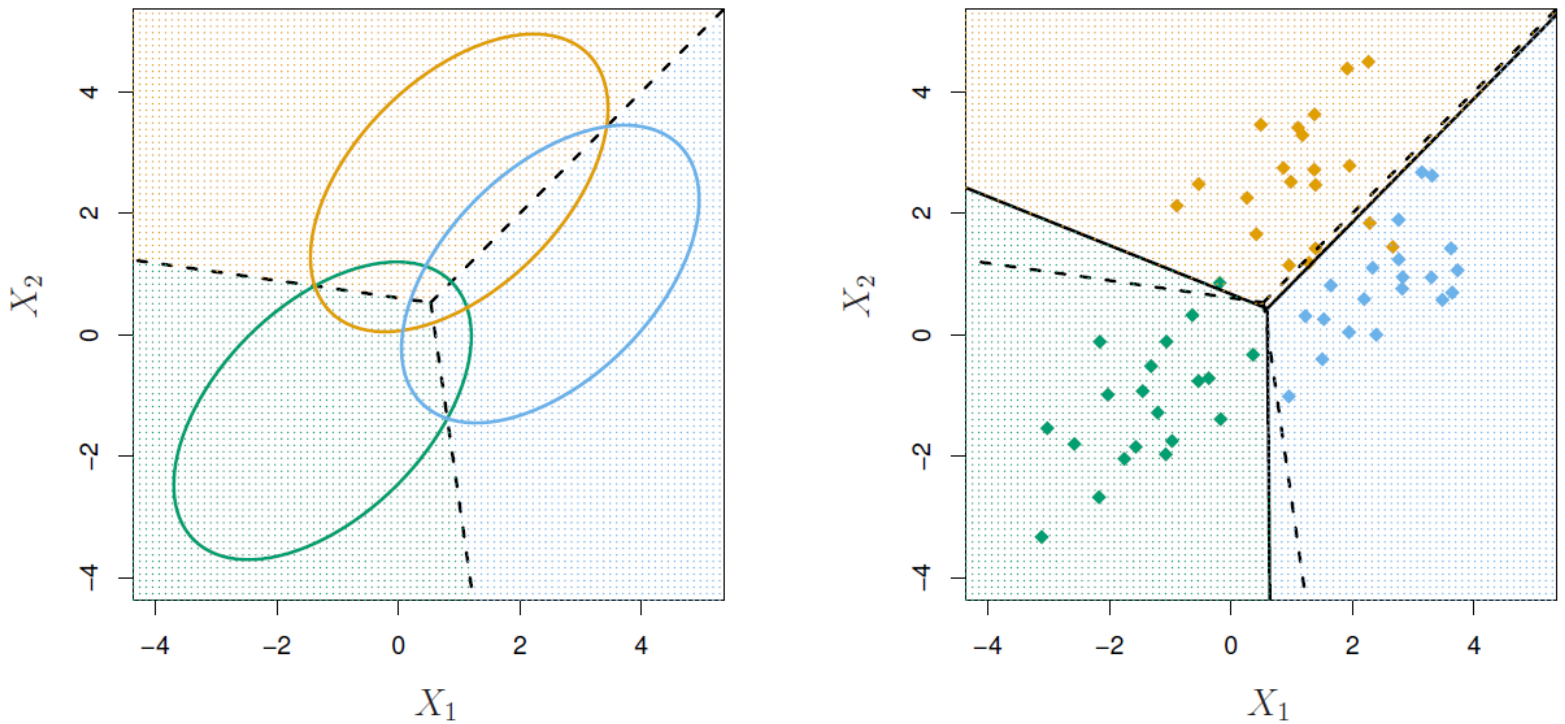


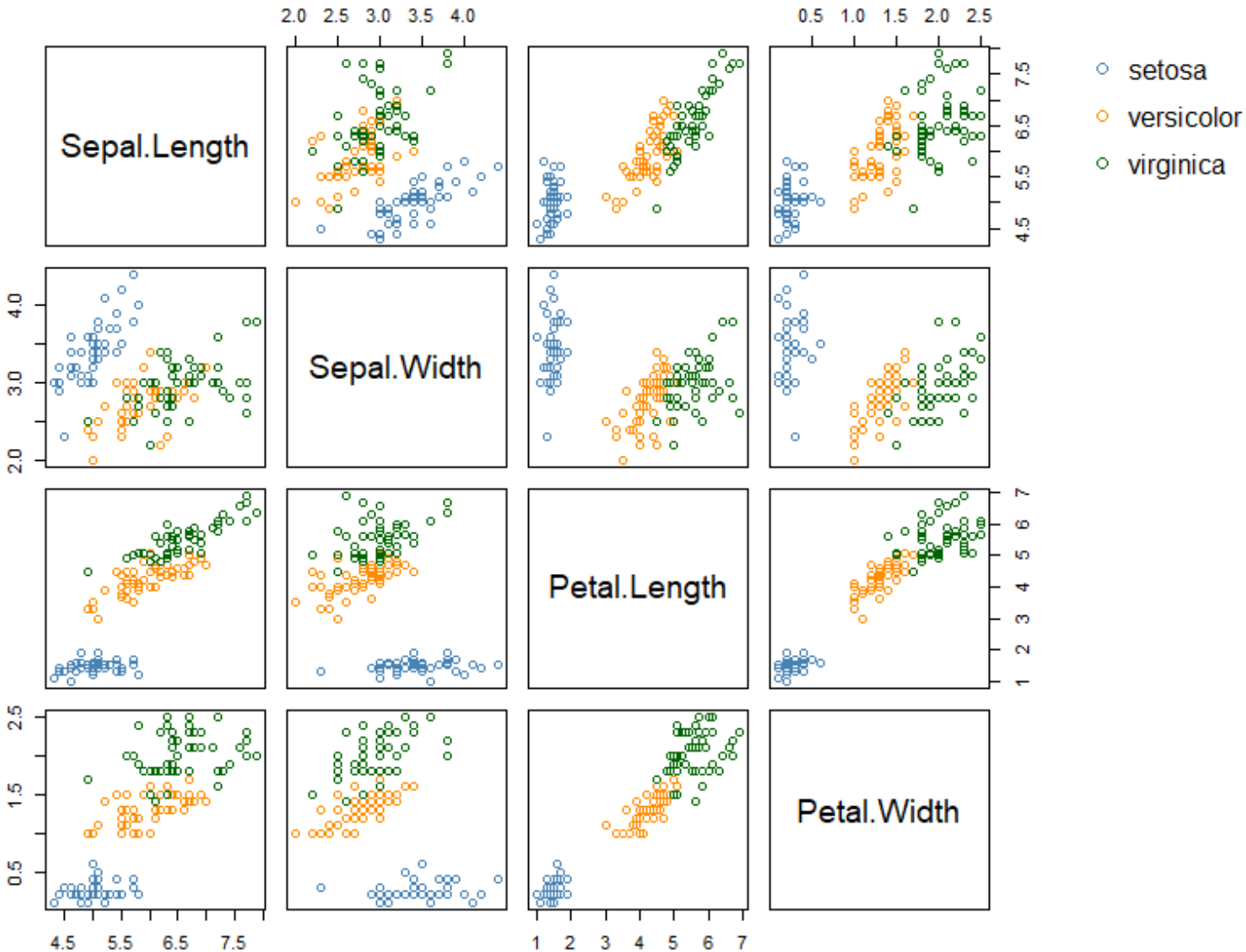
Figure: $\pi_1 = \pi_2 = \pi_3 = 1/3$

Example : IRIS data

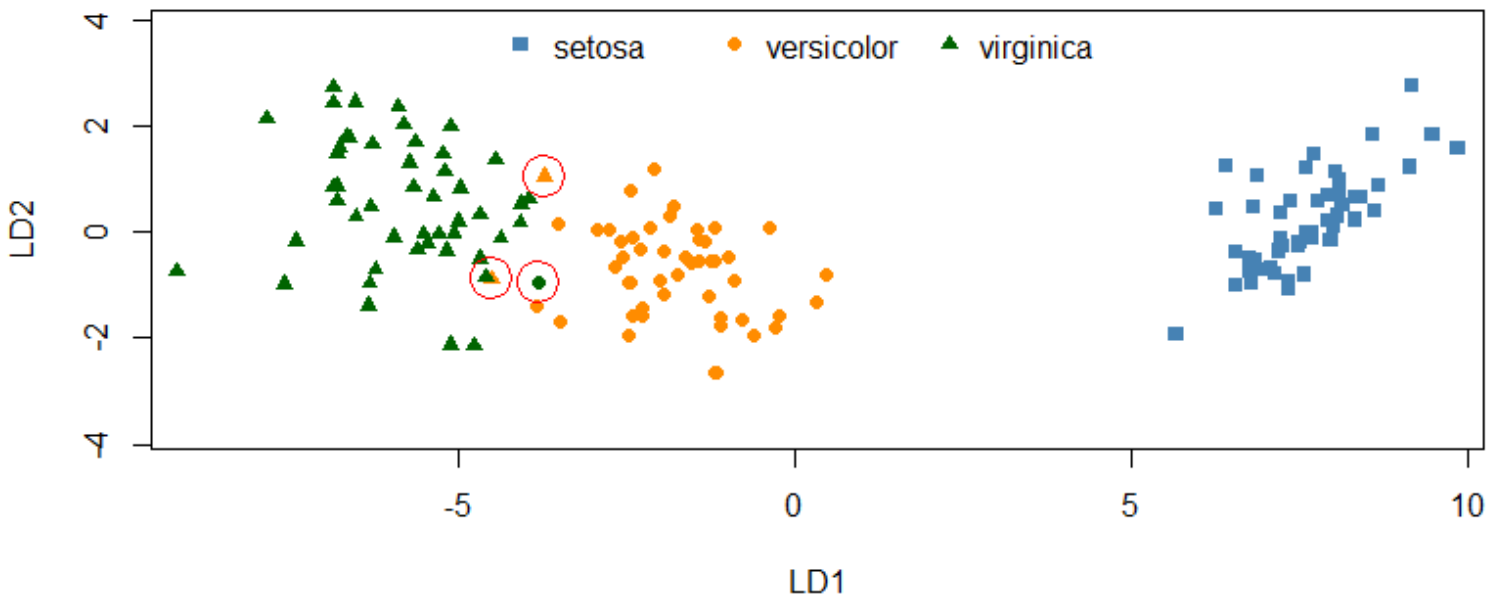
Table: Iris data

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.10	3.50	1.40	0.20	setosa
2	4.90	3.00	1.40	0.20	setosa
3	4.70	3.20	1.30	0.20	setosa
4	4.60	3.10	1.50	0.20	setosa
5	5.00	3.60	1.40	0.20	setosa
6	5.40	3.90	1.70	0.40	setosa
⋮	⋮	⋮	⋮	⋮	⋮

Example : IRIS data



Example : IRIS data - LDA



LDA

- When there are K classes, linear discriminant analysis can be viewed exactly in a $K-1$ dimensional plot.
- Why? Because it essentially classifies to the closest centroid, and they span a $K-1$ dimensional plane.
- Even when $K>3$, we can find the "best" 2-dimensional plane for visualizing the discriminant rule.

LDA

- $\widehat{p_k(x)} = \widehat{P}(Y = k|X = x) = \dfrac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}$
- $\hat{\delta}_k(x)$를 가장 크게하는 범주로 분류하는 것은 사후확률을 가장 크게 하는 범주로 분류하는 것과 동일
- $K = 2$ 인 경우, $\widehat{P}(Y = 1|X = x) \geq 0.5$ 이면 범주1로 분류, 아니면 범주2로 분류

LDA on Credit data

		예측결과		Sum
		No	Yes	
관측값	No	9643	24	9667
	Yes	257	76	333
Sum		9900	100	10000

- default $\sim$ Balance
- $\widehat{P}(Y = 1|X = x) \geq 0.5$ 이면 'Yes'로 분류
- 오분류율 (misclassification rate) = (24+257)/10000 = 2.81%

Type of errors

		Predicted class		Sum Total
		Negative	Positive	
True class	Negative	TN	FP	N
	Positive	FN	TP	P
Sum		N^*	P^*	

- 오분류율 $= \dfrac{FP + FN}{N + P}$

Type of errors

- Type I error (False Positive rate) $= \frac{\mathrm{FP}}{\mathrm{N}}$
- Specificity $=$ 1-Type I error $= \frac{\mathrm{TN}}{\mathrm{N}}$
- Sensitivity (recall, power, 1-Type II error) $= \frac{\mathrm{TP}}{\mathrm{P}}$
- Precision $= \frac{\mathrm{TP}}{\mathrm{P}^*}$

Type of errors

- 질병 양성(Positive), 음성(Negative) 판단
- Specificity = 실제 음성인 사람을 음성이라고 판단할 확률
- Sensitivity = 실제 양성인 사람을 양성이라고 판단할 확률
- Precision = 양성이라고 판단한 경우 중 실제 양성인 사람의 비율

Type of errors

- Type I error (False Positive rate) $= \frac{24}{9667} = 0.25\%$
- Specificity $=$ 1-Type I error $= \frac{9646}{9667} = 99.78\%$
- Sensitivity $= \frac{76}{333} = 22.82\%$
- Precision $= \frac{76}{100} = 76\%$

LDA on Credit data

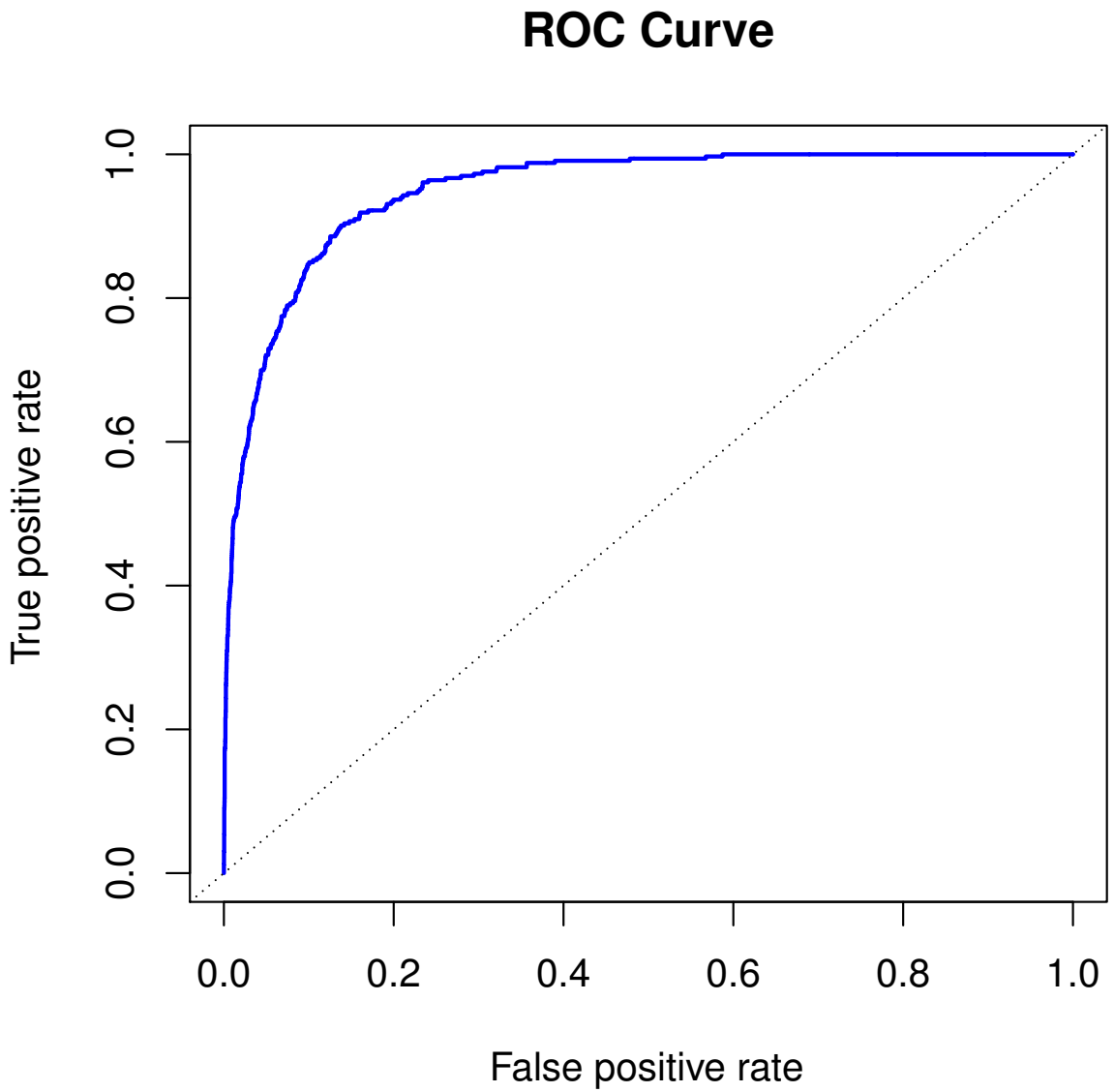
		예측결과		Sum
		No	Yes	
관측값	No	9431	236	9667
	Yes	138	195	333
Sum		9569	431	10000

- default $\sim$ Balance
- $\widehat{P}(Y = 1|X = x) \geq 0.2$ 이면 'Yes'로 분류
- 오분류율 (misclassification rate) = (236+138)/10000 = 3.74%

Type of errors

- Type I error (False Positive rate) $= \frac{236}{9667} = 2.44\%$
- Specificity $=$ 1-Type I error $= \frac{9431}{9667} = 97.56\%$
- Sensitivity $= \frac{195}{333} = 58.56\%$
- Precision $= \frac{195}{431} = 45.24\%$

ROC Curve


ROC Curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

다른 형태의 판별 분석

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

- LDA : $f_k(x)$ - Gaussian densities, 각 그룹의 공분산은 동일하다고 가정

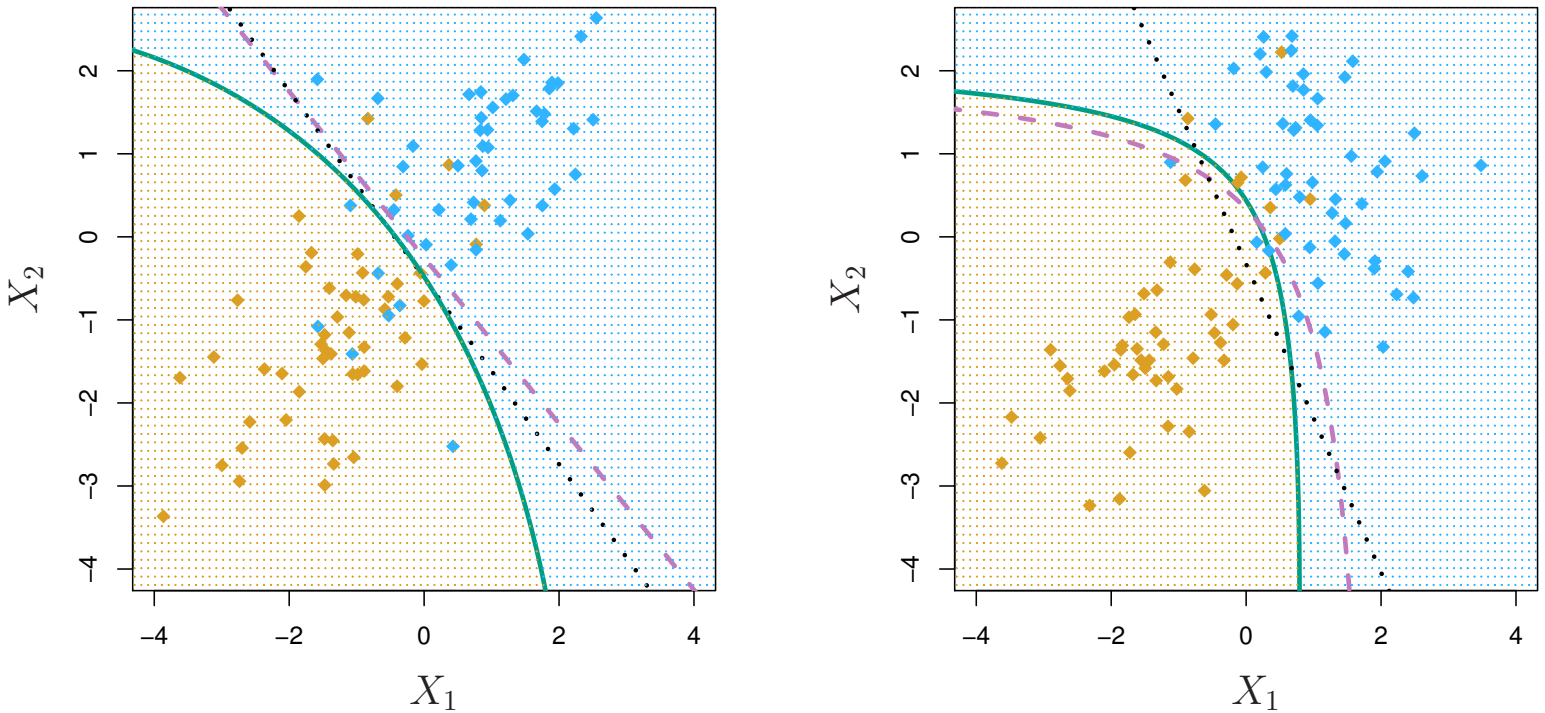
$$\Sigma_1 = \cdots = \Sigma_K = \Sigma$$

- QDA (quadratic discriminant analysis) :

  $f_k(x)$ - Gaussian densities, 공분산 같지 않음

- naive Bayes (conditional independent)

QDA



$$\delta_k(x) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) + \log \pi_k - \frac{1}{2}\log |\Sigma_k|$$

Naive Bayes

- 설명변수들이 독립이라고 가정 : $f_k(x) = \prod_{j=1}^{p} f_{jk}(x_j)$
- p가 커서 DQA나 LDA를 사용하기 어려울 때 유용
- 설명변수가 양적자료가 아닌 범주형 자료일 때도 사용 가능
- Gaussian naive Bayes (가정 : Σ - diagonal)

$$\begin{aligned} \delta_k(x) &\propto \log\left[\pi_k \prod_{j=1}^{p} f_{jk}(x_j)\right] \\ &= -\frac{1}{2}\sum_{j=1}^{p}\left[\frac{(x_j-\mu_{kj})^2}{\sigma_{kj}^2} + \log\sigma_{kj}^2\right] + \log\pi_k \end{aligned}$$

Naive Bayes

- 예) 문서분류

  - $w_1, \ldots, w_N$: 특정 문서에 들어 있는 단어
  - 이 단어들이 포함된 문서가 임의의 범주 k에 속할 확률

$$
\begin{aligned}
p(Y = k | w_1, \ldots, w_N) & \\
&= \frac{p(w_1, \ldots, w_N | Y = k) P(Y = k)}{p(w_1, \ldots, w_N)} \\
&\propto p(w_1, \ldots, w_n | Y = k) P(Y = k) \\
&= p(w_1 | Y = k) \times \cdots \times p(w_N | Y = k) \times P(Y = k)
\end{aligned}
$$

Example

- 영화 장르 분류

	words	Class
1	fun, couple, love, love	Comedy (C)
2	fast, furious, shoot	Action (A)
3	couple, fly, fast, fun, fun	Comedy
4	furious, shoot, shoot, fun	Action
5	fly, fast, shoot, love	Action

- 새로 주어진 영화 소개서가 (fast, fun)이라는 단어를 포함한다면, 이 영화는 어떤 장르로 분류될 수 있을까?

Summary

- Logistic regression is very popular for classification, especially when $K = 2$.
- LDA is useful when n is small, or the classes are well separated, and Gaussian assumptions are reasonable. Also when $K > 2$.
- Naive Bayes is useful when p is very large.