Statistical Learning
Chapter II Exercises

ex 1

2.4 Exercises

Conceptual

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.

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

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

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

   (d) The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.

a) flexible is better ✓ (n large)
b) flexible is worse ✓ (n small)
c) flexible is better ✓ (f is highly non linear
d) flexible is worse ✓ (with high Var(ε)
   linear model can fit data simpler and well enough

b) Worse. a flexible method overfit a small number of observations

c) better - with more degrees of freedom a flexible model would obtain a better fit

d) worse - flexible method fit to the noise in the error term an increase variance

ex2

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.

   (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.

   (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.

   (c) We are interesting in predicting the % change in the US dollar 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 dollar, the % change in the US market, the % change in the British market, and the % change in the German market.

a) ✓ regression ; inference $n = 500$ ✓ $p = 3$ ✓
Q is which factors explain CEO salary

b) ✓ classification ; prediction $n = 20$ ✓ $p = 13$ ✓

c) ✓ regression , prediction $n = 52$ ✓ $p = 3$ ✓

ex 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

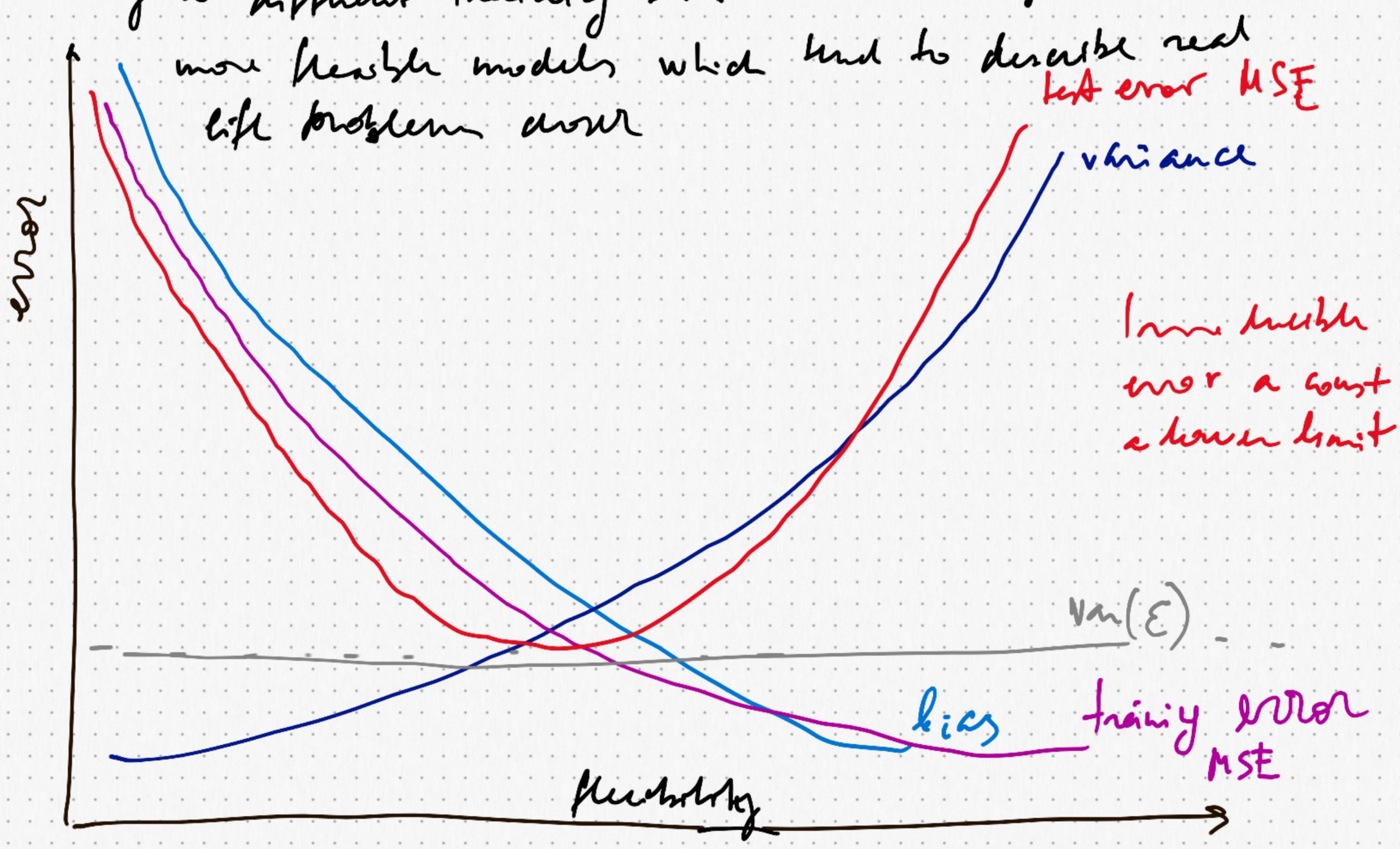
the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

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

bias: approximation of a real life problem by a much simpler model

bias big in simpler less flexible models

variance: by how much the model $\hat{f}$ changes if it is estimated by a different training set. variance is higher in more flexible models which tend to describe real life problem closer



training error is monotonically decreasing in test data because flexible methods fit the data better and overfit it

test error has a minimum above the irreducible error

It grows because very flexible models overfit the data.

ex 5

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?

1. More flexible methods tend to fit better complex (non-linear models) data relationships. Their bias is smaller. The disadvantage is a higher risk of an overfit. Hence the importance to seek a bias-variance tradeoff

2. A more flexible approach is preferred when there is a lot of data and the residual error is not too large. (more flexible: more interest in prediction)

3. A less flexible approach is preferred when the data has a highly linear relationship. When we are interested in the interpretability of the results.

Ex 6

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?

1. Parametric approach consists selecting a linear model and estimating its $p+1$ parameters. Non-parametric approach does not take any assumptions about the form of $\hat{f}$! and so requires a very large number of obs

2. the advantages of a parametric approach is that linear model parameters can be estimated relatively easily. and not as many observations are needed

3. The disadvantages: the parametric model can miss capturing the complex data relationship and to have a less predictive power

ex 7

7. The table below provides a training data set containing six observations, 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 $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

obs	dist
1	3
2	2
3	3.11
4	2.23
5	1.4
6	1.7

(b) What is our prediction with $K = 1$? Why?

(c) What is our prediction with $K = 3$? Why?

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the *best* value for K to be large or small? Why?

a) Euclidian distance for each observation to test point (0,0,0)

b) for k = 1 prediction in green (class of 1 nearest neighbor)

c) for k = 3 prediction is red (most frequent class of 3 nearest neighbors)

d) if the Bayes decision boundary is highly non linear the best value for K has an optimal value due to a concave form of the test error rate curve

K should be small. A small K would be flexible for a non-linear decision boundary, whereas a large K would try to fit a more linear boundary.