Machine Learning

Ch4. Classification

김광수

2024년 2학기

분류 (classification)

- 반응변수가 질적(qualitative) / 범주형(categorical) 변수
- 예 - 응급실에 오는 사람들을 증상을 바탕으로 세 가지 범주로 분류함.

  - 사용자의 IP주소, 과거 거래내역 등을 근거로 온라인 거래 승인여부를 결정함.

  - 어떤 DNA 돌연변이가 질병의 발생에 유해한지를 알아냄.

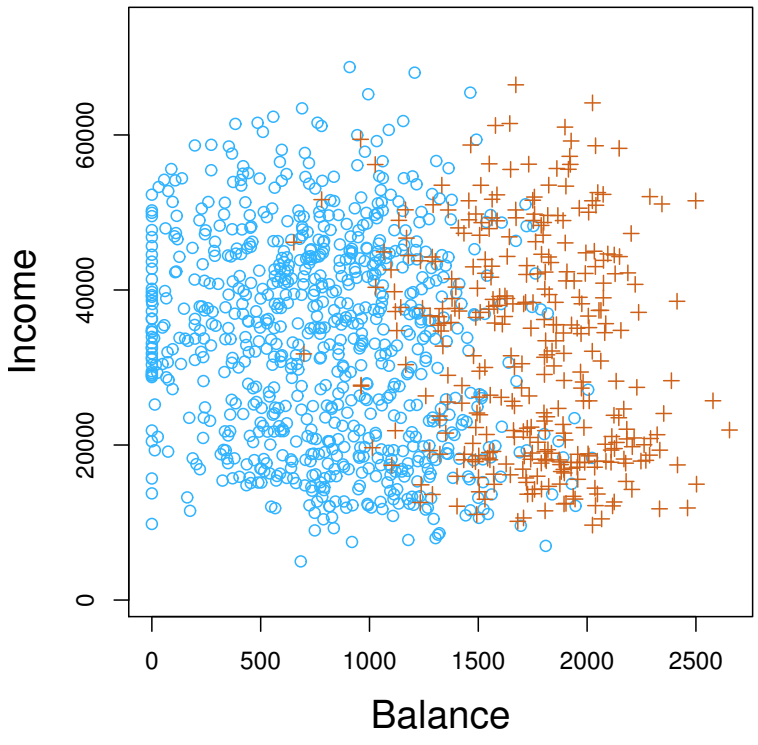
  - 이미지를 보고 개/고양이, 질병유무등을 판별해냄
- 회귀분석과 같이 n개의 훈련자료 $(x_1, y_1), \ldots, (x_n, y_n)$를 관측했다고 하고, 이에 근거하여 분류기를 생성하고자 . 함
- 훈련자료에서 뿐 아니라 평가자료에서도 좋은 성능을 가지기를 기대함.

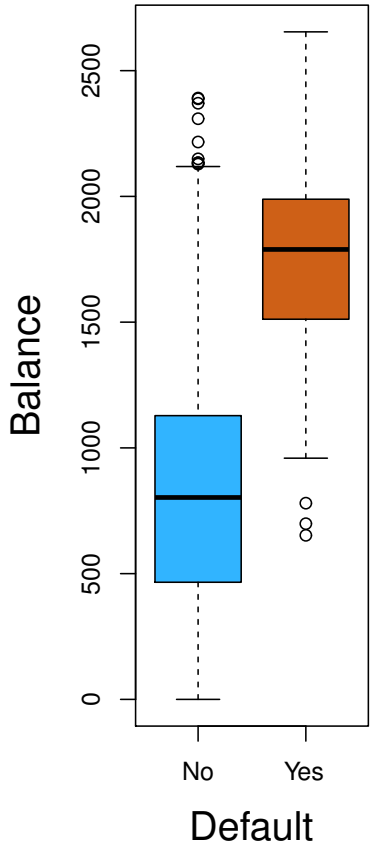
Why not linear regression?

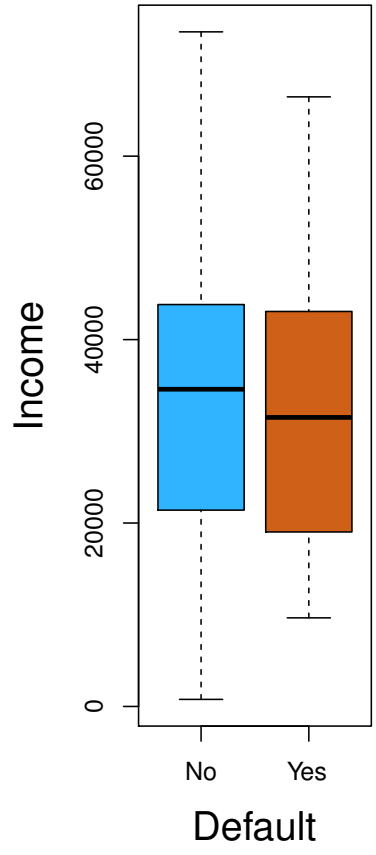
- 환자들을 증상에 따라 1,2,3의 범주에 할당함.
- 1,2,3은 범주의 구분을 위한 label일 뿐 숫자로써의 의미는 전혀 가지지 않음.
- 예를 들어 2번과 1번 사이의 차이 3번과 2번 사이의 차이가 같다는 보장이 전혀 없음.

로지스틱 회귀모형 (logistic regression)

- `Default` dataset: 금융정보를 바탕으로 채무불이행 여부를 예측하기 위한 모의실험 자료.







로지스틱 회귀모형 (logistic regression)

- 로지스틱 모형은 Y에 대한 직접적 모형화가 아닌 Y가 특정 범주에 포함될 확률을 모형화 함.

- 반응변수가 두 범주 중 하나로만 결정되는 경우로 binary response 임.

- 만약 채무불이행여부(default)를 지불잔액(balance)을 이용해서 모형화 하기 원한다면 추정 대상은

$$P(default = YES|balance) = p(balance) \in (0, 1)$$

- 임계치(threshold) : 특정 고객에 대해 $p(balnace) > 0.5$인 경우 default를 YES로 예측할 수 있다. 하지만, 보수적인 결정을 내리고자 하는 경우 예컨대 $p(balnace) > 0.1$인 고객에 대해 YES로 예측하는 것도 가능하다. 0.5,0.1과 같은 값을 임계치라 한다.

모형

- 관심이 되는 반응변수 값을 보통 1, 나머지를 0이라 하면, 주어진 예측변수 X에 대하여 추정대상은

$$Pr(Y = 1|X) = p(X) \in (0, 1)$$

- 위 확률을 X에 대하여 선형으로 모형화 한다면? 예측확률이 음수가 되는 경우 발생이 가능함.

$$p(X) = \beta_0 + \beta_1 X$$

- 이로부터 $p(X)$를 $(0, 1)$ 사이의 값으로 예측해주는 모형화 필요. 대표적인 것이 아래 로지스틱 함수임.

$$p(X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

모형

- 앞선 모형은 간단한 계산을 통해 다음과 같이 표현된다.

$$\frac{p(X)}{1-p(X)} = \exp(\beta_0 + \beta_1 X) \in (0, \infty)$$

- 좌측 항을 'odds'라 하며, 배팅을 하는 분야에서 확률 대신 많이 쓰이는 측도이다.

- 위 모형은 다시 다음과 같이 쓰여진다.

$$\log \frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X \in (-\infty, \infty)$$

- 좌측 항을 'logit'이라 한다.

선형모형 vs 로지스틱모형

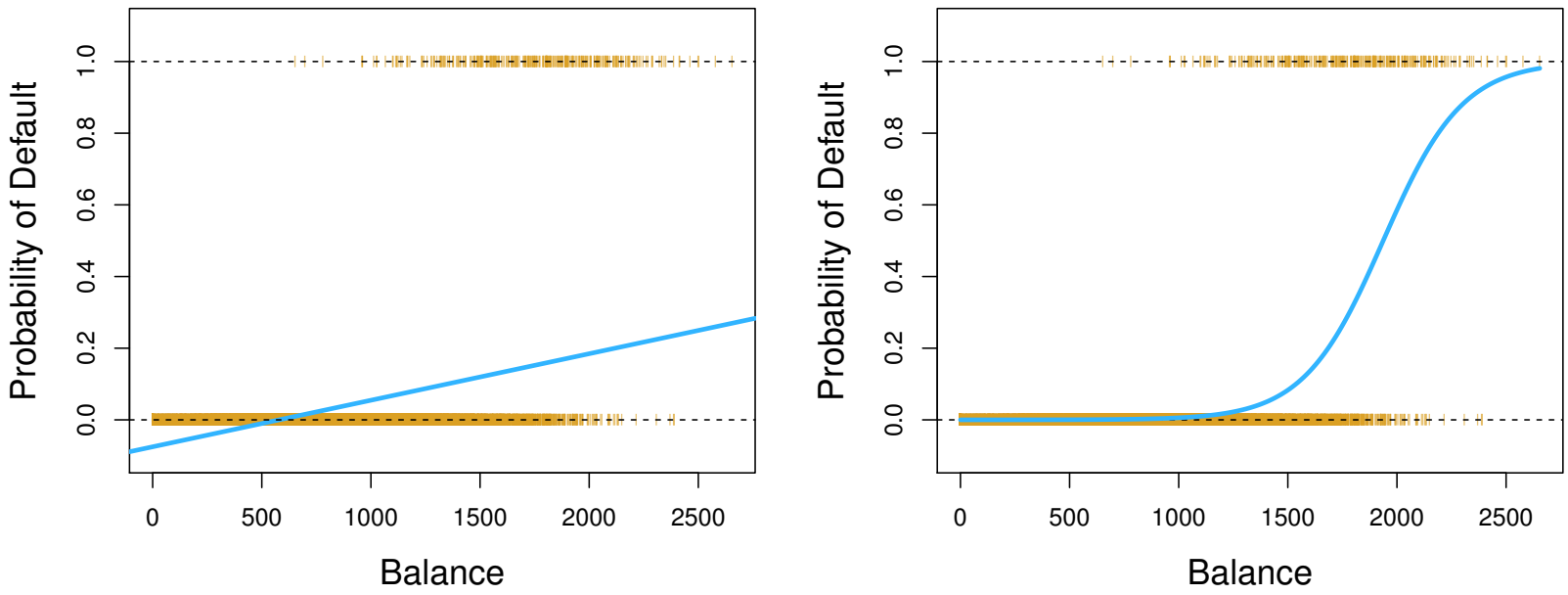


Figure 1: 선형모형 vs 로지스틱모형

추정

- 반응변수는 범주형변수이기에 최소제곱법을 사용하는 것은 절절하지 않을 수 있음.
- 반응변수의 조건부 분포를 이용한 최대우도추정법(maximum likelihood estimation)으로 보통 추정하게 됨.
- 반응변수가 베르누이(혹은 이항)분포임을 이용하여 가능도함수를 기술함.
- 보통 명시적 해가 존재하지 않아 (반복을 통한) 수치적인 접근을 통하여 추정치를 얻게 됨

예측

- 모수에 대한 추정이 이루어지면, 추정치와 주어진 예측변수를 이용하여 확률을 예측하게 됨.

- balnace 로 default 여부를 예측하는 모형에서 $\hat{\beta}_0 = -10.6513$, $\hat{\beta}_1 = 0.0055$으로 주어졌고, 어떤 고객의 지불잔액이 $1,000$이라면,

$$\hat{p}(X \mid X = 1000) = \frac{\exp(-10.6513 + 0.0055 \times 1000)}{1 + \exp(-10.6513 + 0.0055 \times 1000)} = 0.00576$$

- 한편 지불잔액이 1 증가하면, odds가 $\exp(0.0055)$배 증가하하게 된다.

다중로지스틱모형

- 여러 개의 예측변수를 사용하는 경우도 마찬가지로 확장하여 생각할 수 있다.
- 한 개의 예측변수를 사용할 때와 여러 개의 예측변수가 동시에 사용될 때 효과의 형태가 다르게 나타날 수도 있다(balance와 student confounding).

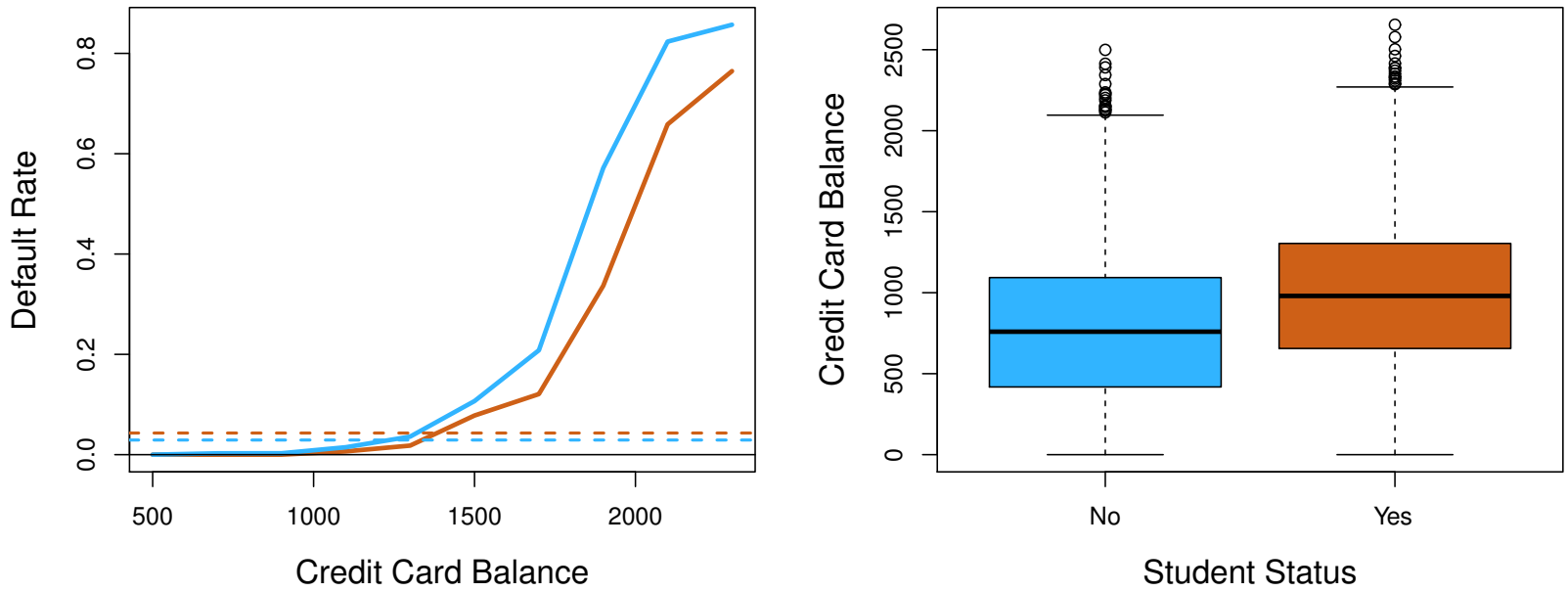


Figure 2: 다중로지스틱모형

다항반응변수에서의 로지스틱모형

- 반응변수가 셋 이상의 범주를 가질 때, 로지스틱 모형을 확장하여 고려할 수 있다.
- 그러나 다른 대안들의 존재로 인하여 이 경우 로지스틱 모형은 폭넓게 쓰이지는 않는다.
- 대표적인 대안 중 하나는 선형판별분석이다.

판별분석 (discriminant analysis)

- 반응변수와 예측변수들의 결합분포에 기반한 방법임.
- 범주들이 잘 분리되어 있을 때, 로지스틱 모형은 불안정하나 판별분석은 그렇지 않다.
- n이 작고 X의 분포가 정규분포에 가까울 때 판별분석의 성능이 로지스틱보다 좋다.
- 다범주 반응변수의 경우 로지스틱에 비해 더 간단하다.

베이즈 정리

- k번째 범주로부터 관측된 X의 분포를 $f_k(x) \equiv Pr(X = x|Y = k)$라 하고, π_k를 각 범주의 사전확률이라 하면, 베이즈 정리에 의하여 다음을 얻는다.

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

- 조건부 확률 $p_k(x)$의 추정치는 π_k와 $f_k(x)$에 대한 추정치로부터 얻을 수 있다.

- π_k의 추정보다는 $f_k(x)$에 대한 추정이 더 어렵다.

- 위에 대한 좋은 추정치를 얻을 수 있다면, Bayes 분류기에 가까운 분류기를 얻을 수 있을 것이다.

선형판별분석 (Linear discriminant analysis : LDA)

- $p = 1$, $f_k(x)$가 정규분포임을 가정함. 즉, $(N(\mu_k, \sigma_k^2))$

- 각 범주에 해당하는 분포의 분산은 동일하다 가정. 즉, $\sigma_1^2 = \cdots = \sigma_K^2 = \sigma^2$

- 이 경우, 조건부 확률 $p_k(x)$를 최대화 시키는 k를 찾는 것은

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

  를 최대로 하는 k를 찾는 것과 같음을 보일 수 있다.

- $K = 2$, $\pi_1 = \pi_2$인 경우, 분류를 위한 경계치는

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

  와 같이 간단히 된다.

예] $p = 1, K = 2.$

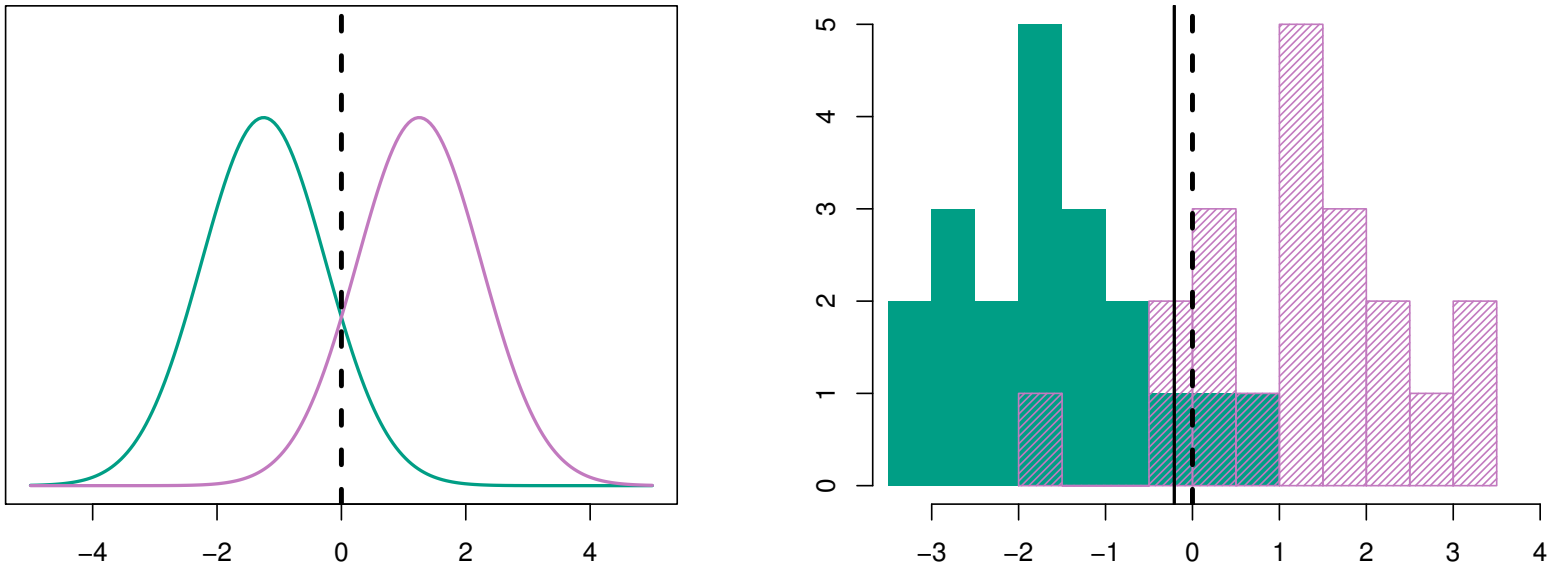


Figure 3: 선형판별분석 $K = 2$

베이즈 분류기로의 근사

- $p = 1$인 경우, LDA는 다음과 같이 베이즈 분류기를 근사하게 된다.

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

  $\hat{\delta}_k$를 최대화시키는 범주 k로 관측치를 할당.

- 'linear'라는 단어는 판별함수 $\hat{\delta}_k(x)$가 x의 선형함수로 표현된다는 사실로부터 유래함.

선형판별분석 ($p > 1$)

- 예측변수가 다변량인 경우 k번째 범주에서 예측변수가 다변량정규분포 ($N(\mu_k, \Sigma),\ \mu \in R^p, \Sigma : p \times p$)를 따른다는 가정으로부터 출발하게 됨.

- $p = 1$일때와 마찬가지로 모든 범주에서 공분산 행렬 Σ는 동일하다고 가정함.

- 이 경우 판별함수는 다음과 같이 주어지게 됨.

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

예] $p = 2, K = 3.$

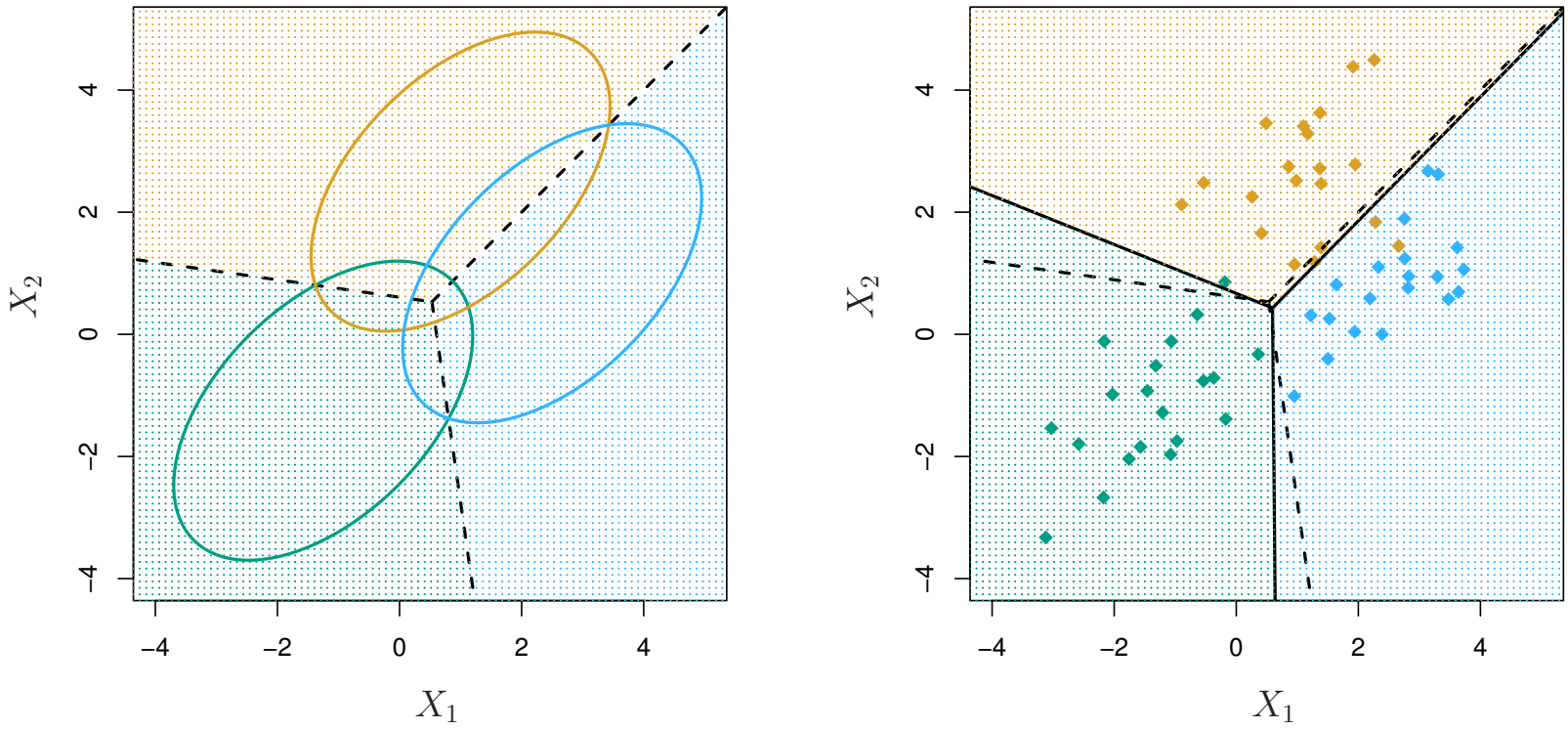


Figure 4: 선형판별분석 $K = 3$

예] ‘default’ dataset

- LDA 적용결과 10,000개의 훈련자료 중 2.75% 만이 오분류되었다. 즉, training error rate = 2.75%.

- 과연 분류가 잘 된 것인가?

  - 평가자료에서의 분류가 잘 되지 않을 수 있다.

  - 본 훈련자료의 3.33%만이 채무불이행(default=YES)이므로, 모든 관측치를 무조건 default=NO로 분류하는 분류기를 사용하더라도 오직 3.33%의 오류만이 발생할 것이다. 이는 2.75%와 큰 차이가 나지 않는다.

- 전체적인 오분류율만이 아닌 좀 더 특수한 상황에서의 평가지표들이 필요할 수 있음.

예] ‘default’ dataset

- Confusion matrix (혼동행렬)

		True default status		
		No	Yes	Total
Predicted default status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

Figure 5: Confusion matrix

- 실제 채무불이행자들에 대해서는 75%가 넘는 오분류가 나타남
- 손해를 끼칠 위험이 큰 고객들을 분별해 내야 하는 회사 입장에서는 위와 같은 결과는 바람직하지 않을 것임.
- 베이즈 분류기에서도 전체 오류율을 최소화시키려 하기 때문에 특정 범주에 대해서는 좋은 성능을 담보하지 않을 수 있음.

민감도(sensitivity), 특이도(specificity)

- 범주가 0과 1로 분류된다고 할 때 두 측도는 아래와 같이 정의됨.

  - 민감도: 1을 1로 분류하는 비율 / 특이도: 0을 0으로 분류하는 비율

- 'default' dataset의 경우, 민감도는 24.3% (81/333), 특이도는 99.8% (9644/9667)로 민감도가 매우 떨어진다.

- 임계치 조정 - 범주가 두 개인 경우 베이즈 분류기는 다음과 같은 조건을 만족하는 관측치에 한하여 default=YES로 분류함.

$$Pr(default = YES|X = x) > 0.5$$

  - 민감도를 향상시키기 위해서 임계치를 0.5에서 0.2로 바꾼다면

$$Pr(default = YES|X = x) > 0.2$$

  를 만족하는 관측치는 모두 default=YES로 분류함. 이 경우 민감도와 특이도가 변하게 됨.

임계치 조정

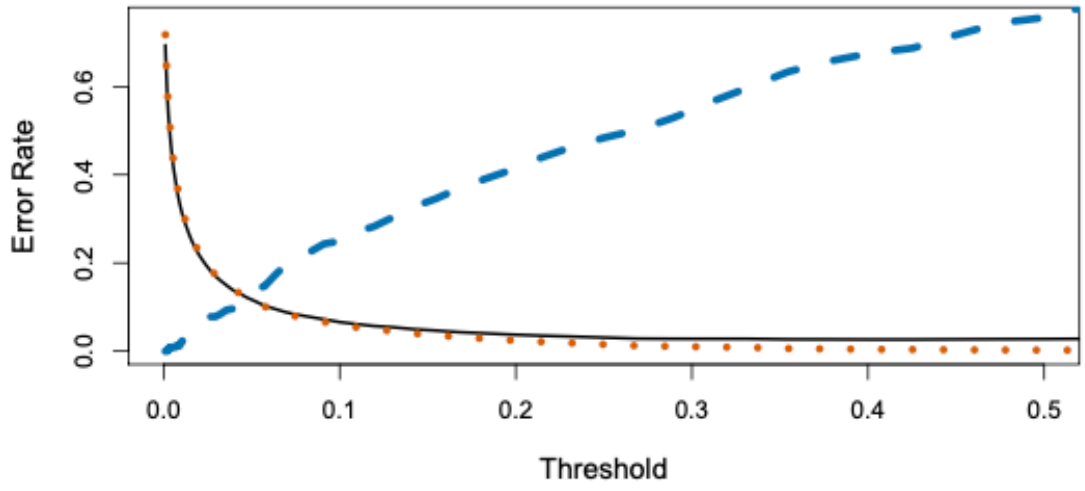


Figure 6: 임계치 조정 결과

- 민감도는 24.3%에서 58.6%로 증가함.
- 치러야 할 손실들.

  - 특이도는 99.8%에서 97.6%로 감소함.

  - 오분류율은 2.75%에서 3.73%로 증가함.
- 민감도의 극적인 향상을 위해서 약간의 손해는 감수할 수도 있음.
- 임계치의 설정에는 domain knowledge가 중요함.

ROC (receiver operating characteristics) curve

- 민감도와 특이도를 동시에 나타내어 분류기의 성능을 평가하는 대표적인 시각화 방법.
- 임계치를 변화시키면서 (1-특이도,민감도)를 2차원 좌표평면 상에 나타낸 곡선임.
- 이상적으로는 왼쪽 상단을 통과하는 것이 좋음. 이 경우 곡선 아래 면적이 1이 됨.
- 곡선 아래면적이 1에 가까울 수록 분류기의 성능이 좋은 것으로 볼 수 있음.
- 곡선 아래면적을 AUC (Area Under Curve)라 하며 ROC curve와 함께 분류기의 성능을 나타내는 지표 중 하나로 활용됨.

ROC curve 예시

- 민감도 (True positive rate), 1-특이도 (False positive rate)

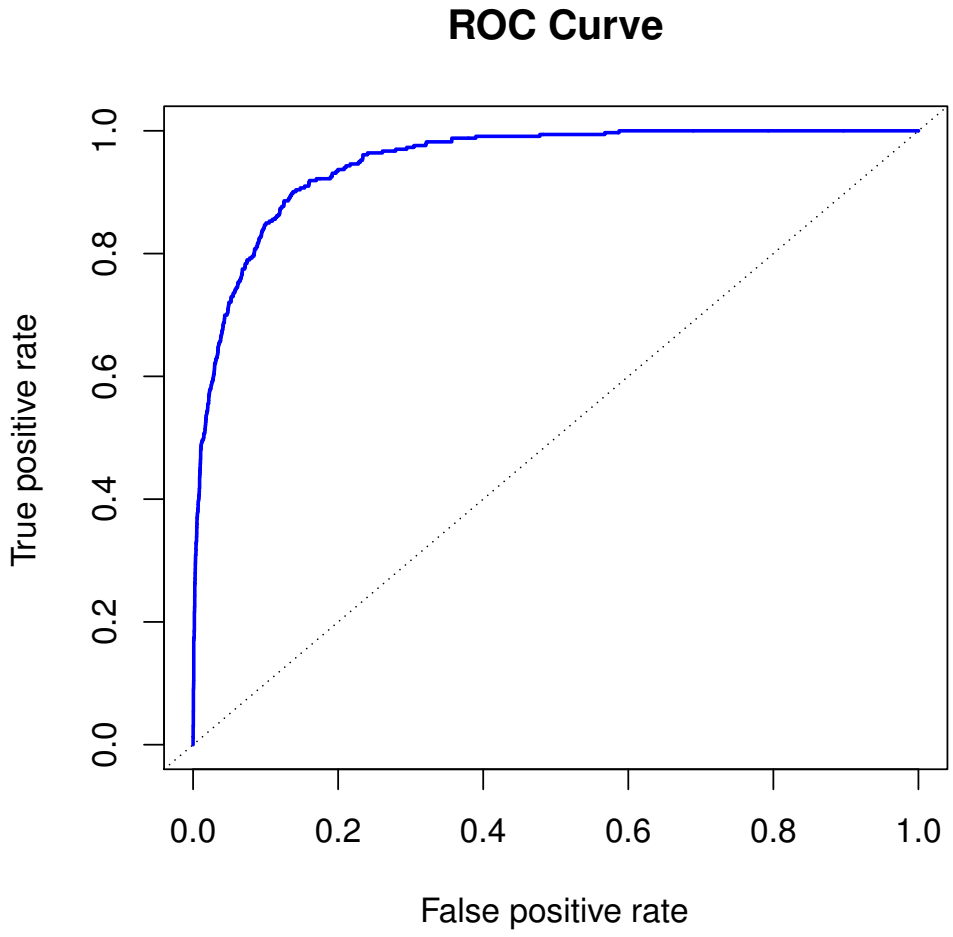


Figure 7: ROC curve

Quadratic discriminant analysis (QDA)

- 각 범주를 특정하는 정규분포의 분산에 이질성을 허용함. 즉, $X \sim N(\mu_k, \Sigma_k)$

- 이 경우 판별함수는 다음과 같은 형태가 됨이 알려져 있다.

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

- 판별함수가 x에 대한 2차식의 형태로 주어짐.

- LDA에 비해 유연한(복잡한) 모형임.

LDA vs QDA 예시

- 보라색 점선이 Bayes decision boundary 가 됨.

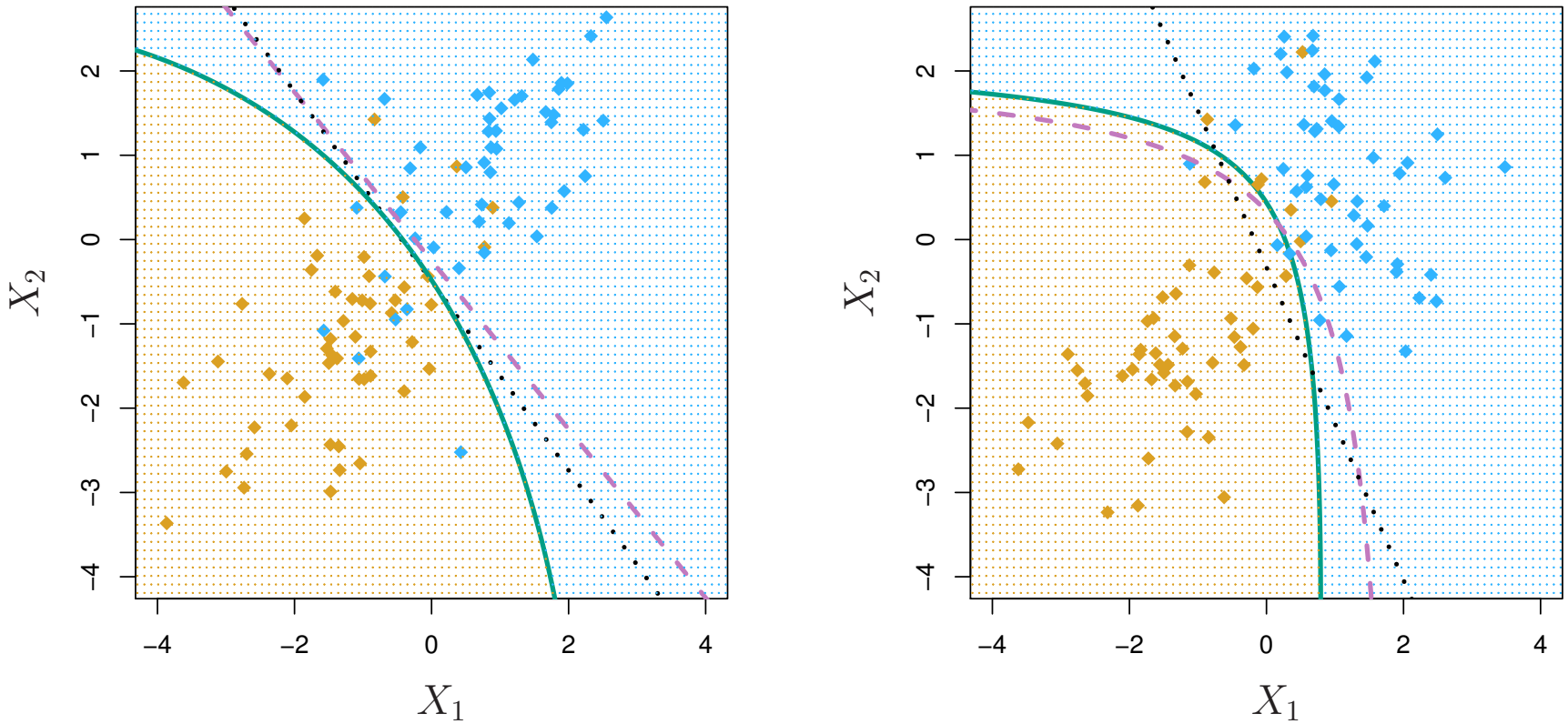


Figure 8: LDA vs QDA

Naive Bayes

- Bayes 규칙에 따라 주어진 예측변수들이 $x_1, \dots, x_p$이고 분류의 대상 변수가 $A \in \{0, 1\}$라고 할때,

$$Pr(A = a \mid x_1, \dots, x_p) = \frac{Pr(x_1, \dots, x_p \mid A = a) Pr(A = a)}{Pr(x_1, \dots, x_p)}$$

- 여기에서 $Pr(x_1, \dots, x_p \mid A = a)$는 p가 큰 경우 계산이 어려움, 이 경우 우리가 느슨하게 아래와 같은 (유사)확률을 사용하면 이것을 Naive Bayes 분류기라고 함.

$$Pr(x_1, \dots, x_p \mid A = a) = \prod_{i=1}^{p} Pr(x_i \mid A = a)$$

K-nearest neighbors

- 조건부 확률을 인접한 K개의 data points의 상대비율로 추정함.

$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

  $\mathcal{N}_0$: x_0와 가장 가까운 K개의 자료의 집합임. 그리고 거리를 계산할 때, y는 사용하지 않음에 유의.

- 위 확률을 최대로 하는 j로 관측치를 분류함.

- K의 선택이 분류기의 성능을 결정하는 데 매우 핵심적인 역할을 함.

분류기의 성능 비교

- 범주가 2개일 때, LDA와 로지스틱모형은 선형적인 decision boundary를 생성한다는 측면에서 유사함.
- 각 범주의 분포가 정규분포로 잘 근사되는지 여부에 의해서 두 방식의 성능이 엇갈릴 수 있음.
- KNN은 decision boundary에 어떠한 가정도 하지 않음. 즉, decision boundary가 비선형인 경우 위 두 방식에 비해 우월성을 보일 수 있음.
- QDA는 qudratic decision boundary를 설정한다는 면에서 KNN과 LDA 혹은 로지스틱의 중간쯤에 위치하는 방법으로 볼 수 있음.
- 모형의 복잡도(유연성)는 LDA $\approx$ 로지스틱 $<$ QDA $<$ KNN 임.

분류기의 성능 비교

- 예측변수, 반응변수 범주 각각 2개

- S1 : 각 범주내의 변수들은 서로 독립인 정규분포.

- S2 : 각 범주내 변수들의 상관계수가 -0.5. 다른 조건은 S1과 동일.

- S3 : 각 범주내의 변수들은 t분포.

- S4 : 각 범주내의 변수들은 각각 상관계수가 0.5, -0.5인 정규분포.

- S5 : 각 범주내의 변수들은 서로 독립인 정규분포. 반응변수가 두 변수의 이차 다항식을 이용한 결합으로 생성됨.

- S6 : 각 범주내의 변수들은 서로 독립인 정규분포. 반응변수가 두 변수의 복잡한 비선형결합으로 생성됨.

- 5개의 분류기

- KNN, $K = 1$ / KNN, K는 CV로 결정 / LDA / 로지스틱 모형 / QDA

분류기의 성능 비교

- 각 시나리오에서 100번씩 훈련자료를 생성하여 분류기를 얻어 test error 산출함 (훈련 데이터는 작은 크기이며 검증 데이터는 큰 크기임, 자세한 내용은 교재 참고).

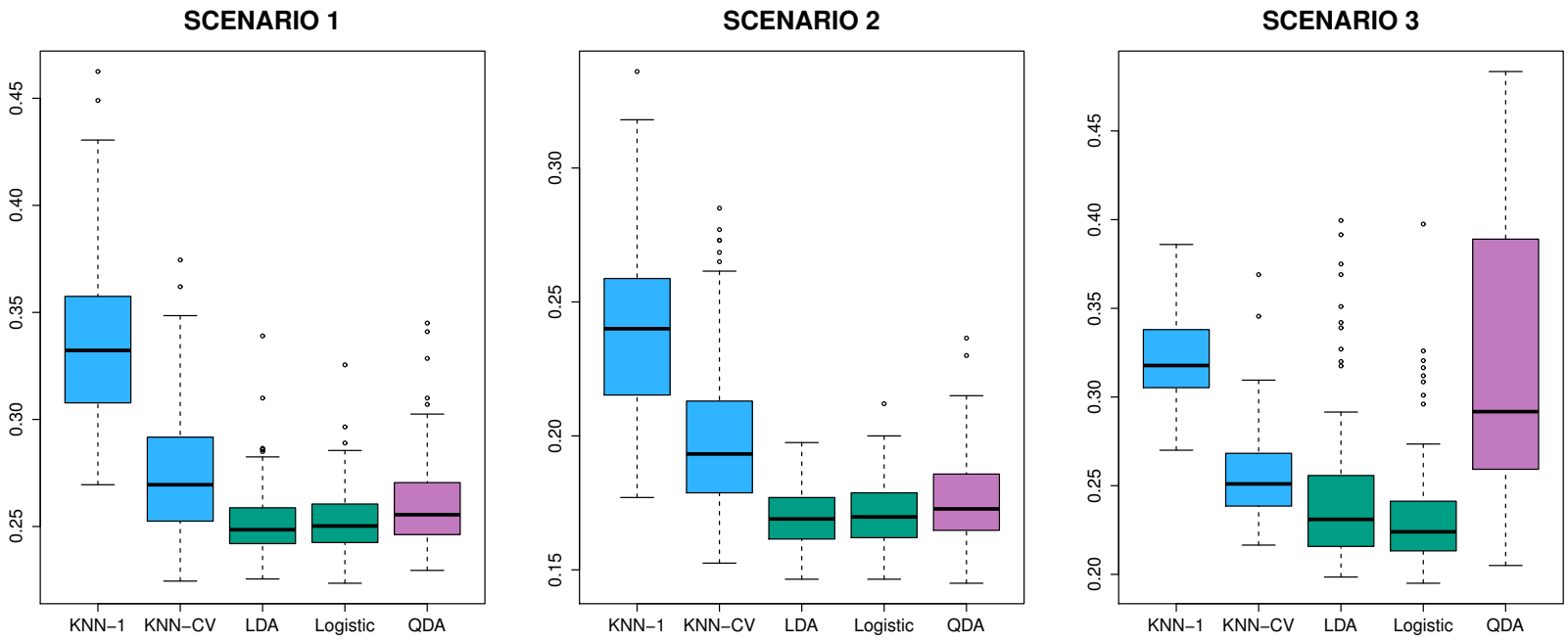


Figure 9: Boxplots of test error rates

분류기의 성능 비교

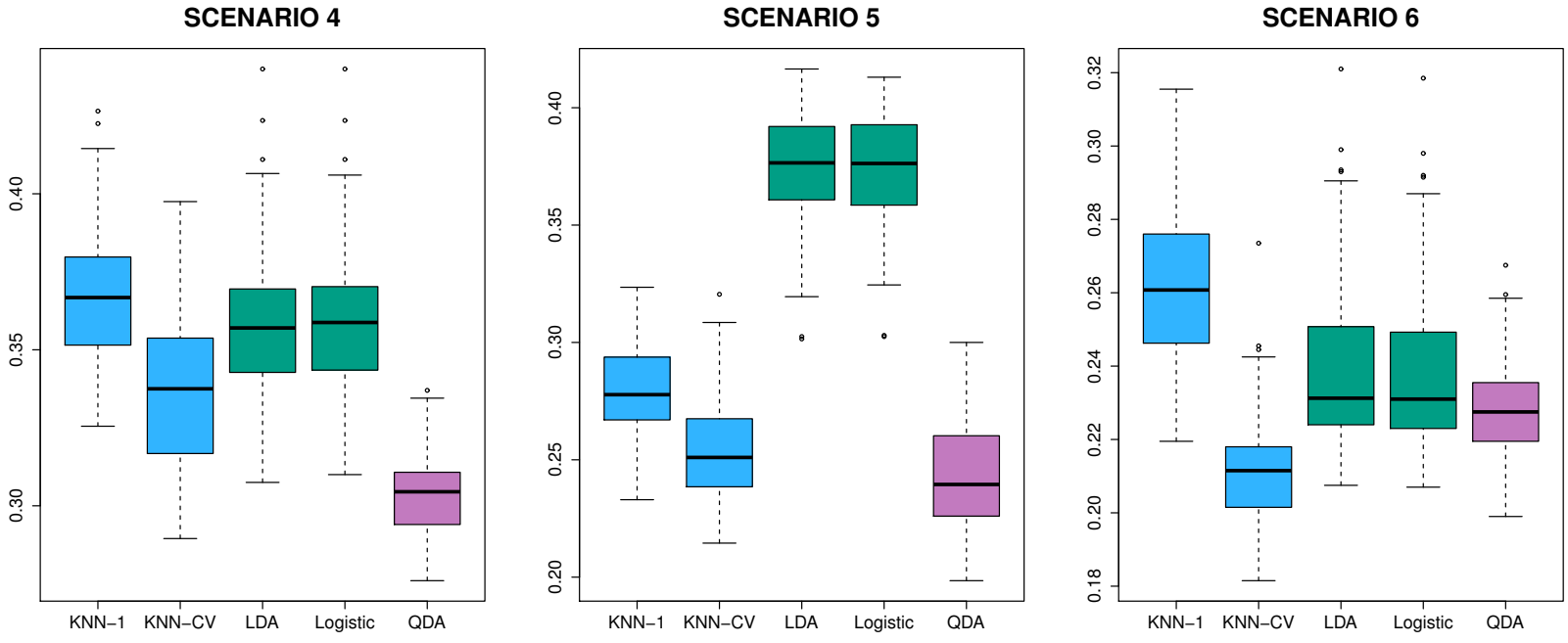


Figure 10: Boxplots of test error rates

분류기의 성능 비교

- 모든 상황(데이터 내지 분포)에서 우월한 분류기는 없다
- 비교적 간단한 상황에서는 LDA나 로지스틱 모형이 우수
- 복잡한 상황에서는 KNN이나 QDA가 우수
- KNN에서도 적절한 수의 K를 설정하는 것이 필요함