ASSESSING MODEL ACCURACY

Chapter 02 – Part II

Slides Inspired by content from IOM 530 “Applied Modern Statistical Learning Methods” – Gareth James (one of the authors of our book)

Outline

- Assessing Model Accuracy
  - Measuring the Quality of Fit
  - The Bias-Variance Trade-off
  - The Classification Setting

In-class exercise Part 1

- Complete the in-class exercise worksheet front side for Day 4 (problems 1 – 7)

Measuring Quality of Fit

- How do we evaluate a regression model's performance?
- One way: mean squared error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Where $\hat{y}_i$ is the prediction our method gives for the observation in our training data.

- **CONCEPT CHECK: What is n in the equation?**
- **Which is better – a higher, or a lower MSE?**
- **Why do we use mean *squared* error instead of mean error?**

Training vs. Test set performance

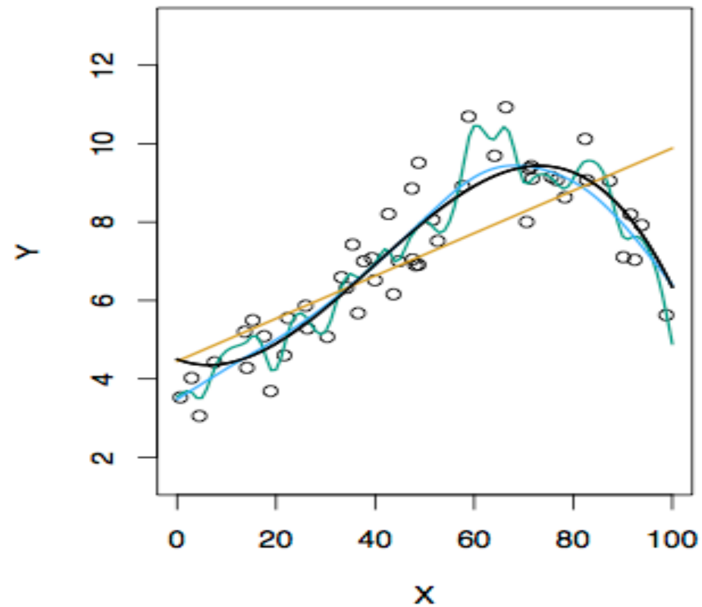
- Model Fitting: Choose model parameters such that MSE is minimized on the **Training Data**
- What we really care about is how well the method works on data that was not used for training (i.e. **Test Data**). Test data performance indicates the model's ability to *generalize*.
- There is no guarantee that the method with the smallest *training* MSE will have the smallest *test* MSE. If training performance is good and test performance is bad, the model has *failed to generalize*.

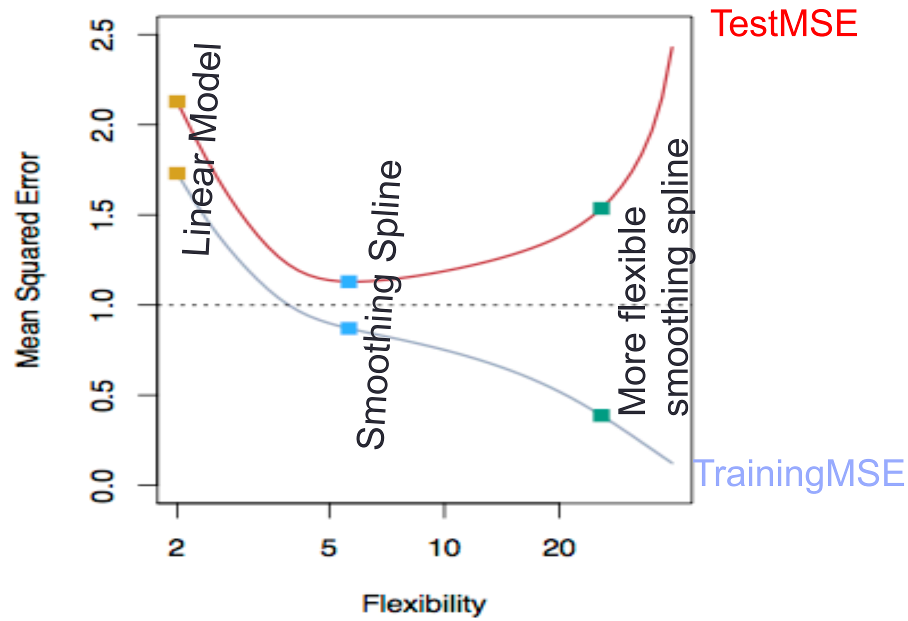
Training vs. Test errors

- In general the more flexible a method is the lower its training MSE … it will “fit” or explain the training data very well.

- In a more flexible model, the test MSE may be higher than in a less flexible model.
- **CONCEPT CHECK: Why would test MSE be larger than train MSE in a flexible model?**

Examples with Different Levels of Flexibility: Example 1



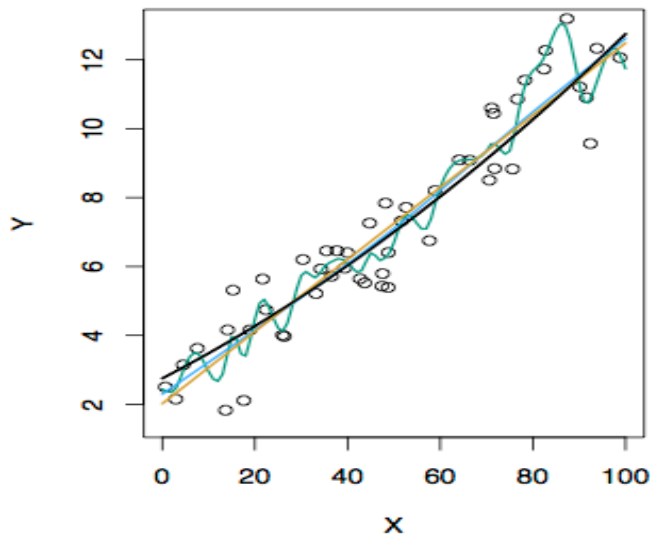


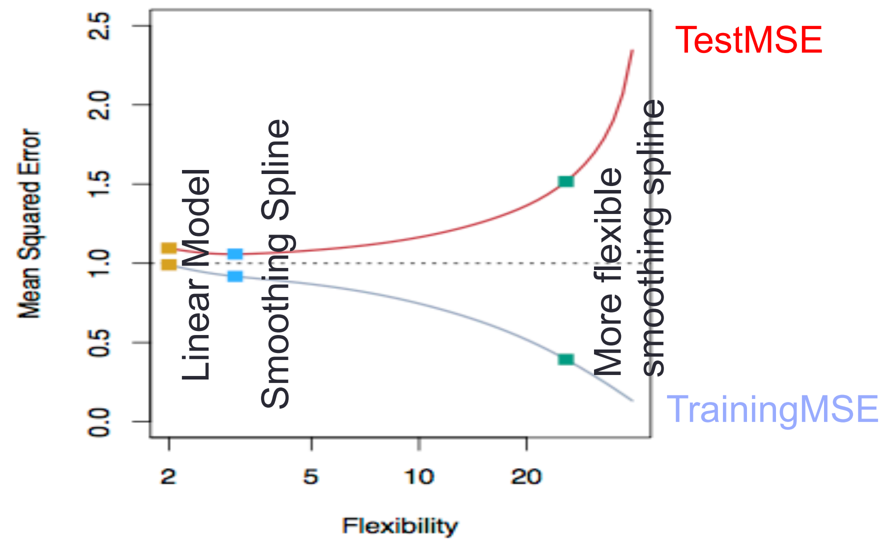
Black: Truth
Orange: Linear Estimate
Blue: smoothing spline
Green: smoothing spline (more flexible)

RED: Test MSE
Grey: Training MSE
Dashed: Minimum possible test MSE (irreducible error)

CONCEPT CHECK:
Where does "irreducible error" come from?

Examples with Different Levels of Flexibility: Example 2

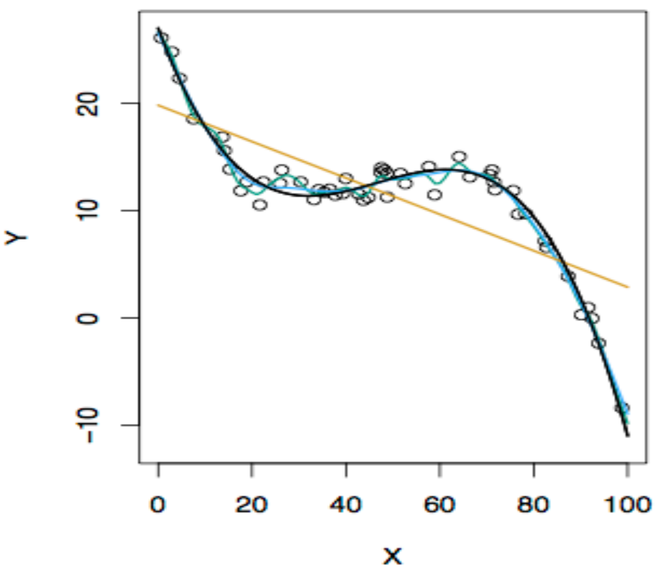


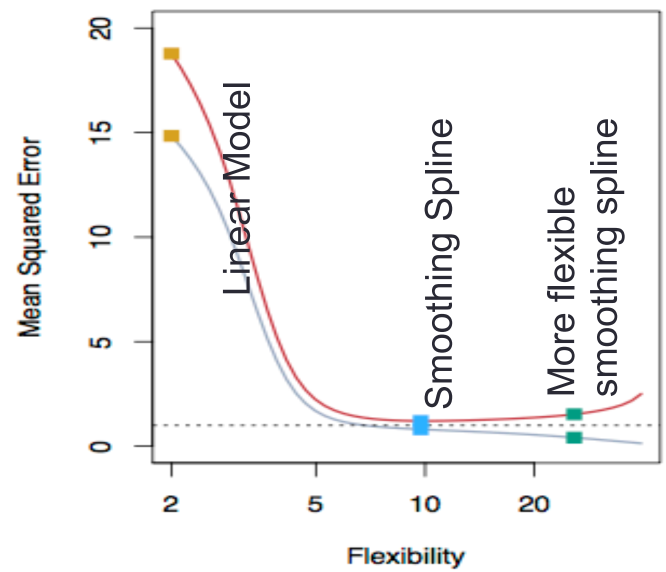


Black: Truth
Orange: Linear Estimate
Blue: smoothing spline
Green: smoothing spline (more flexible)

RED: Test MSE
Grey: Training MSE
Dashed: Minimum possible test MSE (irreducible error)

Examples with Different Levels of Flexibility: Example 3





TestMSE

TrainingMSE

Black: Truth
Orange: Linear Estimate
Blue: smoothing spline
Green: smoothing spline (more flexible)

RED: Test MSE
Grey: Training MSE
Dashed: Minimum possible test MSE (irreducible error)

Bias / Variance Tradeoff

- The previous graphs of **test** versus **training** MSE's illustrate a very important tradeoff that governs the choice of statistical learning methods.

- There are always two competing forces that govern the choice of learning method: *bias* and *variance*.

Bias of Learning Methods

- The *Inductive Bias* of a (machine) learning algorithm is the set of assumptions used to predict outputs given inputs it has not encountered (Tom Mitchell, 1980).
- Intuition- Higher Bias: quicker to generalize well… but more likely to overgeneralize
- In our text, **Bias** refers to the error that is introduced by modeling a real life phenomenon by a model that does not match the phenomenon.
- Often we are talking about a real world phenomenon that is complicated *being modeled by a much simpler model*.
- For example, linear regression assumes that there is a linear relationship between Y and X. It is unlikely that, in real life, the relationship is exactly linear so *some bias* will be induced when we select a linear model.
- The more flexible/complex a method is the less bias it will generally have.

Variance of Learning Methods

- Variance refers to the model's sensitivity to error caused by small fluctuations in the training data.
- Intuition – Variance tells us how sensitive the model is to having trained with a different set of training data from the same original phenomenon
- **Concept check: Why does choosing a higher variance model lead to a performance improvement on the training set but worse performance on the test set?**

Bias-Variance Trade-off

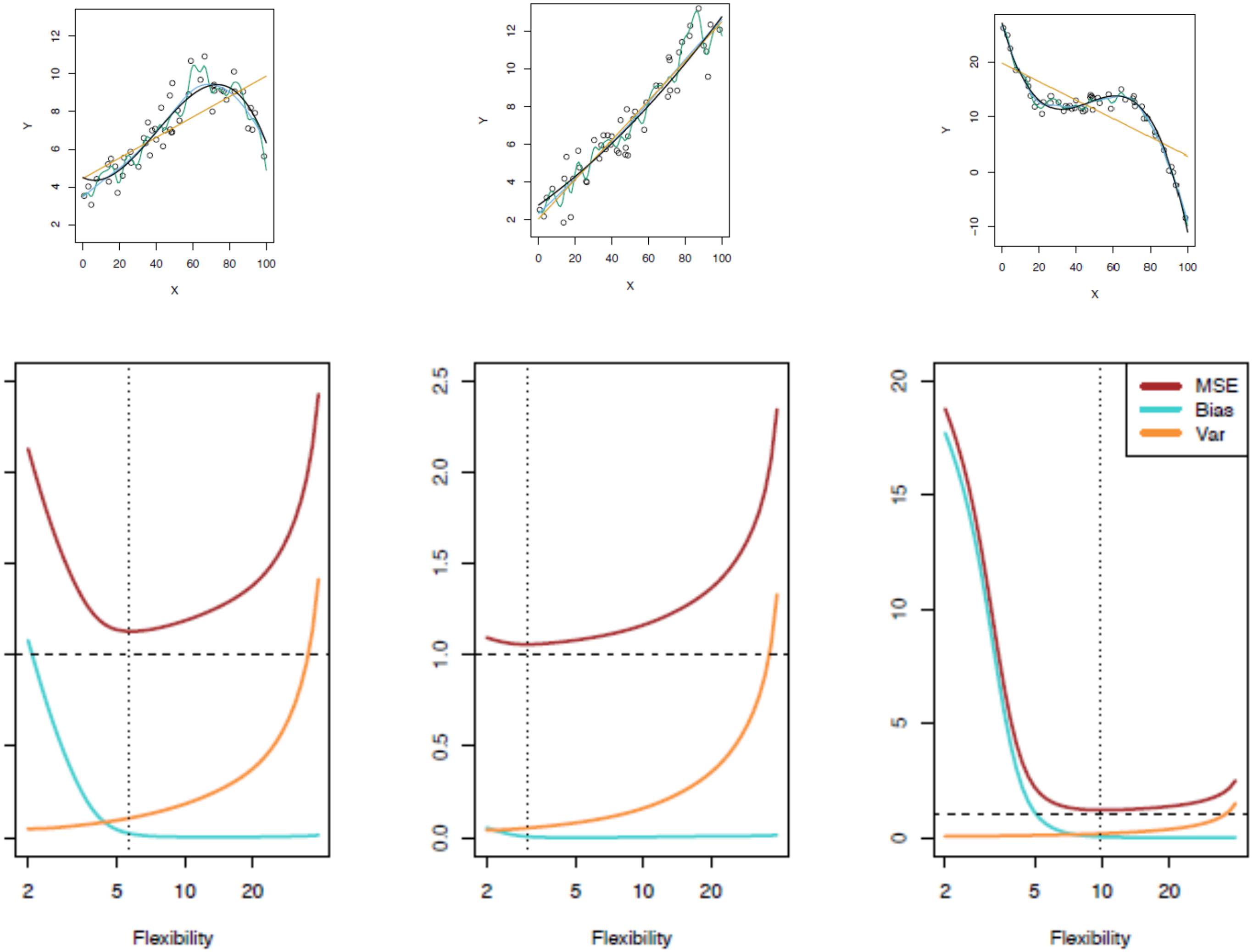
- It can be shown that for any given, X=x_0, the expected test MSE for a new Y at x_0 will be equal to

Irreducible error

$$ExpectedTestMSE = E(Y - f(x_0))^2 = Bias^2 + Var + \sigma^2$$

- … As a modeling method gets more flexible the bias will decrease and the variance will increase but expected test MSE may go up or down
- The mathematical details of the derivation are in the "Elements of Statistical Learning" book but are not covered in our course (https://web.stanford.edu/~hastie/ElemStatLearn/)

Test MSE, Bias and Variance



Assessing ***Classification*** Performance

- For a regression problem, we used the MSE to assess the accuracy of the statistical learning method
- For a classification problem we can use the error rate i.e.

$$Error\ Rate = \sum_{i=1}^{n} I(y_i \neq \hat{y}_i)/n$$

- $I(y_i \neq \hat{y}_i)$ is an *indicator function*, which will give 1 if the condition $(y_i \neq \hat{y}_i)$ is true, otherwise it gives a 0.

- Thus the error rate represents the fraction of incorrect classifications, or misclassifications

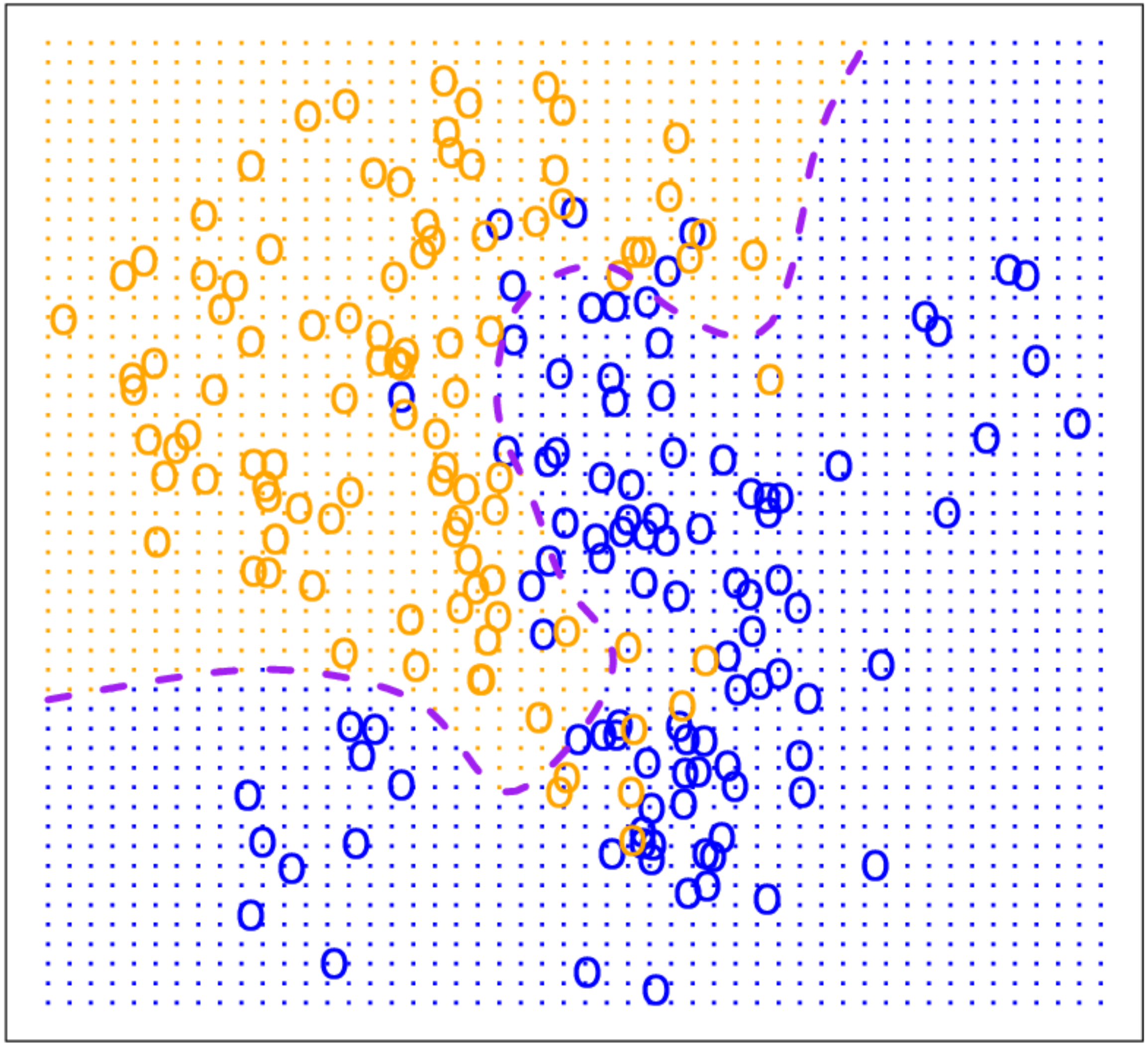
Bayes Error Rate

- The Bayes error rate refers to the lowest possible error rate that could be achieved if somehow we knew exactly what the "true" probability distribution of the data looked like.

- On test data, no classifier (or statistical learning method) can get lower error rates than the Bayes error rate.

- *In many real life problems the Bayes error rate can't be calculated exactly*. **Why not?**

Bayes Optimal Classifier



K-Nearest Neighbors (KNN)

- K-Nearest Neighbors is a flexible approach for classification
  - can be used to *estimate* the Bayes Classifier.

- For any given X_i we find the k closest neighbors to X_i in the training data, and examine their corresponding Y labels.

- The class of X_i is predicted to be the class of the majority (or plurality if more than 2 classes) of its neighbors

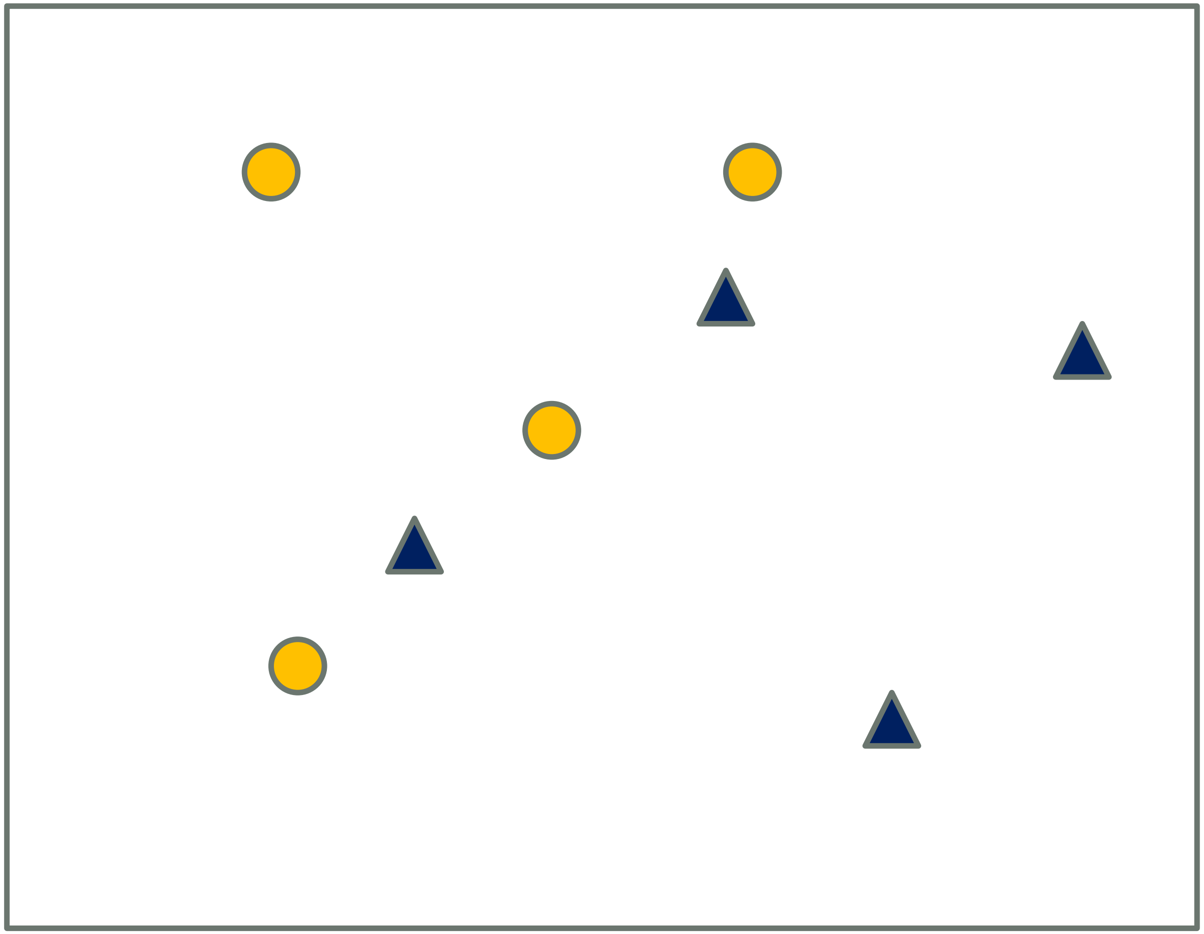
- The smaller that k is the more flexible the method will be.

KNN Worksheet

- Using your knowledge of test and training set performance, complete problem # 8 on the worksheet for Day 4

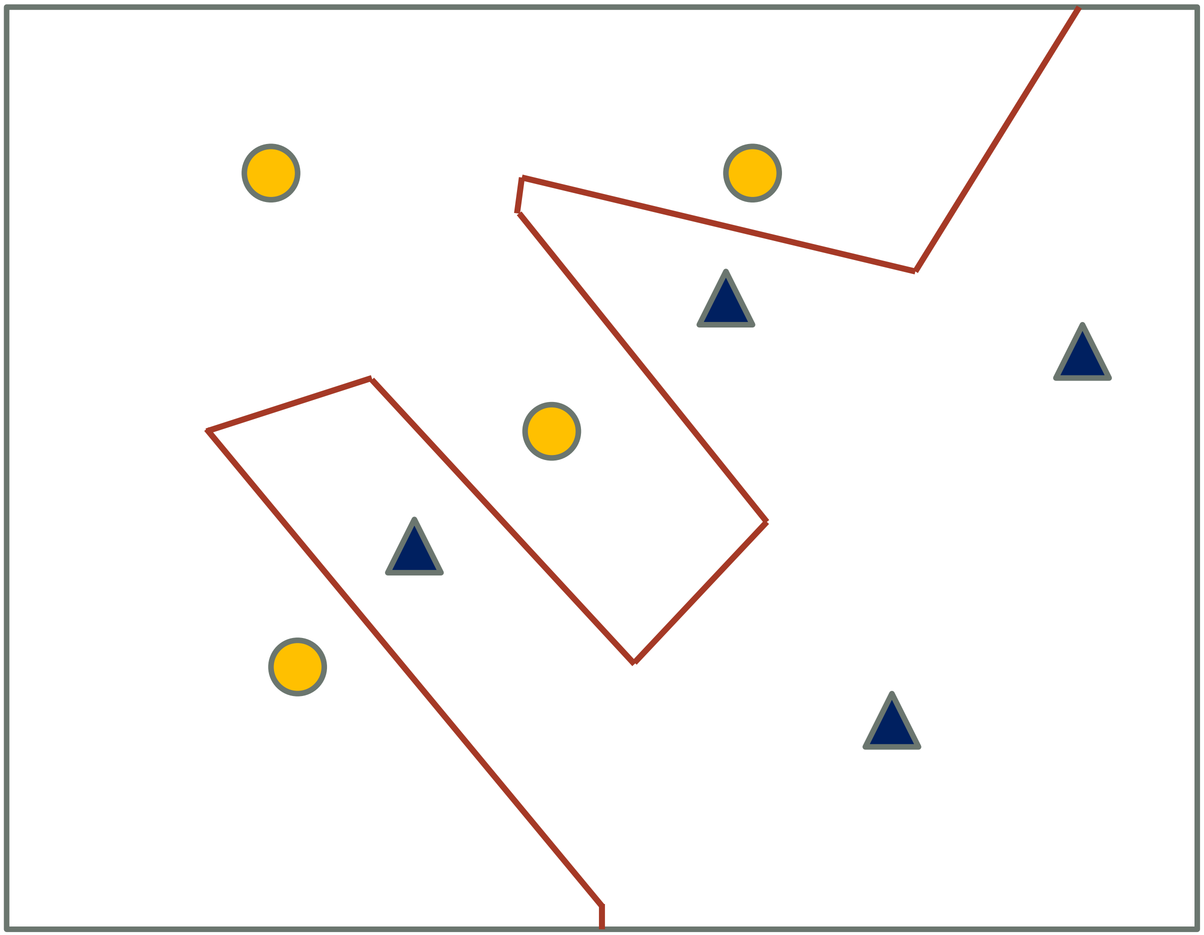
KNN visual – decision boundaries

- What do the decision boundaries look like when K=1?

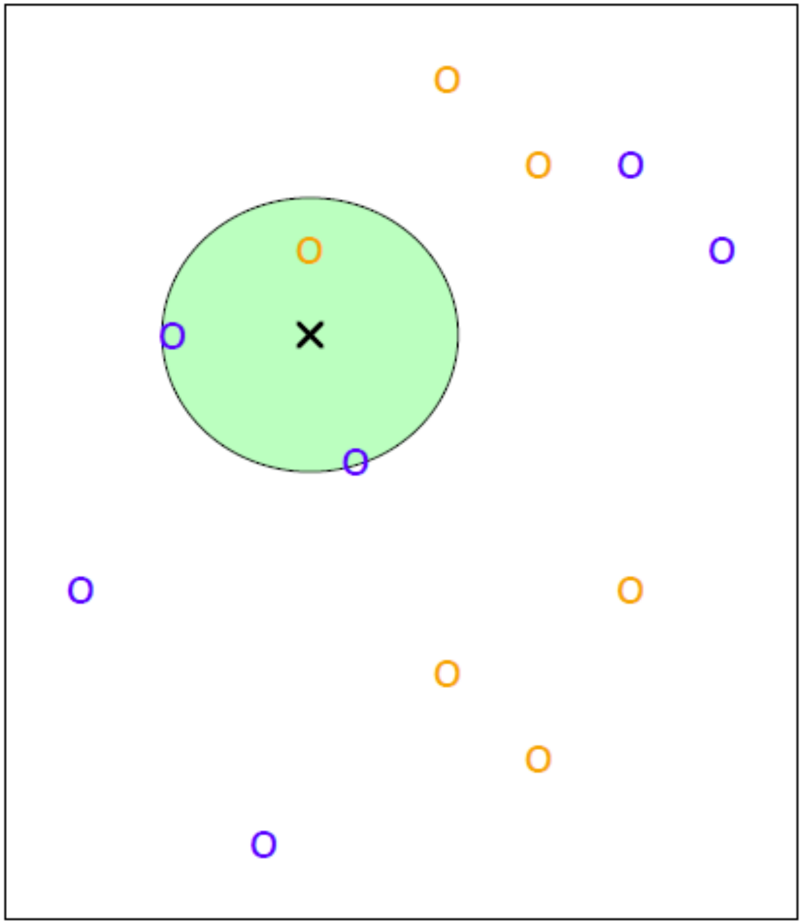


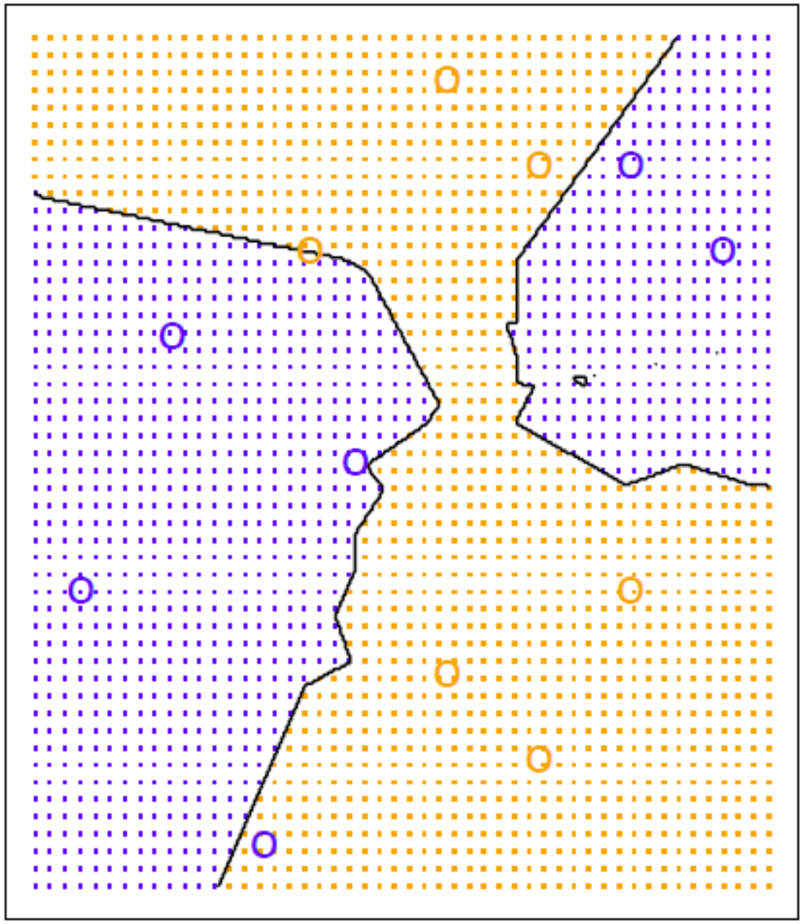
KNN Demo – decision boundaries

- What do the decision boundaries look like when K=1?

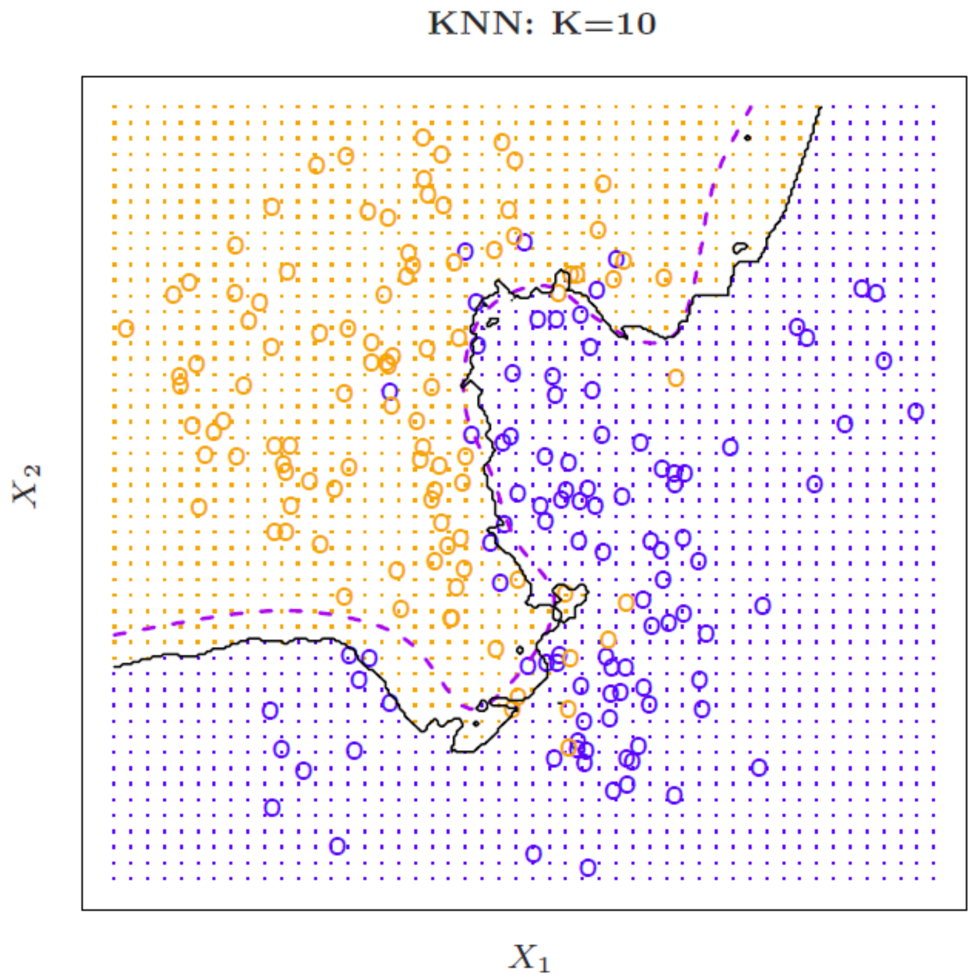


KNN Example with k = 3

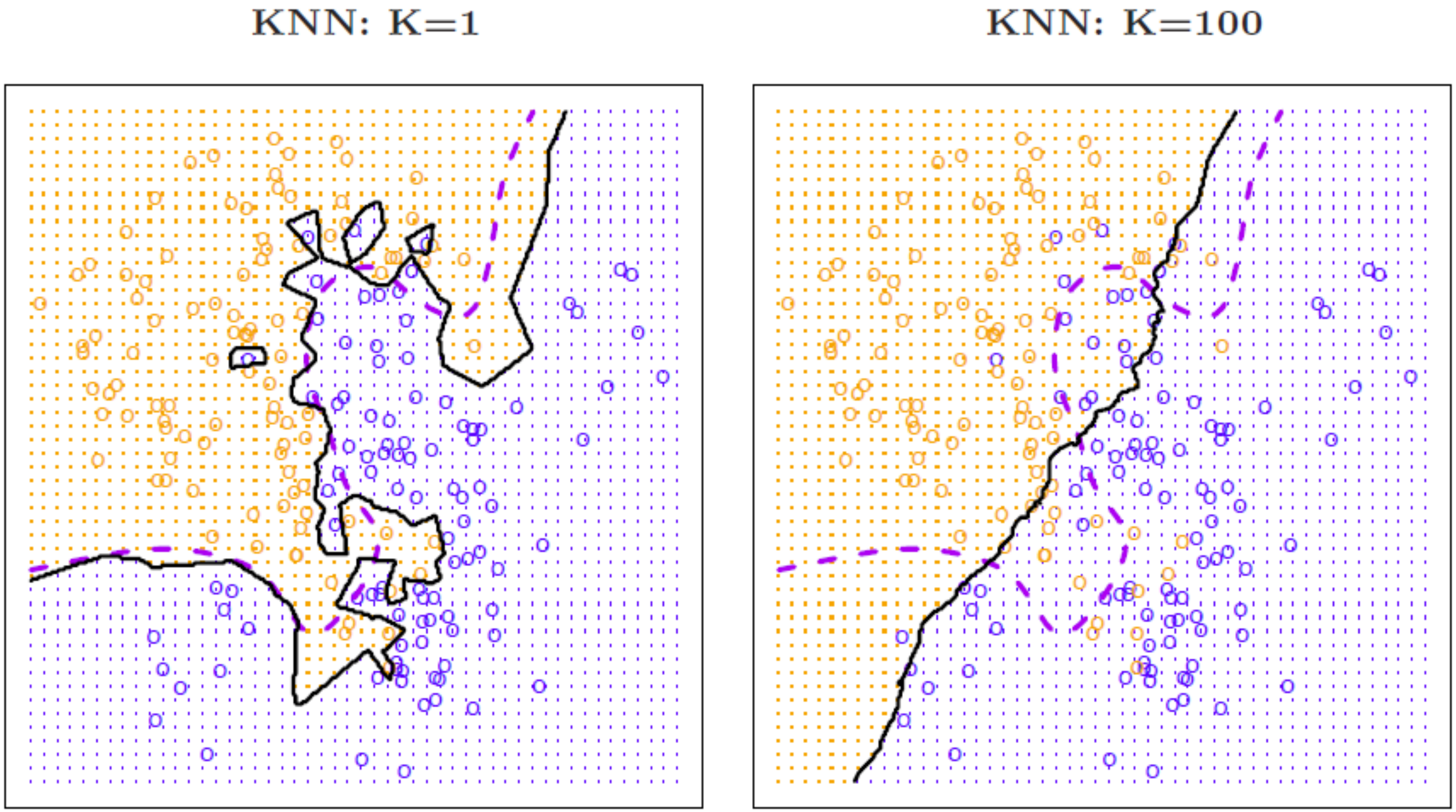




Simulated Data: K = 10



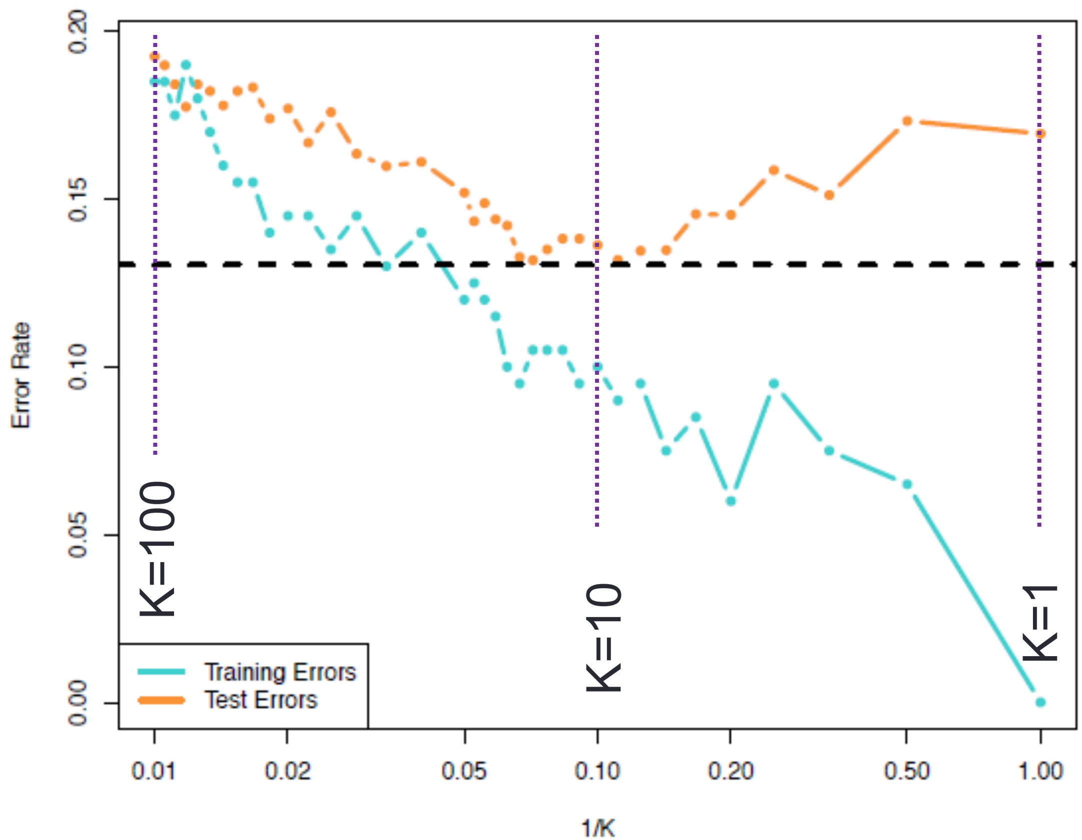
K = 1 and K = 100



Training vs. Test Error Rates on the Simulated Data

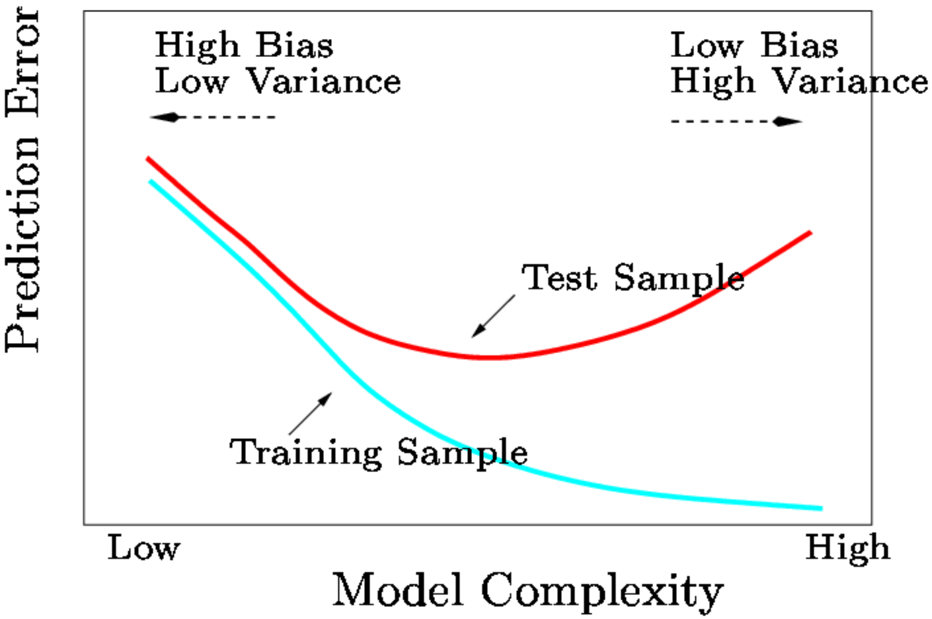
- Notice that training error rates keep going down as k decreases or equivalently as the flexibility increases.

- However, the test error rate at first decreases but then starts to increase again. **WHY?**



Model complexity & Performance

- As model complexity increases, ***training*** error declines*
- As complexity increases, ***test*** errors will decline at first (as reductions in bias dominate) but will then start to increase again (as increases in variance dominate)
- Where test error is minimized, the model has a good complexity



Find the model with the *right* complexity

More flexible/complicated is not always better