exercise 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.

exercise 1

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

exercise 1

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

better: *because a large dataset will reduce the variance of the model, making it easier for a more flexible model to approximate the true function.*

exercise 1

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

exercise 1

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

worse: *the model will have high variance, and since it has a high number of predictors probably it will tend overfit all the parameters*

exercise 1

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

exercise 1

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

depends: *the inflexible method will have high bias, but if we also have few data-points the model will have very high variance and the total error might be higher than a linear model. If we have a high number of data-points, then the more flexible model should have better performance*

exercise 1

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

exercise 1

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

worse: *the inflexible method will try to fit the very variable points and find patterns that aren't present*

exercise 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.

exercise 2

(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.

exercise 2

(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.

scenario: regression | *interest:* inference| *n:* 500 | *p:* 3

exercise 2

(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.

exercise 2

(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.

scenario: classification | *interest:* prediction| *n:* 20 | *p:* 13

exercise 2

(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.

exercise 2

(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.

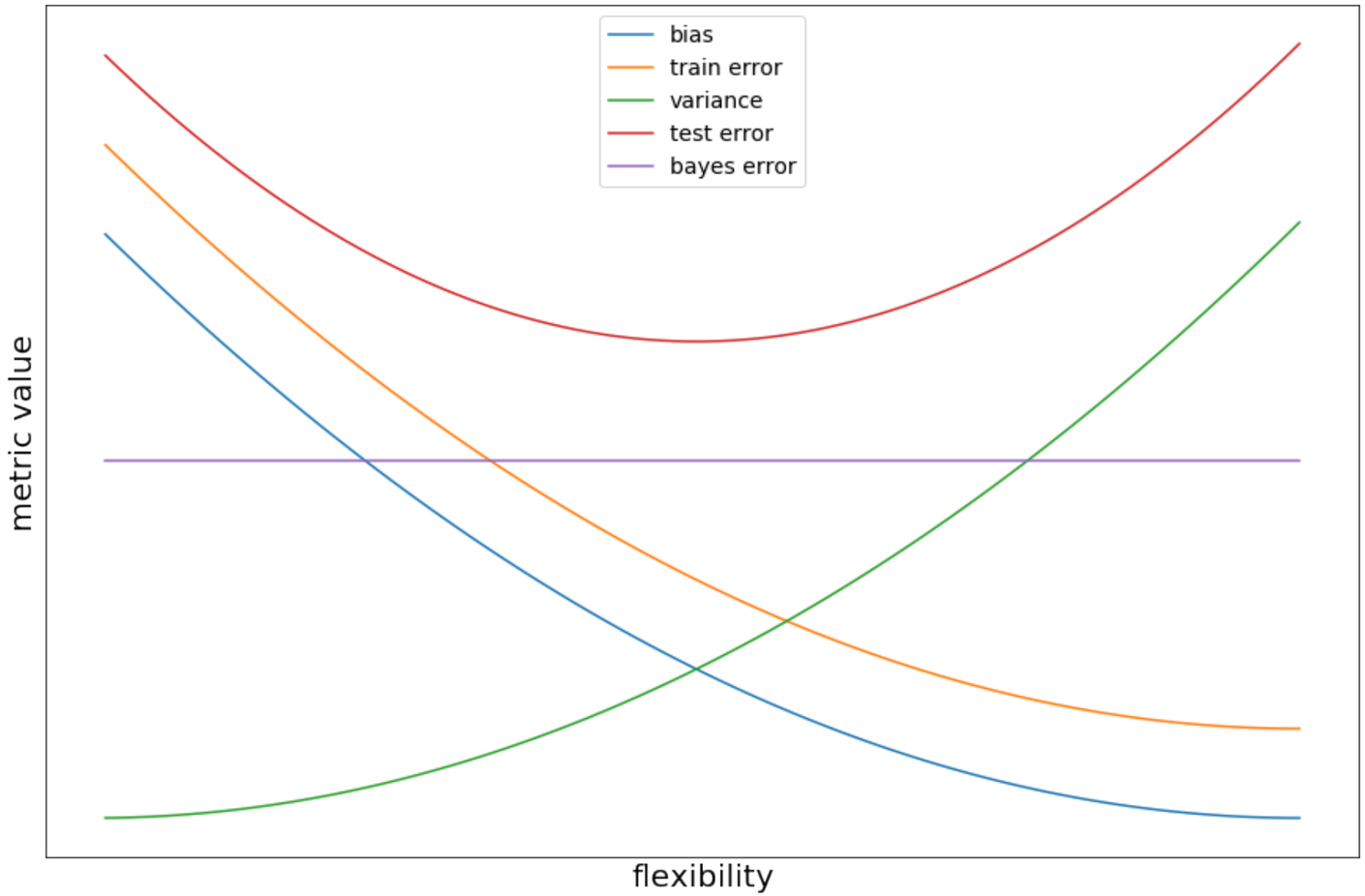
scenario: regression | *interest:* prediction| *n:* 52 | *p:* 3

exercise 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.

exercise 3


bias
train error
variance
test error
bayes error
metric value
flexibility

exercise 3

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

exercise 3

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

bayes error: *the irreducible error is constant for any model*

exercise 3

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

variance: *the more flexible the model is, the more it will try to adjust to the data-set*

exercise 3

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

bias: *the more flexible the model is, the more complex relationships it can capture*

exercise 3

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

test error: *it starts to decrease together with the model bias, but once the variance starts to increase it increases due to overfitting*

exercise 3

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

training error: *it decreases as more flexible models start to "learn" the data*

exercise 4

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

exercise 4

(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.

exercise 4

(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.

i: *predict whether a person will have cancer in the next 5 years given his/her age, gender and how many cigarettes smokes a day*

exercise 4

(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.

ii: *estimate how eating fruits affect the probability of getting stomach cancer*

exercise 4

(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.

iii: *detect whether a prospect client will fail or not given its wage and job title*

exercise 4

(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.

exercise 4

(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.

i: *estimate how the gender, years of experience and education affect the income or a person*

exercise 4

(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.

ii: *predict the price of a house given the size in sq meters, number of bedrooms, and parking slots*

exercise 4

(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.

iii: *predict what will be the order value that a user will make given the products he has seen*

exercise 4

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

exercise 4

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

i: *group different groups of clients but the kind of products they buy*

exercise 4

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

ii: *decide where to open comic stores in a city given the lat and lon of schools*

exercise 4

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

iii: *group the location of stores in clusters to pre-process data to use it as input in a model*

exercise 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?

advantages: *capture more complex relationships yielding a lower bias*

disadvantages: *needs more data to yield a good model; are harder to interpret, making them worse suited for inference; more computation required to fit them; prone to overfit*

exercise 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?

a less flexible approach is preferred: *when we want to make inference, when we have a small data-sample, and the model is linear*

a more flexible approach is preferred: *when we want to make predictions, we have a huge data-sample, and the model is not linear*

exercise 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: *assumes a functional form of f defined by a set of parameters, so reduces the problem from estimating f to estimate a set of parameters that define f*

exercise 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?

non-parametric: *a non-parametric approach makes no assumptions about the functional form of* f*, but it finds it through grouping similar observations from the data-sample*

exercise 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?

advantages: *a parametric approach converges much more quickly to a set of parameters, thus requires a smaller data-set for training*

disadvantages: *since it makes assumptions about the form of* f *, these assumptions might be wrong which leads to inaccurate estimations of* f

exercise 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

exercise 7

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

exercise 7

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.

exercise 7

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.

obs	**1**	**2**	**3**	**4**	**5**	**6**
dist	3	2	3.16	2.23	1.41	1.73

exercise 7

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

exercise 7

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

Green: *because the 5th point is the closest one*

exercise 7

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

exercise 7

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

Red: *because the three closest points have [Red, Green, Red] as labels, since Red is the most common, thats the prediction*

exercise 7

(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?

exercise 7

(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?

small: *because the decision boundary with a small value of K would be more flexible*

that's it