***Lab 01 – Exercises***

*Conceptual*

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

*a.* Large *n* small *p*: Better. More flexibility will allow a better fit to a greater *n*.

*b.* Large *p* small *n*: Worse. Higher flexibility and lower *n* could overfit.

*c.* Highly non-linear relationship: Better. Less linearity is better fit by a more flexibility.

*d.* Variance of the error terms is extremely high: Worse. Flexible methods fit to the ***noise*** in the error terms, and increase the variance.

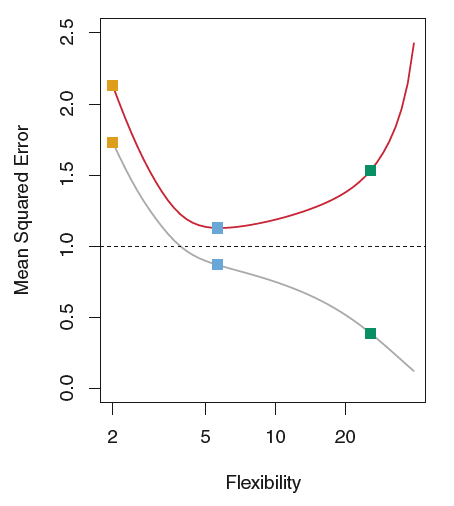
1. *Explain whether each scenario is a* ***classification*** *or a* ***regression*** *problem, and indicate whether we are most interested in inference or prediction. Finally,* ***provide n and p.***

*a.* n = 500, p = 3. Regression

*b.* n = 20, p = 13. Classification

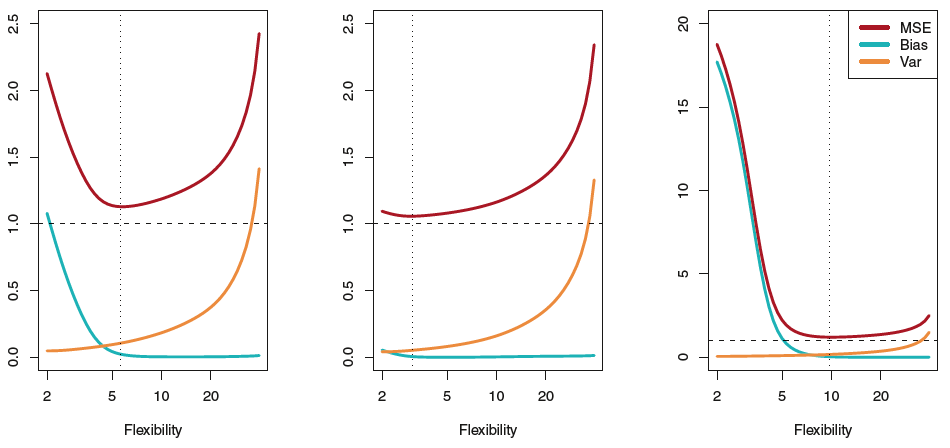
*c.* n = 52, p = 3. Regression

1. *Bias-Variance Decomposition*
2. *Provide a sketch of a* ***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 (label each one).*



**Training MSE** – Grey

**Test MSE** – Red



**Squared Bias** – Blue

**Variance** – Orange

**Bayes/Irreducible/Var()** – Dashed Line

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

Initial curve fitting may have high reducible error and variance. With

increased training and increased flexibility, the error rate will decrease for a time until a minimum is reached. This is likely the desired model. Further increasing flexibility will cause the model to interpret patterns that are not truly there, further increasing the error.

**Training MSE** – The error associated with this curve should decrease as flexibility increases as it has attempted to be fit to a large amount of data. ***[decreases monotonically because increases in flexibility yield a closer fit]***

**Test MSE** – The error initially decreases as flexibility increases, however, overfitting results from increasing the flexibility to high. ***[concave up curve because increases in flexibility yield a closer fit before overfitting]***

**Squared Bias** – “The difference between the averaged prediction at x0 and the true x0”. As flexibility increases the bias will increase because the model fits more of the data points. Variance, on the other hand, increases as flexibility increases (Bias-variance tradeoff.) This results in an initial decrease in the MSE and an increase at higher flexibility (overfitting). [***decreases monotonically because increases in flexibility yield a closer fit]***

**Variance** – “Variance between training datasets” will increase as flexibility increases because no one training data set is going to be similar enough to another to use a highly flexible model that fits one better. ***[increases monotonically because increases in flexibility yield an overfit model]***

**Bayes Error Rate/Irreducible Error/ Var()** – This value remains constant regardless of the flexibility of the model as it is inherently associated with the estimation of *f*. ***[defines the lower limit, the test error is bounded below by the irreducible error due to variance in the error in the output values (0<= value). When the training error is lower than the irreducible error, overfitting has taken place. The Bayes error rate is defined for classification problems and is determined by the ratio of data points which lie at the ‘wrong’ side of the decision boundary, (0 <= value < 1).]***

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

1. *Describe* ***3 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.*
2. Voting Outcome (Prediction) – The response would be either voted for party A or voted for party B. The predictors could be variables such as household income, years of education, and average daily television news consumption.
3. Distinguishing Brazilian Portuguese from Portuguese as spoken in Portugal (Inference) – The response would be a 1 for TRUE and a 0 for FALSE. The predictors would be the characteristic use of a few grammatical constructs, spelling, or specific words.
4. Nice day or Not (Inference) – The response would be a value wither falling in the range of comfortable and sunny or not. The predictors would be temperature and sunlight emission.
5. *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.*
6. Rate of change in volume over time for glaciers (Prediction) – The response would be the volume of glacial ice at a given time, while the predictors would be ambient temperature, surface temperature, solar radiance, etc.
7. Volume of water passing a set point at a given moment (Prediction) – The response would be total volume passing a certain point at the time of observation. The predictors would be water velocity, previous rainfall, and channel size.
8. Student GPA as a function of time spent studying (Prediction) – The response would be GPA and the predictor would be number of hours spent studying.
9. *Describe three real-life applications in which cluster analysis might be useful.*
10. Classifying microbes in a sample based on gene sequences
11. Determining the number of types of ground cover in an area based on the ratio of RG RB or GB in their spectral profiles.
12. Identifying subgroups within wage groups based on level of education and wage.
13. *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?*

The advantage for both is a better fit to the data.

For regression problems, high flexibility in a model may be desirable when a linear fit simply doesn’t describe the data well enough to be used. This would eliminate a larger portion of the bias in the model. However, being overly flexible would result in the model fitting patterns that don’t truly describe the overall trend of the data.

For classification problems, a highly flexible approach may be desirable when there is more variability in the data set. However, overfitting the data could lead to misclassification of some observations.

***[The advantages for a very flexible approach for regression or classification are obtaining a better fit for non-linear models, decreasing bias.***

***The disadvantages for a very flexible approach for regression or classification are requires estimating a greater number of parameters, follow the noise too closely (overfit), increasing variance.***

***A more flexible approach would be preferred to a less flexible approach when we are interested in prediction and not the interpretability of the results. A less flexible approach would be preferred to a more flexible approach when we are interested in inference and the interpretability of the results.]***

1. *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?*

A parametric learning approach makes an assumption about the functional form of *f* before the model is developed. Non-parametric methods don’t make any assumptions beforehand. A major consideration is optimal flexibility without overfitting the data with regard to parametric methods.