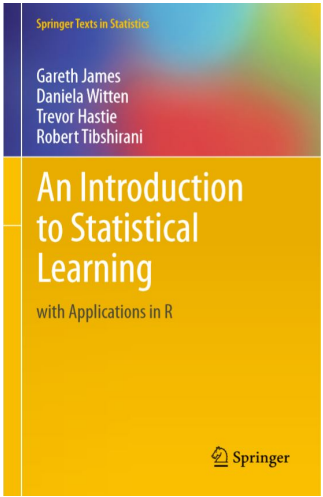
쉽게 배우는 머신러닝

Ch.4 Classification

훈 러닝 (Hun Learning)

January 6, 2020

TEXTBOOK



- **An Introduction to Statistical Learning : with Applications in R**
- 목차:
  1. Intro
  2. Statistical Learning
  3. Linear Regression
  4. Classification
  5. Resampling Methods
  6. Linear Model Selection and Regularization
  7. Moving Beyond Linearity
  8. Tree-based Methods
  9. Support Vector Machines
  10. Unsupervised Learning

CLASSIFICATION OVERVIEW

Regression vs Classification

Regression: Y가 수치형(Quantitative, Numerical)
나이, 몸무게, 소득, 주가 등
Classification: Y가 범주형(Qualitative, Categorical)
성별, 브랜드, 감염 여부 등

- **Classfication에서 y_0의 추정은 y_0를 "분류"하는 것과 같다.**
- 대표적인 Classifier:
  Bayes Classifier (Unattainable Ideal)
  K-Nearest Neighbors
  Logistic Regression
  Linear Discriminant Analysis
  etc.

WHY NOT LINEAR REGRESSION?

- $Y \in \{A, B, C\}$일 때, 다음 중 어떤 것을 쓰냐에 따라 결과가 달라짐.

$$Y = \begin{cases} 1 & \text{if A} \\ 2 & \text{if B} \\ 3 & \text{if C} \end{cases} \qquad Y = \begin{cases} 1 & \text{if C} \\ 2 & \text{if B} \\ 3 & \text{if A} \end{cases}$$

- 설사 $Y \in \{$순한 맛, 약간 매운 맛, 아주 매운 맛$\}$ 이런 거여도 이걸 숫자로 coding하는 통일된 방법이나 근거가 없음.
- **Y의 범주가 2개일 경우(Binary case)**는 $Y \in \{0, 1\}$으로 나태내어 예컨대 $\hat{y_0} > 0.5$ 이면 Y=1로 "분류"할 수 있다.
  이때 $\hat{Y} = X^T\beta$는 조건부 확률 $P(Y = 1|X)$으로 해석할 수도 있다!
  다만 문제는 OLS로 fitting할 시 $P(Y = 1|X)$이 0보다 작거나 1보다 클 수도 있다는 것..
  이 문제를 해결한 것이 **Logistic Regression**

LOGISTIC REGRESSION

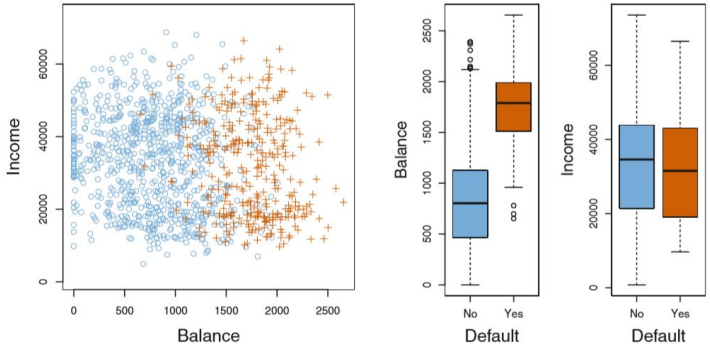


FIGURE 4.1. *The* `Default` *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* `balance` *as a function of* `default` *status.* Right: *Boxplots of* `income` *as a function of* `default` *status.*

LOGISTIC REGRESSION

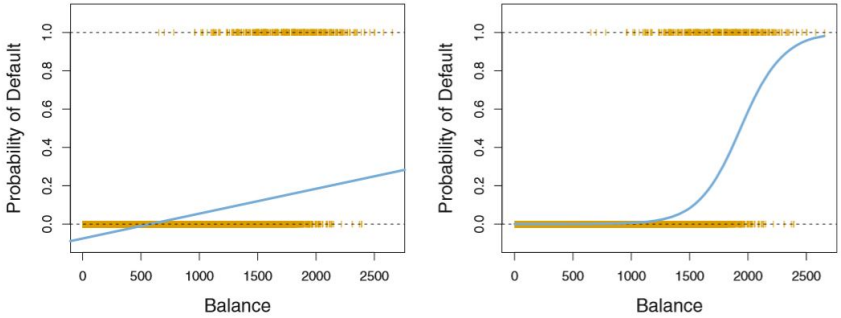


FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default` *(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

The Logistic Model

주어진 데이터 X로 $p(X) \equiv P(Y = 1|X)$를 fitting하자! OLS를 그대로 사용할 시

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

우변의 함수 형태를 다음과 같이 바꿔보자. 그러면 $p(X) \in [0, 1]$임을 알 수 있다.

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

이를 다시 쓰면 다음과 같다. **즉 로지스틱 회귀는 Y=1일 확률의 log-odd에 대한 선형 모델이다!**

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

$$\leftrightarrow \underbrace{log(\frac{p(X)}{1 - p(X)})}_{log\ odds, logit} = \beta_0 + \beta_1 X$$

The Logistic Model: β의 의미

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

위의 식에서 보면 X의 한 단위 변화는 odds에 e^{β_1}을 곱하는 것과 같다. 그러나 다음의 식에서 보듯이 p(X)와 X의 관계는 선형이 아니므로,

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

X의 한 단위 변화에 따른 p(X)의 변화는 1) 회귀계수의 크기, 방향 2) 현재 X의 수준에 다르다.

The Logistic Model: Maximun Likelihood Estimation of β

Logit Regression에서 가정한 개별 y의 sampling density는 다음과 같다.

$$\textit{Sampling Density of } y|X = \begin{cases} p(y=1|X) = p(X) = \frac{exp(\beta_0+\beta_1 X)}{1+exp(\beta_0+\beta_1 X)} \\ p(y=0|X) = 1-p(X) = 1 - \frac{exp(\beta_0+\beta_1 X)}{1+exp(\beta_0+\beta_1 X)} \end{cases}$$

N개의 (y_i, x_i) iid sample에 대한 Joint Sampling Density는 다음과 같다. 여기서 Y는 벡터.

$$\textit{Joint Sampling Density of } p(Y|X) = \prod_{i|y_i=1} p(X_i) \prod_{j|j=0} (1-p(X_j))$$

N개의 (y_i, x_i) iid sample이 주어졌을 때 위 식의 값은 Density의 모수인 회귀계수 β의 값에 따라 달라진다. 이렇게 해석하면 위의 식은 데이터가 주어졌을 때 β의 식, Joint Likelihood으로 볼 수 있다.
이 식을 최대화하는 β를 추정치로 하는 방법이 MLE

$$\textit{Joint Likelihood function } L(\beta_0, \beta_1|Y, X) = \prod_{i|y_i=1} p(X_i) \prod_{j|j=0} (1-p(X_j))$$

The Logistic Model: Maximun Likelihood Estimation of β

상반된 결과? No! 학생일수록 "대출잔고(balance)가 높기 때문에" 부도율이 높은 것. 똑같은 잔고에서는 오히려 학생의 부도율이 낮다.

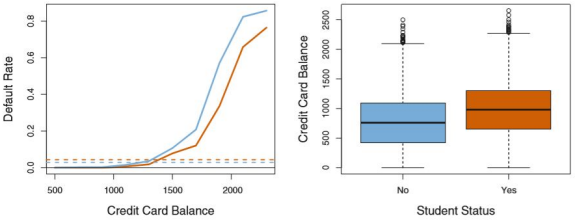


FIGURE 4.3. *Confounding in the* `Default` *data.* Left: *Default rates are shown for students (orange) and non-students (blue). The solid lines display default rate as a function of* `balance`*, while the horizontal broken lines display the overall default rates.* Right: *Boxplots of* `balance` *for students (orange) and non-students (blue) are shown.*

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

TABLE 4.1. *For the* `Default` *data, estimated coefficients of the logistic regression model that predicts the probability of* `default` *using* `balance`*. A one-unit increase in* `balance` *is associated with an increase in the log odds of* `default` *by* 0.0055 *units.*

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

TABLE 4.2. *For the* `Default` *data, estimated coefficients of the logistic regression model that predicts the probability of* `default` *using student status. Student status is encoded as a dummy variable, with a value of* 1 *for a student and a value of* 0 *for a non-student, and represented by the variable* `student[Yes]` *in the table.*

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

TABLE 4.3. *For the* `Default` *data, estimated coefficients of the logistic regression model that predicts the probability of* `default` *using* `balance`*,* `income`*, and student status. Student status is encoded as a dummy variable* `student[Yes]`*, with a value of* 1 *for a student and a value of* 0 *for a non-student. In fitting this model,* `income` *was measured in thousands of dollars.*

LOGISTIC REGRESSION

The Logistic Model: Response with > 2 Classes

- 예컨대 $Y \in \{A, B, C\}$이면 Sampling Density (Likelihood)는 다음과 같다. ($\beta \in \mathbb{R}^{p \times 1}, X \in \mathbb{R}^{n \times 1}$)

$$\textit{Sampling Density of } y|X = \begin{cases} p(y = A|X) = \frac{exp(X^T \beta_A)}{1+exp(X^T \beta_A)} \\ p(y = B|X) = \frac{exp(X^T \beta_B)}{1+exp(X^T \beta_B)} \\ p(y = C|X) = 1 - p(y = A|X) - p(y = B|X) \end{cases}$$

$$\textit{Likelihood } L(\beta_A, \beta_B) = \prod P(A|X) \prod P(B|X) \prod P(C|X)$$

- 이 경우 범주가 추가될 때마다 추정해야하는 모수의 수가 p개 만큼 추가된다.
- 그러나 범주가 2개보다 많은 경우는 대부분 다른 방법을 쓴다. ex) **Linear Discriminant Analysis**

LDA: Intuition

- Classification의 Parameteric Methods는 근본은 다 똑같다. 결국은 아래의 Bayes Classifier 추정하는거임. 이걸 다 아는 하나님은 요거 보고 분류하는거. (unattainable gold standard)

$$Bayes\ Classifier:\ p_k(X) = P(Y = k|X = x)$$

- 결국은 $k \in [K]$ 중에서 Bayes Classifier가 가장 큰 애를 고르는 거니까 위 식의 Monotone trasnformation을 추정해도 됨.

$$Discriminant:\ \delta_k(X) = f(P(Y = k|X = x))$$

- Logit Regression도 결국 log-odd에 대한 선형 모델인 것을 기억하라.

$$\underbrace{log(\frac{p(X)}{1-p(X)})}_{log\ odds, logit} = \beta_0 + \beta_1 X$$

- **LDA도 Discriminant의 선형함수. 그러나 추정 방식이 Logit은 MLE라면 LDA는 베이즈 룰 사용!**

판서로 때우는 Bayes Rule 초간단 복습

LDA: Estimating Bayes Classifier

$$\begin{aligned} \textit{Bayes Classifier } P(Y=k|X=x) &= \frac{P(Y=k \textit{ and } X=x)}{P(X=x)} \\ &= \frac{P(Y=k)P(X=x|Y=k)}{\sum_k P(Y=k)P(X=x|Y=k)} \\ &= \frac{\pi_k f_k(X)}{\sum_{l=1}^{K} \pi_k f_l(X)} \end{aligned}$$

- π_k: Prior Prob of each class k, 예컨대 전체 N개 관측치 중 class k인 관측치 개수의 비율로 쉽게 추정할 수 있다.
- $p_k(X) = P(k|X)$: Predictor X가 주어졌을 때 관측치가 class k에 귀속될 확률, 우리의 target
- $f_k(X) = P(X|k)$**: Likelihood of X in class k. 이것에 대한 가정에 따라 LDA와 QDA**
  "If we can find a way to estimate $P(X|k)$, then we can develop a classifier that approximate the Bayes Classifier."

LINEAR DISCRIMINANT ANALYSIS

LDA: p = 1 One Predictor

- LDA의 요체는 $X|y = k \sim$ *Normal Distribution*

$$f_k(X) = \frac{1}{\sqrt{2\pi\sigma_k^2}} exp(-\frac{(x-\mu_k)^2}{2\sigma_k^2})$$

LDA의 가정

- $\sigma_i^2 = \sigma^2$ for $\forall i \in [K]$, Shared Variance: 각 클래스마다 X의 평균은 다르겠지만 모든 클래스에서의 X의 분산은 동일하다!

$$f_k(X) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu_k)^2}{2\sigma^2})$$

LINEAR DISCRIMINANT ANALYSIS

LDA: p = 1 One Predictor

$$f_k(X) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu_k)^2}{2\sigma^2})$$

위 식을 Bayes Classifier $P(Y = k|X)$에 대입하면

$$P(K|X) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu_k)^2}{2\sigma^2})}{\sum_l \pi_l \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu_l)^2}{2\sigma^2})}$$

결국은 주어진 x_0값에서 어느 class의 $P(Y = k|X)$가 가장 큰지만 알면 되니까 아래처럼 간단히 단조변환하면 Discriminant. x에 대해 Linear하다!

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

LINEAR DISCRIMINANT ANALYSIS

LDA: p = 1 One Predictor

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

문제는 위 식에서 모수 $(\mu_1, ...\mu_k)$, σ^2, $(\pi_1, ..., \pi_k)$를 어떻게 추정할 것인가이다.

- $\hat{\mu}_k$는 class k인 관측치 x의 평균으로
- $\hat{\sigma^2}$는 각 클래스 내에서의 편차 제곱을 모두 더해 N-k로 나눈 것으로
- $\hat{\pi_k}$는 전체 관측치 중 class k의 비율로 추정

$$Sample\ Linear\ Discriminant : \hat{\delta_k}(x) = x\frac{\hat{\mu_k}}{\hat{\sigma^2}} - \frac{\hat{\mu_k}^2}{2\hat{\sigma}^2} + log(\hat{\pi_k})$$

여기서 만일 모든 클래스 별 관측치의 개수가 동일하면 $log(\hat{\pi_k})$는 의미가 없음.
관측치 x_0에 대해 각각의 class k에 대한 $\hat{\delta_k}(x)$를 모두 계산해 그 값이 가장 큰 class k로 관측치를 분류!

LINEAR DISCRIMINANT ANALYSIS

LDA: $p > 1$ **Multiple Predictor**
$p > 1$의 경우 $X|y = k \sim$ *Class specific* $MVN(\mathcal{M}_k, \Sigma)$

$$f_k(X) = \frac{1}{(2\pi^2)^{p/2}|\Sigma|^2} exp(-frac12(X - \mathcal{M}_k)^T \Sigma^{-1}(X - \mathcal{M}_k)$$

$$\delta_k(X) = x^T \Sigma^{-1} \mathcal{M}_k - \frac{1}{2} \mathcal{M}_k^T \Sigma^{-1} \mathcal{M} + log(\pi_k)$$

- $\hat{\mathcal{M}}_k$는 class k에서 각 x_i 관측치의 변수별 평균으로
- $\hat{\Sigma^2}_{i,j}$는 각 클래스에서의 i변수와 j변수의 편차 곱의 합을 모두 더해 N-k로 나눈 것으로
- $\hat{\pi}_k$는 전체 관측치 중 class k의 비율로 추정

Bayesian Decision Boundary: set of $X = (x_1, x_2, ..., x_p)$ where $\delta_k(X) = \delta_l(X)$ for $k \neq l$

LDA: p = 1 One Predictor

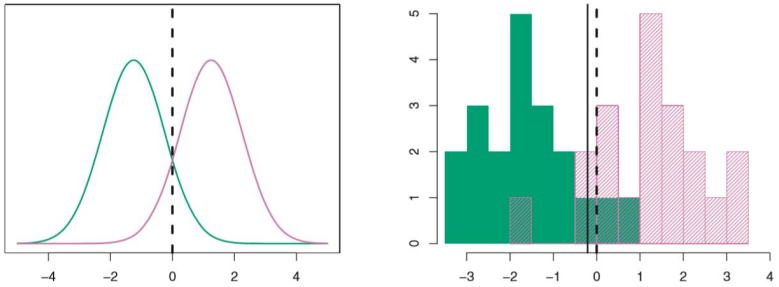


FIGURE 4.4. Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.* Right: *20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.*

LDA: $p > 1$ Multiple Predictor

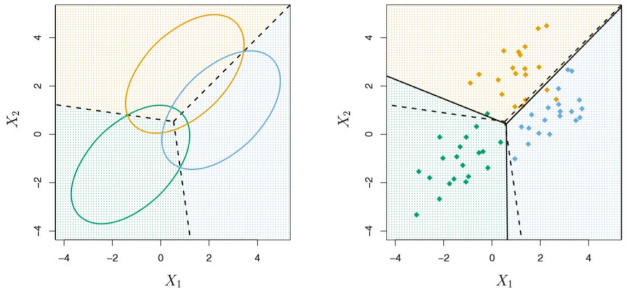


FIGURE 4.6. *An example with three classes. The observations from each class are drawn from a multivariate Gaussian distribution with $p = 2$, with a class-specific mean vector and a common covariance matrix.* Left: *Ellipses that contain 95 % of the probability for each of the three classes are shown. The dashed lines are the Bayes decision boundaries.* Right: 20 *observations were generated from each class, and the corresponding LDA decision boundaries are indicated using solid black lines. The Bayes decision boundaries are once again shown as dashed lines.*

LINEAR DISCRIMINANT ANALYSIS

LDA: Training Error Rate

- LDA에서 Training ER은 대부분 Test ER보다 낮다. Training data로 Bayes Classifier를 추정하기 때문. 때문에 만일 모수의 개수와 관측치의 비율 p/N이 크면 Overfitting 문제가 발생할 수 있다.
- Null Classifier: Y = 0 or 1 에서 그냥 모든 관측치를 Y=1로 분류하는 경우, 이 때의 Error Rate은 전체 관측치에서 0의 비율이다. 모델의 Training ER이 이거보다 낮으면 안 쓰느니만 못한 거.

LDA: Confusion Matrix

- 똑같은 에러라도 중요도가 다를 수 있다. 예컨대 신용카드 회사 입장에서는 파산하지 않을 사람을 파산으로 예측하는 것보다, 파산할 사람을 파산하지 않을 것으로 예측하는 것이 더 큰 문제. **내가 쓴 Classifer의 오류가 어떤 종류인지를 나타내는 것이 Confusion Matrix**

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

LDA: Threshold and ROC, AUC

- 이전 장의 Confusion Matrix를 보면 "파산예측O|파산No"은 23/9667로 낮지만 "파산예측X|파산Yes"은 252/333으로 거의 75%가 넘는다. **실제 파산한 333명의 고객 중에서 252명은 파산하지 않을 것으로 예측됐다는 거다.** 이런 classifier 가지고 갖다간 카드사한테 뺨 맞는다. 왜 이런 결과가 나타난 걸까?
- 파산하지 않을 사람한테 파산할거에요 라고 예측하는 것보다 **파산할 사람한테 파산하지 않을 거에요라고 예측하는 것이 크나큰 문제.** 그러나 Bayes Classifier에서 계산하는 Error Rate은 두 종류의 오류를 동등하게 취급한다. 때문에 이전의 경우처럼 실제로 파산하지 않은 사람의 비율이 압도적으로 많을 경우, Overall Error Rate를 최소화하게 되면 경우 파산하지 않을 사람에게 파산할 거에요라고 예측하는 비율을 줄이게 된다. 똑같은 비율이어도 분모가 크면 분자도 크니까.
- **이 경우 Classifier의 Threshold를 수정하여 class-specific performance을 향상시킬 수 있다!** Binary Case에서 임계치는 0.5이다. 이걸 낮추면 "X가 주어진 고객의 파산 확률이 (파산하지 않을 확률보다 낮아도) 내가 설정한 임계치보다 높으면 파산으로 분류"하는 것.

$$P(Default = Yes|X = x) > Threshold$$

LDA: Threshold and ROC, AUC

많이 헷갈리니까 한 눈에 정리하면 다음과 같다.

$$\text{실제 파산X} \begin{cases} \text{파산 예측O (False Positive, Type I Error)} \\ \text{파산 예측X (Specificity)} \end{cases}$$

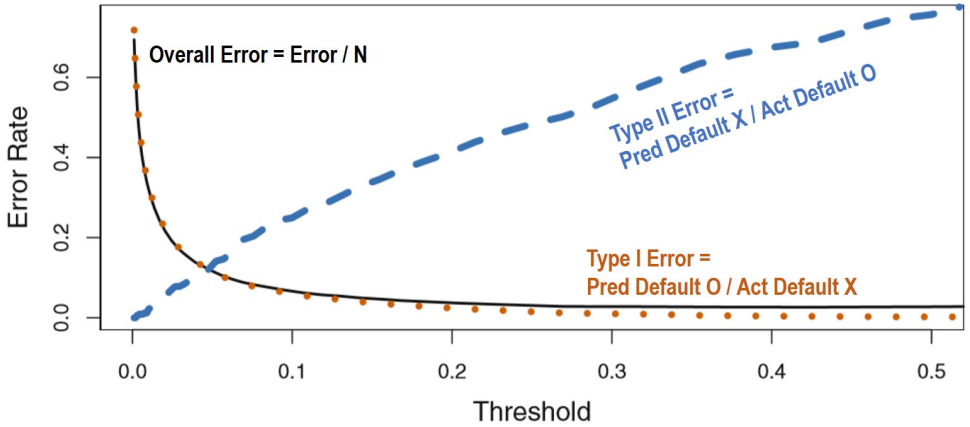
$$\text{실제 파산O} \begin{cases} \text{파산 예측O (True Positive, Power, Sensitivity)} \\ \text{파산 예측X (Type II Error)} \end{cases}$$

이걸 우리가 익숙한 "H_0: 파산X vs H_1 : 파산O" 으로 일반화하면

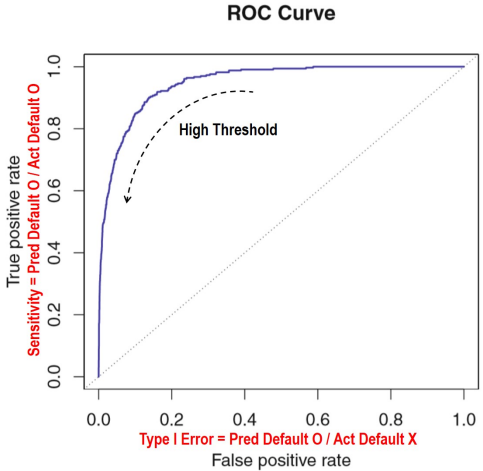
$$H_0 \text{ true} \begin{cases} \text{Reject } H_0 \text{ (False Positive, Type I Error)} \\ \text{Accept } H_0 \text{ (Specificity)} \end{cases}$$

$$H_1 \text{ true} \begin{cases} \text{Reject } H_0 \text{ (True Positive, Power, Sensitivity)} \\ \text{Accept } H_0 \text{ (Type II Error)} \end{cases}$$

LDA: Threshold and ROC, AUC



LDA: Threshold and ROC, AUC



QUADRATIC DISCRIMINANT ANALYSIS

Quadratic Discriminant: Different Var per class

- LDA에서의 Class-specific $f_k(X)$는 각 class 별로 평균은 다르나 분산은 같다. 만일 각 class 별로 다른 분산을 가정하면 QDA. 이 경우 Discriminant가 X에 대한 Quadratic 함수가 되기 때문.

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

- QDA를 사용할 경우 Σ에서 추정할 모수의 개수가 $K\frac{p(p+1)}{2}$이다. 때문에 QDA는 LDA에 비해 덜 유연하며, training 데이터가 적은 경우 모델의 분산이 굉장히 클 수 있다.
- 반면에 실제로 Discriminant이 비선형이면 QDA를 사용함으로써 모델의 Bias를 줄일 수 있다. **(Var-Bias Tradeoff)**

QUADRATIC DISCRIMINANT ANALYSIS

Quadratic Discriminant: V-B Tradeoff

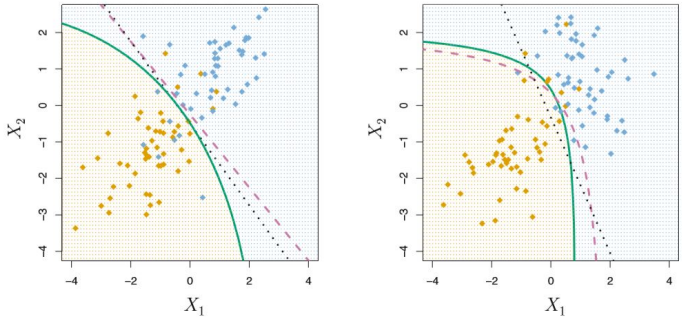


FIGURE 4.9. Left: *The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with* $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. *The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA.* Right: *Details are as given in the left-hand panel, except that* $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. *Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.*

A COMPARISON OF CLASSIFICATION METHODS

Classification 방법 별 비교

- **Logit vs LDA:** 둘 다 Bayes Classifier의 선형 모수 추정이라는 점에서 유연하나 가정이 다르다. LDA는 각 class에서 X의 분포가 정규 분포임을 가정한다. 예컨대 파산한 자의 소득과 잔고가 이변수 정규분포를 따르는 것을 가정하는 것. 이 가정이 맞을 경우 LDA의 성능이 더 좋으나 아니면 이러한 가정이 없는 Logit이 더 좋을 수 있다.
- **KNN:** KNN은 Bayes Classifier의 비모수 추정이다. 이때 주변에 몇 개의 점을 볼 것인가에 따라 굉장히 유연할 수도 경직적일 수도 있다.
- **QDA:** QDA는 모수 개수를 추가함으로써 LDA. Logit과 KNN의 사이에 있다고 할 수 있다. 실제 Decision Boundary가 비선형일 때는 적은 수의 표본으로도 QDA가 좋은 성능을 낼 수 있다.

결론은 각 사례 별로 길고 짧은 것은 대봐야 안다는 것!

Classification Case Studies: Linear Decision Boundaries

Scenario 1: There were 20 training observations in each of two classes. The observations within each class were uncorrelated random normal variables with a different mean in each class. The left-hand panel of Figure 4.10 shows that LDA performed well in this setting, as one would expect since this is the model assumed by LDA. KNN performed poorly because it paid a price in terms of variance that was not offset by a reduction in bias. QDA also performed worse than LDA, since it fit a more flexible classifier than necessary. Since logistic regression assumes a linear decision boundary, its results were only slightly inferior to those of LDA.

Scenario 2: Details are as in Scenario 1, except that within each class, the two predictors had a correlation of -0.5. The center panel of Figure 4.10 indicates little change in the relative performances of the methods as compared to the previous scenario.

Classification Case Studies: Linear Decision Boundaries

Scenario 3: We generated X_1 and X_2 from the *t-distribution*, with 50 observations per class. The t-distribution has a similar shape to the normal distribution, but it has a tendency to yield more extreme points—that is, more points that are far from the mean. In this setting, the decision boundary was still linear, and so fit into the logistic regression framework. The set-up violated the assumptions of LDA, since the observations were not drawn from a normal distribution. The right-hand panel of Figure 4.10 shows that logistic regression outperformed LDA, though both methods were superior to the other approaches. In particular, the QDA results deteriorated considerably as a consequence of non-normality.

Classification Case Studies: Linear Decision Boundaries

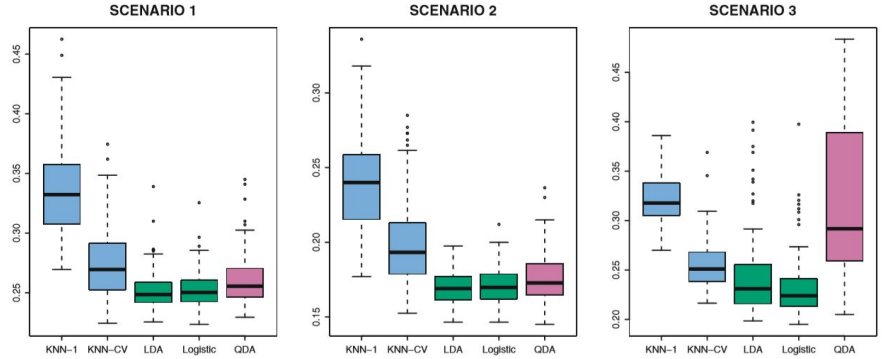


FIGURE 4.10. *Boxplots of the test error rates for each of the linear scenarios described in the main text.*

Classification Case Studies: Non-Linear Decision Boundaries

Scenario 4: The data were generated from a normal distribution, with a correlation of 0.5 between the predictors in the first class, and correlation of -0.5 between the predictors in the second class. This setup corresponded to the QDA assumption, and resulted in quadratic decision boundaries. The left-hand panel of Figure 4.11 shows that QDA outperformed all of the other approaches.

Scenario 5: Within each class, the observations were generated from a normal distribution with uncorrelated predictors. However, the responses were sampled from the logistic function using X_1^2, X_2^2, and $X_1 \times X_2$ as predictors. Consequently, there is a quadratic decision boundary. The center panel of Figure 4.11 indicates that QDA once again performed best, followed closely by KNN-CV. The linear methods had poor performance.

Classification Case Studies: Non-Linear Decision Boundaries

Scenario 6: Details are as in the previous scenario, but the responses were sampled from a more complicated non-linear function. As a result, even the quadratic decision boundaries of QDA could not adequately model the data. The right-hand panel of Figure 4.11 shows that QDA gave slightly better results than the linear methods, while the much more flexible KNN-CV method gave the best results. But KNN with $K = 1$ gave the worst results out of all methods. This highlights the fact that even when the data exhibits a complex non-linear relationship, a non-parametric method such as KNN can still give poor results if the level of smoothness is not chosen correctly.

Classification Case Studies: Non-Linear Decision Boundaries

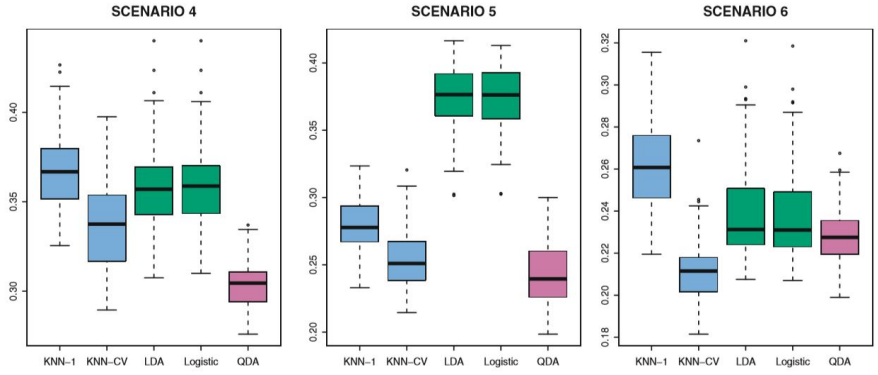


FIGURE 4.11. *Boxplots of the test error rates for each of the non-linear scenarios described in the main text.*