Yonsei University, Department of Applied Statistics

Chapter 2 Statistical Learning

YBIGTA Science Team 신보현

January 20, 2019

Contents

2.1 What is Statistical Learning? **3**

2.1.1 Why Estimate f? 3

2.1.3 How Do We Estimate f ? 4

2.1.3 the trade-off between prediction accuracy and model interpretability 5

2.1.4 Supervised learning vs UnSupervised learning 6

2.1.5 regression vs classification problems 6

2.2 assessing model accuracy **6**

2.2.1 measuring the quality of fit 6

2.2.2 the bias-variance trade-off 8

2.2.3 the classification setting 9

2. Statistical Learning

2.1 What is Statistical Learning?

간략하게 용어 정리

X: input variable, predictors, independent variable, features, just variable, 독립변수

Y: output variable : response, dependent variable, 종속변수

relationship between Y and X which can be written as $Y = f(X) + \epsilon$

여기서 $f(X)$는 some fixed but unknown function이라고 보면 된다. 즉, 우리가 관심있는 변수, 독립 변수와 종속 변수간의 알려지지 않은 관계를 $f(X)$라고 보면 된다.

ϵ: random error term which is independent of x and has mean zero. 우리가 아무리 열심히 독립 변수와 종속 변수간의 관계를 설정해도 종속 변수에 관해서 우리가 생각하지 못한 어떤 변수, 또는 관계 없는 변수가 있을 수 있으며 이것을 모두 ϵ에 넣어 버린다.

핵심은 statistical learning은 f를 estimating하기 위한 여러가지 접근 방법을 말한다. 여기서는 f를 측정하기 위한 여러 개념을 배울 것이다.

2.1.1 Why Estimate f?

크게는 prediction과 inference를 위해 측정한다.

* prediction 많은 상황에서는 X는 사용할 수 있지만 output Y는 쉽게 못 얻는다. 그렇기 때문에 우리는 $\widehat{Y} = \widehat{f(x)}$ (error의 mean이 0이기 때문에)을 이용해서 Y를 예측한다. 여기서 $\widehat{f(x)}$은 black box라고 여겨지고 $\widehat{f(x)}$의 정확한 형태를 얻는 것에 관심있는 것이 아니라 Y를 정확히 예측하는지가 주요 관심사이다. 이렇게 Y를 예측하기 위해서 사용하는 $\widehat{Y}$의 정확도는 reducible error와 irreducible error 두 가지에 의존한다. 일반적으로 $\widehat{f(x)}$은 f을 측정하기 위한 완벽한 측정도구가 아니므로 어느정도의 error가 발생한다. 이러한 error를 reducible error라고 하는데 이는 적절한 통계적 learning 을 통해 줄일 수 있다. 이와 반면에, 아무리 완벽한 측정을 했음에도 여전히 error가 있을 수 있다. 왜냐하면 Y는 ϵ의 함수이므로 정의상 X를 이용해서 예측할 수 없기 때문이다. 즉, 주어진 데이터로 열심히 f를 estimate 한다고 해도 random하게 발생하는 error가 존재하고 이를 irreducible error라고 한다.
그렇다면 왜 irreducible error가 0보다 클까? ϵ는 우리가 측정할 수 없지만 Y를 측정하기에 유용한 변수들을 담고 있다. 예를 들어 약의 부작용을 측정하려고 할 때 그날 환자의 기분 등 다른 요인이 중요한 역할을 할 수 있다. 따라서 항상 일어난다고 할 수 있다. 또는 독립변수와

종속변수간의 관계를 알아보고자 할 때, 주어진 데이터로는 100% 완벽히 추론할 수 없으므로 항상 일정한 오차가 일어난다고 생각할 수 있다.
아래는 측정 오차를 reducible error와 irreducible error로 분해한 결과이다. $\hat{f}, X$가 fixed 되었다고 가정하자.

$$
\begin{aligned}
E[(Y-\widehat{Y})^2] &= E[(f(X)+\epsilon-\widehat{f(X)})^2] \\
&= E[(f(X)-\widehat{f(X)})^2+\epsilon^2+(f(X)-\widehat{f(X)})\epsilon] && (1) \\
&= [f(X)-\widehat{f(X)}]^2+E[\epsilon^2] && (2) \\
&= [f(X)-\widehat{f(X)}]^2+Var[\epsilon] && (3) \\
&= Reducible + Irreducible && (2.3)
\end{aligned}
$$

이 책의 주요 내용은 reducible error을 줄이는데 있다.

* inference: 우리는 가끔 X가 변함에 따라 Y가 어떠한 방식으로 변하는지 궁금하다. 이러한 상황에서는 Y를 예측하는 것이 아니라 X와 Y의 관계를 세부적으로 이해하는 것이 목표이다. $\widehat{f(x)}$은 black box이지만 우리는 그것의 정확한 형태를 알고싶다.

* 어떤 모델링은 prediction과 inference모두를 할 수 있다. 이렇게 우리의 관심사가 무엇이냐에 따라서 f를 측정하기 위한 다른 방법들을 사용한다. 예를 들어 linear models은 simple하고 interpretable한 inference로 적당하지만 다른 방법만큼 정확한 예측을 하지는 못한다. 하지만 나중에 non linear models은 꽤 정확한 예측을 하기도 한다.

2.1.3 How Do We Estimate f?

우리는 이 책에서 f를 측정하기 위한 다양한 방법들을 배운다. 그런데 그러한 방법들은 어떠한 특징을 공유하는데 그것을 이 파트에서 알아보자. 우리는 n개의 다른 데이터를 관찰하고 이것을 training data라고 부른다. x_{ij}는 jth predictor for observation i를 뜻한다. (i번째 관찰했을 때의 j번째 예측치) y_i는 i번째 반응 변수이다. 우리는 통계적 learning을 training data에 적용해 unknown function인 f를 예측하는 것이 목표다. 다시 말해서 어떠한 관측지 (X,Y)에 대해서도 $Y \approx \widehat{f(X)}$인 것을 찾고 싶은 것이다. 이러한 task를 위한 통계적 learning은 parametric methods 와 non parametric methods로 나눌 수 있다.

* parametric methods

<책의 원문>
1. First, we make an assumption about the functional form, or shape, of f. For example, one very simple assumption is that f is linear in X:

2. After a model has been selected, we need a procedure that uses the training data to fit or train the model. In the case of the linear model fit (2.4), we need to estimate the parameters $\beta_0, \cdots \beta_p$.

다시 말해서 f를 예측하는 문제(완전히 임의적인 p 차원을 가지는 f를 예측)에서 parameters를 예측하는 문제로 문제가 줄어든다. 일반적으로 parameters을 예측하는 것이 훨씬 쉽기 때문에 문제를 심플하게 만들어주는 장점이 있다. 하지만 단점은 우리가 선택하는 모델이 보통 true unknown form of f와 맞지 않는다는 점이다. 우리는 이러한 문제를 flexible models을 사용함으로써 해결할 수 있다. 하지만 이것은 또 굉장히 많은 parameters를 예측하는 것을 요구한다. 이러한 복잡한 모델은 overfitting문제를 일으킬 수 있다.

* non parametric methods

non-parametric methods는 f의 형태에 대해 명시적인 추측을 하지 않는다. 대신 data points에 가능한 가장 가까이 가는 f의 estimate를 찾는다. 이러한 방법은 특정한 형태의 f 함수를 피함으로써 더 정확한 f의 모양을 가질 수 있다는 장점이 있다. 하지만 major 단점있는데, 이는 정확한 f의 estimate을 얻기 위해 굉장히 많은 관측치가 필요하다는 것이다. 또한 주어진 데이터에 매우 딱 맞게 적합을 하기 때문에, 과적합의 우려가 있다.

참고
적합: 적합을 한다는 것은 주어진 데이터로 f를 추론(estimate)하는 과정이라고 생각하면 된다.
과적합: 말 그대로, 과하게 적합되었다는 뜻이다. 데이터가 주어지면 그에 맞게 적합을 시키는데, 이 때 너무 과도하게 맞춤형으로 적합이 된 상태이다. 주어진 데이터에 이렇게 잘 적합되었는데 오히려 좋은 것이 아닌가 하는 의문이 들 수도 있다. 환자의 데이터를 통해 질병을 예측하는 경우를 생각해보자. 기존의 환자들의 데이터로 모델(모델은 f라고 생각하면 편하다)을 적합할 것이고 이를 통해 f에 대한 estimate, 즉 $\hat{f}$를 도출할 것이다. 하지만 여기서 우리의 관심사는 기존의 환자들의 질병 여부를 예측하는 것이 아니라 새로운 환자의 데이터가 주어졌을 때 이 환자의 질병 여부를 예측하는 것이다. 새로운 환자의 데이터는 기존 환자의 데이터와 비슷하다는 보장이 없기 때문에 f를 주어진 데이터에 딱 맞게 적합을 하면 새로운 데이터에 대한 예측 정확도가 떨어질 수밖에 없다. 이는 이후의 test MSE, training MSE에서 자세히 다룬다.

2.1.3 the trade-off between prediction accuracy and model interpretability

여기서 공부하는 여러 모델은 flexible한것도 있고 restrictive한 것도 있다. 그중 restrictive model을 선호하는 이유가 몇가지 있다. 만약 inference가 목표라면 linear 같은 inflexible한

것이 좋다. 그래야 x와 y의 관계를 설명하기가 용이하다. 하지만 매우 flexible한 모델은 매우 복잡한 estimate로 이어져 어떠한 개개인의 predictor가 반응과 연관되어 있는지 알기가 어렵다. 즉, inference을 하기가 어려워진다.
회사에서 어떤 데이터로 모델을 설계했다고 하자. 우리가 설계한 모델을 데이터 분석에 대해 잘 모르는 상급자에게 설명해야할 상황이 올 것이며 Boosting 기법 같은 복잡한 모델을 사용했다면 설계한 모델에 대한 설명이 어려울 것이다. 이와 같이 상황에 따라서 flexible한 모델과 inflexible한 모델을 선택해야 한다.
하지만 어떤 상황에서는 prediction이 목표이다. 이러한 상황에서는 flexible한 모델을 사용하는 것이 좋다.
다시 말해, flexibility와 interpretability는 서로 상충관계에 있다고 보면 된다. 두 마리의 토끼를 동시에 잡기는 힘드므로 적절한 선택을 하자.
그렇지만 놀랍게도 꼭 이러한 것은 아님. less flexible한 모델로 정확한 prediction을 얻을 수 있다. 이러한 현상은 매우 flexible한 방법에서 과적합의 잠재성과 관련이 있는데 나중에 논의하기로 한다.

2.1.4 Supervised learning vs UnSupervised learning

대부분의 통계적 learning은 이 두가지 범주로 나뉜다. 지도 학습은 반응 변수가 주어져 있고. 비지도 학습은 반응 변수가 주어져 있지 않다. 비지도 학습의 예로는 cluster가 있다. cluster의 목적은 어떠한 x_i를 확연한 그룹으로 분류하는 것이다. 여기서 각 데이터는 어떤 cluster에 속하는지 사전에 label이 주어지지는 않는다. 따라서 이를 비지도 학습의 한 예라고 볼 수 있다.
때대로는 어떠한 분석이 지도학습인지 비지도학습인지 구분하는 것이 덜 중요할 때가 있다. 이러한 상황을 semi-supervised learning이라고 한다.

2.1.5 regression vs classification problems

반응변수가 연속형 변수일 때, regression이라고 하며 이산형 변수일 때, classification이라고 한다.
regression의 기법으로는 Linear Regression, Smoothing 등이 있고 classification으로는 LDA, SVM 등이 있다.

2.2 assessing model accuracy

이 파트에서는 통계적 기법을 선택하는 과정에서 일어나는 중요한 개념들을 짚을 것이다.

2.2.1 measuring the quality of fit

어떠한 통계적 기법이 잘 사용되는지 알려면 그것의 predictions가 얼마나 잘 관측된 데이터와 맞는지 알려주는 척도가 필요하다. regression에서는 가장 보편적으로 이것을 위해 mean squared error(mse)를 사용한다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f(x_i)})^2 \tag{2.5}$$

MSE는 i번째 관측치와 예측치가 가까울수록 작아지고 멀수록 커질 것이다. 위의 공식은 training data로 계산이 되기 때문에 training MSE라고 부르는 것이 더 정확하다. 하지만 일반적으로 우리는 training data가 아니라 unseen test data에 대해서 얼마나 잘 작동하는지가 관심사다. 따라서 우리는 가장 적은 test MSE를 주는 방법을 선택한다. 보통 데이터 분석을 할 때, 모델을 적합시키는데 사용하는 데이터를 training data, 모델의 정확도를 평가하는데 사용하는 데이터를 test data라고 한다.(5장에서 자세히 다룬다) 다른 말로 우리는 average squared prediction error for test observations을 계산해야 한다.

$$Ave(y_o - \widehat{f(x_o)})^2 \tag{2.6}$$

어떤 상황에서는 test data가 이용가능할 것이다. 그렇다면 이를 계산해서 가장 적은 test MSE를 산출하는 모델을 고르면 된다. 그런데 만약 test 관측지가 없다면? 그렇다면 걍 training MSE가 가장 낮은 애를 고르는 것으로 생각할 수 있다. 왜냐하면 training MSE와 test MSE가 서로 가깝게 연관된다고 쉽게 생각할 수 있기 때문이다. 하지만 가장 낮은 training MSE가 가장 낮은 test MSE를 보장한다고 말할 수 없다.

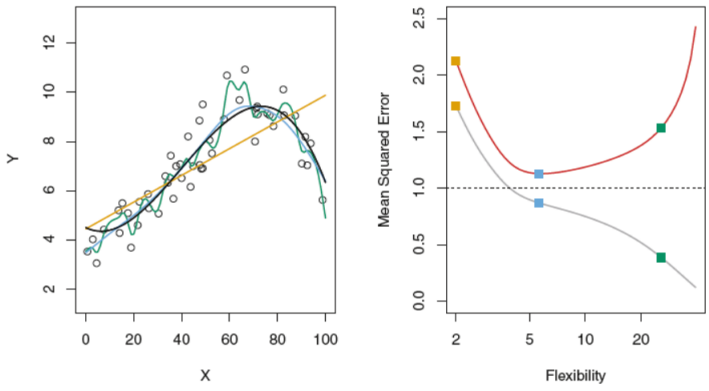

FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

figure 2.9에서 오른쪽은 flexible에 따른 training MSE와 test MSE의 변화 추이이다. 다시 말하면 자유도 (degrees of freedom) 에 따른 변화 추이이다.[자유도란 모델의 flexibility와 관련된 개념이라고 이해하면 된다.] training MSE는 지속적으로 낮아지지만 test MSE는 낮아지다가 어느 순간이 되면 다시 올라간다. 한편, 수평선은 irreducible error이고 앞서 살펴본 $Var(\epsilon)$ 이다. test MSE의 최소값이 수평선보다 위에 있는 것으로 보아 irreducible error의 의미와 일맥상통함을 확인할 수 있다.

figure 2.9에서 볼 수 있듯이 통계적 방법의 flexibility가 올라갈수록 training MSE는 지속적으로 낮아지지만 test MSE는 U자 모양이다. 이것은 통계적 learning의 근본적인 특성에서 기인한다. 만약 주어진 방법이 small training MSE but large test MSE라면 과적합되었다고 한다. 왜냐하면 통계모형이 training data에서만 패턴을 너무 열심히 찾았기 때문에 진정한 unknown function f의 특성을 잘 모르고 새로운 데이터가 주어질 때 정확한 예측을 할 수 없는 것이다.

주목할 만한 것은 test MSE는 항상 training MSE보다는 크다. 왜냐하면 통계적 모형이 직간접적으로 training MSE를 최소화하는 것을 찾기 때문이다. 하지만 이에 속지 말자. 우리의 관심사는 주어진 데이터가 아니라 새롭게 주어진 데이터에 대해서 관심있는 종속변수를 얼마나 잘 예측하는지임을 명심하자.

2.2.2 the bias-variance trade-off

test MSE는 아래의 분해에 의해 U자 모양을 취한다.

$$E[(y_o - \widehat{f(x_0)})^2] = Var(\widehat{f(x_0)} + [Bias(\widehat{f(x_o)}]^2 + Var(\epsilon) \tag{2.7}$$

<증명>

$$\begin{aligned}
E[(y_o - \widehat{f(x_0)})^2] &= E[(y_o - f(y_0) + f(y_0) - \widehat{f(x_0)})^2] \\
&= E[(y_o - f(y_0))^2] + E[(f(y_0) - \widehat{f(x_0)})^2] + 2E[(y_o - f(y_0))(f(y_0) - \widehat{f(x_0)})] \\
&= Var(\epsilon) + MSE(\widehat{f(x_0)}) + 2E[y_o - f(y_0)]E[f(y_0) - \widehat{f(x_0)}] \\
&= Var(\epsilon) + E[(\widehat{f(x_0)} - E[\widehat{f(x_0)}] + E[\widehat{f(x_0)}] - f(x_0))^2] + 2E[\epsilon]E[f(x_0) - \widehat{f(x_0)}] \\
&= Var(\epsilon) + E[(\widehat{f(x_0)} - E[\widehat{f(x_0)}])^2] + E[(E[\widehat{f(x_0)}] - f(x_0))^2] \\
&+ 2E[(\widehat{f(x_0)} - E[\widehat{f(x_0)}])(E[\widehat{f(x_0)}] - f(x_0))] \\
&= Var(\epsilon) + Var(\widehat{f(x_0)}) + [Bias(E[\widehat{f(x_0)}] - f(x_0))]^2
\end{aligned}$$

이 공식에 의해서 작은 test MSE를 가지고 싶으면 낮은 var와 낮은 bias를 가져야 한다.

여기서 주목해야 할 점은 var와 bias의 제곱은 음수가 될 수 없으므로 애초에 test MSE는 $Var(\epsilon)$ 이상이라는 것이다. 다시 한번 말하면, $Var(\epsilon)$ 는 irreducible error이다.

그렇다면 var와 bias의 뜻은 무엇일까. var은 우리가 다른 training set을 넣음으로써 생기는 $\hat{f}$의 변화라고 생각하면 된다. 하지만 이상적으로 다양한 training set에 대해서 var은 작은게 좋다. 즉 변화게 덜 일어나는게 좋다. 높은 var은 데이터에 작은 변화만 있어도 $\hat{f}$에 큰 변화를 준다. 일반적으로 더 flexible 한 모델이 더 높은 var을 가진다.(flexible 하다는 것은 wiggly 하다는 것이고 그것은 하나의 데이터가 기존 training set에 다르게 나온다면 변화의 폭이 더 커질 것이다)

반대로 bias는 실제 문제에서 측정하는데 생기는 error이며 오차라고 생각하면 된다.

일반적으로 더 flexible 할수록 더 bias가 더 적고 덜 flexible 할수록 bias가 더 크다. 더 flexible 하다는 것은 구불구불하다는 것이고 var은 크지만 그만큼 주어진 데이터에 잘 적합되었다는 뜻이므로 bias 즉 오차는 더 작다. 덜 flexible하다는 것은 linear line처럼 직선의 형태이고 이렇다면 var은 작은 대신 bias는 큰 것이다.

우리가 한 모델의 flexibility을 늘릴수록 bias는 처음에는 var가 올라가는 것보다 더 빨리 내려간다. 따라서 test MSE가 작아진다. 하지만 어느 시점에가서는 flexibility을 올리는 것이 bias에 거의 영향이 없고 var이 급격하게 올라가게 됨. 그러면 test MSE는 증가한다. 이러한 bias, var, test MSE 사이의 관계를 bias - var trade-off라고 부른다. var와 bias가 둘다 낮은 모델을 찾는 것이 바로 어려운 과제이다. 이 책에서 계속 trade-off를 다루게 될 것임. 실제로는 이 세 가지를 계산하기 어렵지만(test MSE를 계산하는 경우는 많지 않으므로) 그럼에도 이 trade-off관계를 명심해야한다!

2.2.3 the classification setting

classification문제에서 $\hat{f}$의 정확도를 측정하기 위해서 사용되는 지표는 the training error rate이다.

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \tag{2.8}$$

하지만 우리는 regression에서처럼 training이 아니라 test data에 대한 error rate가 궁금하다. 좋은 classifier은 test error rate가 가장 작은 것이다.

$$Ave(I(y_o \neq \hat{y_0})) \tag{2.9}$$

* the bayes classifier

It is possible to show (though the proof is outside of the scope of this book) that the test error rate given in (2.9) is minimized, on average, by a very simple classifier that assigns each observation to the most likely class, given its predictor values. In other words, we should simply assign a test observation with predictor vector x_0 to the class j for which $P(Y = j \mid X = x_0)$ is largest.

bayes classifier은 가장 낮은 test error rate를 만든다. 이것은 bayes error rate라고 불린다.

$$1 - E(max_j Pr(Y = j \mid X)) \tag{2.11}$$

bayes classifier은 x_0가 주어진 상태에서 해당 범주에 속할 확률이 가장 큰 애를 그 범주로 보내기 때문에 error는 1에서 빼면 된다.

* K-Nearest Neighbors

이론적으로 항상 bayes classifier를 사용하고 싶지만 실제에서는 conditional distribution of y given x를 모르므로 사용하는 것은 불가능하다. 그래서 KNN 같은 것으로 대체를 한다. 자연수 K와 test observation x_0가 주어진다면 KNN은 x_0와 가장 가까운 training data K 개를 찾고 걔네들이 class j와 같은 확률을 계산한다. 즉, test observation x_0가 주어졌다는 점에서 conditional prob로 계산하는 것이다. 그리고 마지막으로 bayes rule을 적용해 가장 큰 확률을 가지는 class로x_0를 분류한다.

$$Pr(Y = j \mid X = x_o) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j) \tag{2.12}$$

위의 수식은 사실 =보다는 ≈라고 보는 것이 타당하다. 조건부 분포를 알지 못하기 때문에 정확한 확률을 계산할 수 없어 상대도수로 확률을 estimate 한 것이다.KNN은 매우 간단하지만 bayes classification과 비슷할 정도의 성능을 보일 때가 많다. KNN에서는 K의 선택이 KNN 성능에 지대한 영향을 끼친다. K = 1일 때는 매우 flexible 해서 bias는 낮지만 var은 매우 높다. K가 커질수록 less flexible해지고 decision boundary는 linear에 가까워진다. 이것은 low var but high bias에 해당한다.

regression과 마찬가지로 training error rate와 test error rate간에 강력한 관계는 없다. regression 이든 classification이든 적절한 수준의 flexibility를 선택하는 것은 통계적 모형의 성공에 있어서 아주 중요하다.

* 참고사항

bias : 예측한 값의 평균(기댓값)과 실제 값의 차이. 모델의 bias는 초록색선 하나만 가지고 말하는 것이 아니라 다른 data set를 가지고 모델을 만들고 그런 식을 초록색선을 평균낸 것과 실제 값이 차이임. 하나의 data set으로 구한 모델과의 차이가 아니라 여러번 표본을 뽑았을 때 각각 적합시키고 걔네들과 각각 실제 f와의 차이임.$Bias = E[\hat{f}] - f(x)$ 우리가 가지고 있는 data set에서가 아니라 모집단에서 뽑은 여러 data set을 통해서 만들어지는 것임. 다시 말해 실제로 구현할 수가 없음. 다시말해, 추상적인 얘기임.