Exercise 2

1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

   편향과 분산의 차이는 무엇일까?
   편향: 모델과 실제 값 사이의 차이→ 편향이 클 수록 fitting이 어렵다
   분산: 모델이 다른 데이터 셋에 대해 얼마나 민감한지→ 따라서 분산이 높다는 것은 학습 데이터에만 너무 맞춰져 새로운 데이터(예측)를 집어 넣을 경우 성능이 급감→ overfitting(모델이 학습 데이터셋에 너무 특화되어, 학습 데이터에는 잘 맞지만 새로운 데이터에 대해 일반화되지 못하는 상황)

   모형이 inflexible할 수록 (즉, linear) variance는 작지만 error(bias)는 커진다. 반대로 모형이 flexible할 수록(즉, Non-linear) error(bias)는 작지만, variance는 크다

   가장 좋은 모델은 bias도, variance도 모두 작은 모델이지만, 일반적으로 둘은 trade-off(상충관계)에 있기 때문에 그런 모델을 찾는 것은 쉽지 않다.

   (a) The sample size n is extremely large, and the number of predictors p is small.

   샘플의 크기가 매우 크고 예측해야할 변수가 작은 경우엔 flexible model이 더 적합하다. 샘플 사이즈가 매우 크기 때문에 fitting을 매우 잘 시킬 수 있고(과적합 우려 X) floxible model이기 때문에 bias의 위험도 낮다.

   (b) The number of predictors p is extremely large, and the number of observations n is small.

a와 반대로 샘플의 수가 작은데 예측해야 할 변수가 큰 경우엔 inflexible model이 더 적합하다. 우선 샘플 사이즈 수가 적기 때문에 flexible model로 했다간 과적합의 문제에 봉착할 수 있다.

(c) The relationship between the predictors and response is highly non-linear.

둘 사이의 관계가 non-linear할 경우, flexible model이 더 적합하다. 비선형적인 관계는 flexible model이 더 fitting이 잘 되기 때문이다. 만약 이 경우에 inflexible model을 사용할 경우, bias가 너무 커질 위험이 있다.

(d) The variance of the error terms, i.e. $\sigma 2 = Var(\varepsilon)$, is extremely high.

분산이 굉장히 높을 경우, 다른 training data를 사용했을 때 기존 data와의 편차가 굉장히 커진다. 즉, 과적합 문제가 발생할 수 있기 때문에 이런 데이터에 대해 flexible model을 사용하는 것은 옳지 못하다

---

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

   classification이냐 regression이냐의 문제는 해당 모델의 변수가 수치형(quantitative)인지 아니면 범주형(qualitative)인지에 따라 결정된다. 일반적으로 수치형인 경우엔 regression 문제로, 범주형인 경우엔 classification 문제로 처리한다

   (a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

   →500개의 기업을 뽑고, 각 기업의 이익(수치형), 직원수(수치형) 업계, CEO 연봉(수치형)을 변수로 활용하기 때문에 Regression으로 볼 수 있다. 이 경우, n=500, p=3(이익, 직원수, ceo연봉)으로 볼 수 있다.

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

→
성공할지, 실패할지(범주형) 알아보고 싶어하기 때문에 이는 classification으로 볼 수 있다. 이경우 n=20, p=13

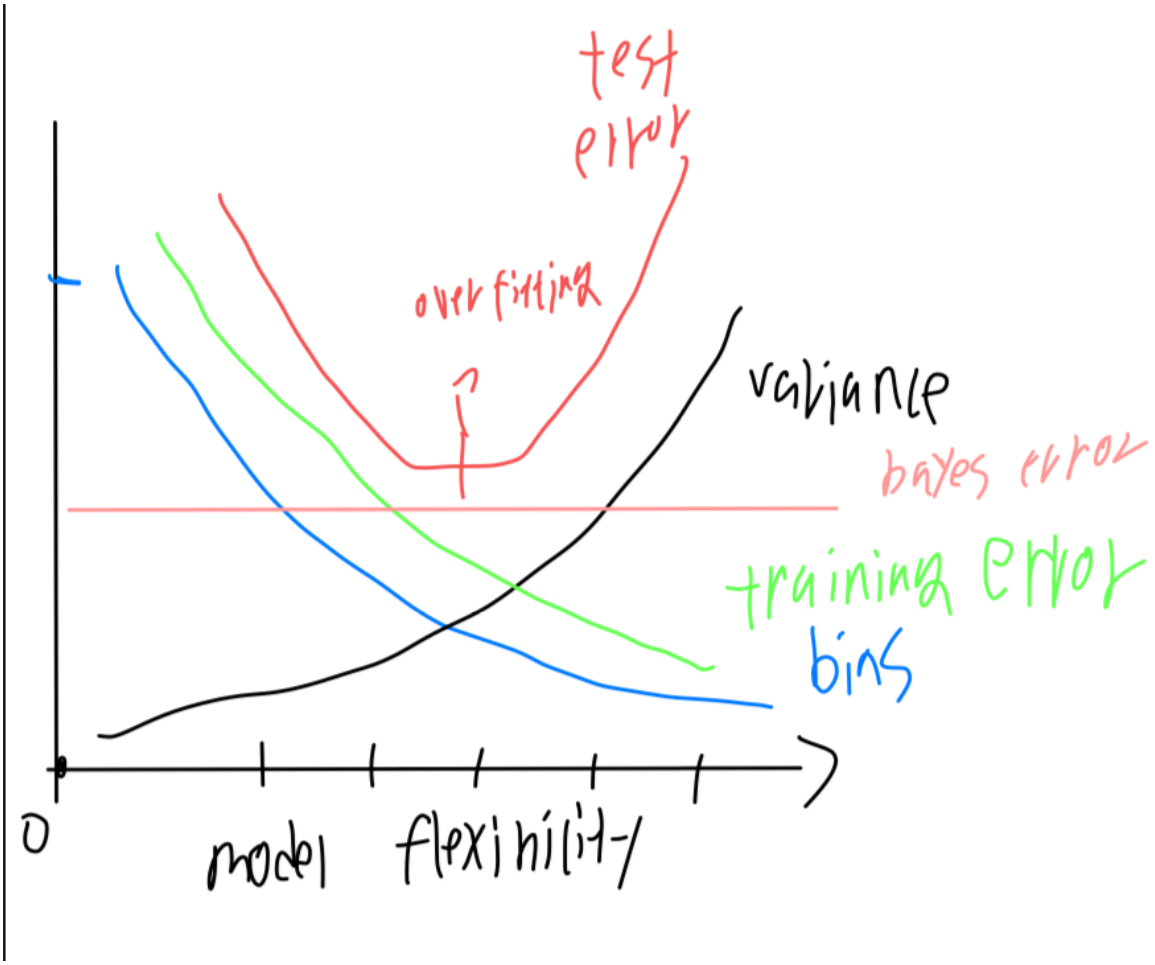
(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

→ 2012년 전체의 weekly data를 수집하고, 각 주마다 US market, British market, German market의 변동을 변수로 하기 때문에 이는 Regression(그중에서도 prediction)에 해당한다. 이 경우 n=52(1년은 52주), p=3

---

3. We now revisit the bias-variance decomposition.

(a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(b) Explain why each of the five curves has the shape displayed in part (a).

1. bias
   flexibility가 커질 수록, 차수가 점점 증가하기 때문에 fitting이 잘 될 수 있고 fitting 이 잘 된다는 것은 bias가 작아진다는 것을 의미한다

2. variance
   bias와 반대로 flexibility가 커질 수록, fitting이 잘되는 것은 맞지만, 그만큼 학습 데이터에만 익숙해져 다른 새로운 데이터가 등장했을 때 편차가 매우 커질 수 있다

3. Training error
   학습데이터에서 발생하는 에러는 당연히 flexibility가 커지면 fitting이 잘되기 때문에 줄어든다

4. test eroor
   flexibility가 높아지면 자칫 학습데이터의 오차마저 반영해버리는 overfitting문제가 발생할 수 있기 때문에 test error은 어느 정도 작아지다가 overfitting이 발생한 후

부터는 다시 점차 오르게 된다

5. bayes error
   얘는 모델의 유연성과는 관계가 없기 때문에 일정한 값을 가진다.

---

4. You will now think of some real-life applications for statistical learning.

   (a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

   금융쪽에서 classification이 쓰이는 경우는 아마 주가조작 같은걸 탐지할 때 쓰이지 않을까 생각한다 금융감독원 같은 기관에서 이상 매수신호 같은 걸 감지하면 경보 같은게 뜨는 걸로 알고 있는데 그런게 classification이 쓰이는 경우가 아닐까? 이 경우엔 종속변수 매수가격이 될 것이고 독립변수로는 시계열 데이터로 이루어진 그 동안의 해당 기업 매수 가격이 될 것 같다

   그리고 또 다른 것

   (b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

   (c) Describe three real-life applications in which cluster analysis might be useful.

---

5. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

   flexibile한 모델의 경우 inflexible 모델에 비해 fitting이 잘된다는 것과 그렇기 때문에 bias는 낮다는 장점이 있지만, 동시에 overfitting의 문제, varience가 높은 것, data가 많이 필요하다는 단점이 있다

   따라서 샘플수가 많고 predictors수는 적거나 X, Y 두 변수간에 비선형적인 패턴을 보일 때는 flexible한 모델이 적절하고, 반대로 샘플수는 적은데 predictors 수는 많

거나 X,Y 두 변수 사이에 선형적인 패턴이 있을 때는 inflexible한 모델이 더 적합하다.

---

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?

   parametric methods의 경우 함수를 추정해야 하는 문제에서 각 parameters인 β값들을 찾는 것으로 문제가 단순화된다는 장점과 그렇기 때문에 딱히 많은 observations가 필요하지 않지만 우리가 각 Parameters 값을 찾아서 얻은 $\hat{f}$값이 실제 f와 맞지 않을 수 있다는 단점이 있다.
   반대로 non-parametric methods의 경우 최대한 실제 f값과 가깝게 되도록 fitting이 들어가기 때문에 실제 f값과의 차이는 굉장히 적지만, 정확하게 이를 추정하기 위해선 굉장히 큰 수의 observation이 필요하다.

   classificaiton은 정확한 분류가 들어가야 하기 때문에 non-parametric methods가 적절할 것 같고 regression은 parametric methods가 적절할 것으로 보인다

---

7. The table below provides a training data set containing six observa- tions, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	−1	0	1	Green
6	1	1	1	Red

   Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point,
X1 =X2 =X3 =0.
→ 테스트 포인트 (0,0,0)과 각각의 Obs와의 거리를 구하면 아래와 같다

$D1 : \sqrt{9} = 3$
$D2 : \sqrt{2^2 + 0^2 + 0^2} = \sqrt{4} = 2$
$D3 : \sqrt{1^2 + 3^2} = \sqrt{10}$
$D4 : \sqrt{1^2 + 2^2} = \sqrt{5}$
$D5 : \sqrt{(-1)^2 + 1^2} = \sqrt{2}$
$D6 : \sqrt{1^2 + 1^2 + 1^2} = \sqrt{3}$

(b) What is our prediction with K = 1? Why?
→ K=1인 경우 Test point인 (0,0,0)과 가장 가까운 점은 D5이고 D5는 Green이므로 (0,0,0) 역시 Green일 것이라 예상할 수 있다

(c) What is our prediction with K = 3? Why?
→ K=3인 경우 Test point인 (0,0,0)과 가장 가까운 점 3개는 각각 D5, D6,D2이고 이는 Green 1개 Red 2개 이므로 Green일 확률이 1/3, Red일 확률이 2/3이기 때문에 Red라고 예측할 수 있다

(d) If the Bayes decision boundary in this problem is highly non- linear, then would we expect the best value for K to be large or small? Why?
→ K를 올릴수록 이 방법은 inflexible해지고 따라서 분산은 낮아지고 bias는 높아지기 때문에 결국 decision boundary는 linear한 모양과 가까워진다. 따라서 decision boundary가 non-linear한 경우인 K는 small해야한다.