MSSC 6250 / Statistical Machine Learning

Instructor: Mehdi Maadooliat

What is Statistical Learning?

- Chapter 02 – Part I

Department of Mathematics, Statistics and Computer Science

OUTLINE

- What Is Statistical Learning?
  - Examples
  - Why estimate f?
  - How do we estimate f?
  - The trade-off between prediction accuracy and model interpretability
  - Supervised vs. unsupervised learning
  - Regression vs. classification problems

What is Statistical Learning?

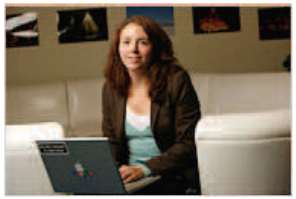
For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

SIGN IN TO RECOMMEND
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

Enlarge This Image



Thor Swift for The New York Times

Carrie Grimes, senior staff engineer at Google, uses statistical analysis of data to help improve the company's search engine.

"People think of field archaeology as Indiana Jones, but much of what you really do is data analysis," she said.

Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession as a place for dronish number nerds. They are finding themselves increasingly in demand — and even cool.

"I keep saying that the sexy job in the next 10 years will be statisticians," said Hal Varian, chief economist at Google. "And I'm not kidding."

Multimedia

Quote of the Day, New York Times, August 5, 2009

"I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding." — HAL VARIAN, chief economist at Google.

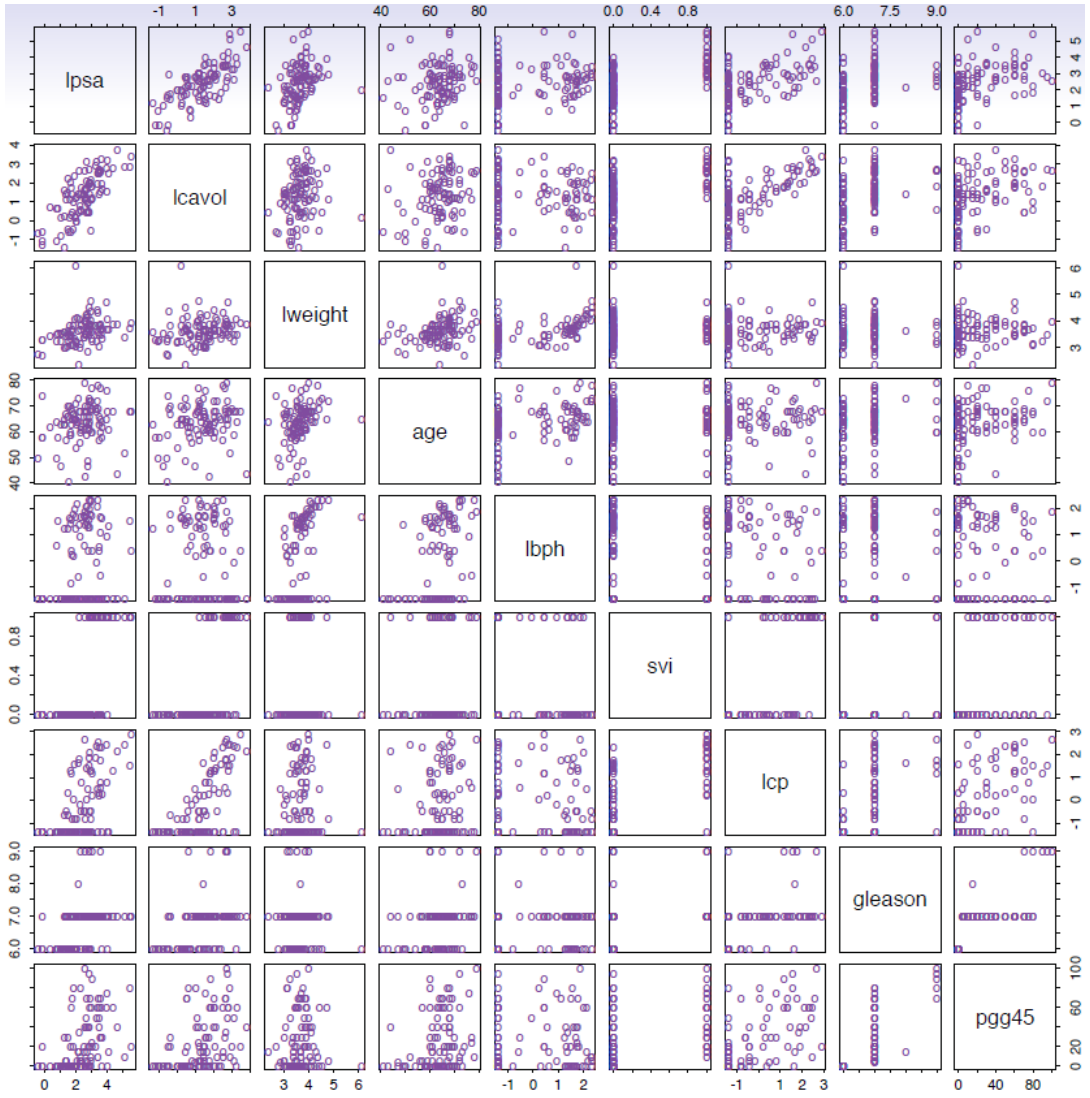
What is Statistical Learning? (Examples)

How IBM built Watson, its *Jeopardy*-playing supercomputer by Dawn Kawamoto DailyFinance 02/08/2011

Learning from its mistakes According to David Ferrucci (PI of Watson DeepQA technology for IBM Research), Watson's software is wired for more that handling natural language processing.

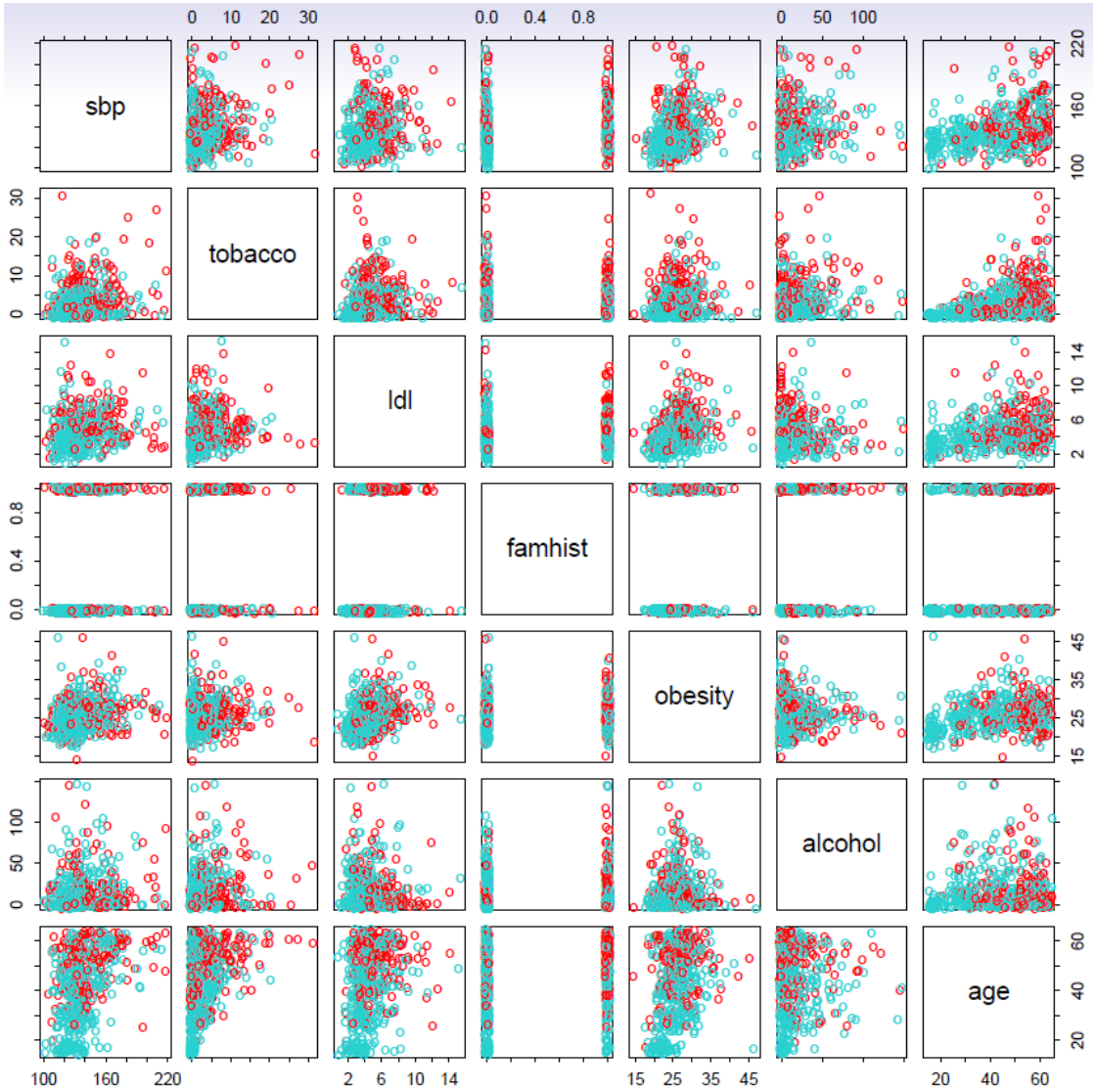
"It's machine learning allows the computer to become smarter as it tries to answer questions — and to learn as it gets them right or wrong."

IDENTIFY THE RISK FACTORS FOR PROSTATE CANCER.



MARQUETTE UNIVERSITY
Be The Difference.

PREDICT WHETHER SOMEONE WILL HAVE A HEART ATTACK ON THE BASIS OF DEMOGRAPHIC, DIET AND CLINICAL MEASUREMENTS.



MARQUETTE UNIVERSITY
Be The Difference.

CUSTOMIZE AN EMAIL SPAM DETECTION SYSTEM.

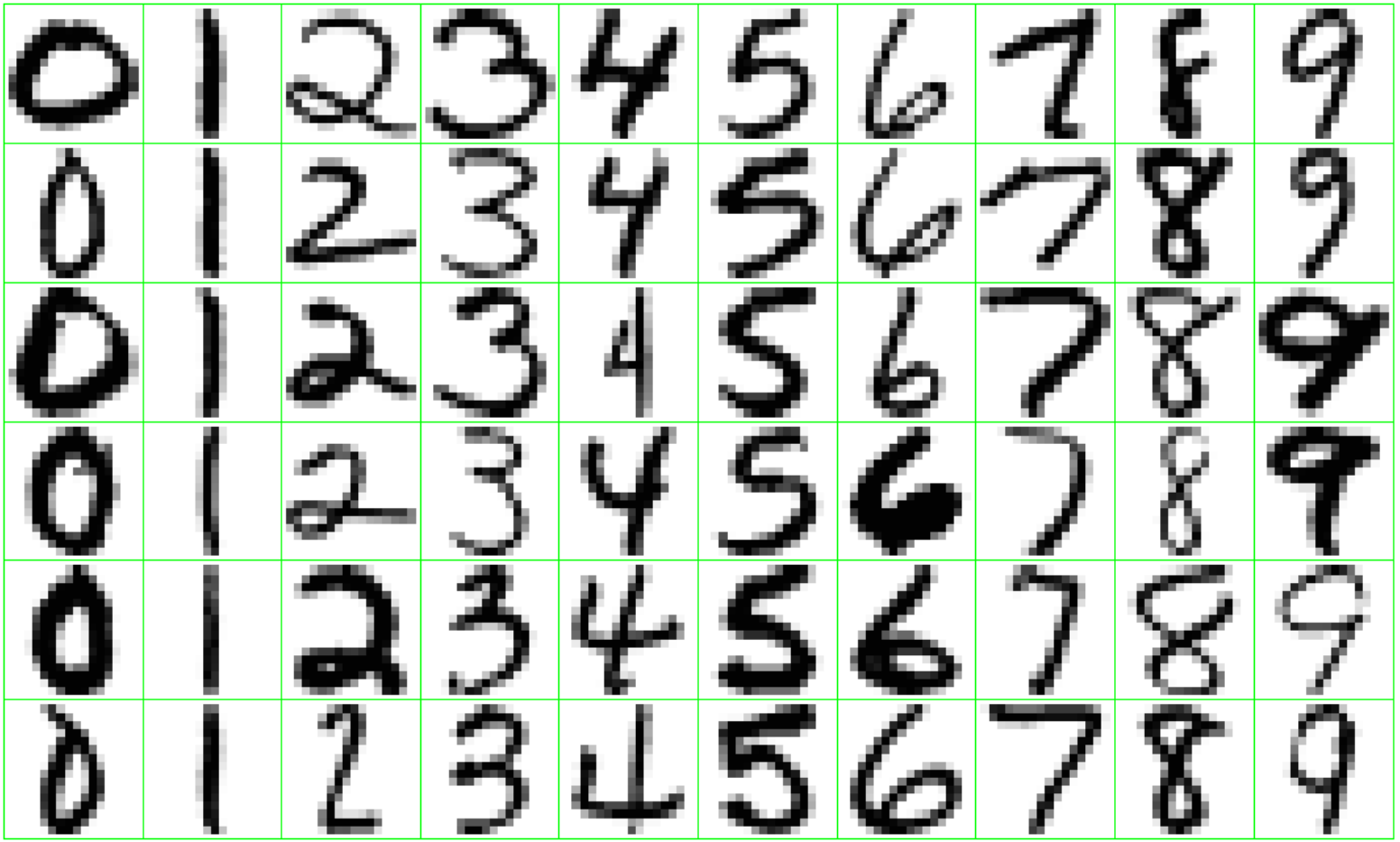
- data from 4601 emails sent to an individual (named George, at HP labs, before 2000). Each is labeled as *spam* or *email*.
- goal: build a customized spam filter.
- input features: relative frequencies of 57 of the most commonly occurring words and punctuation marks in these email messages.

	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

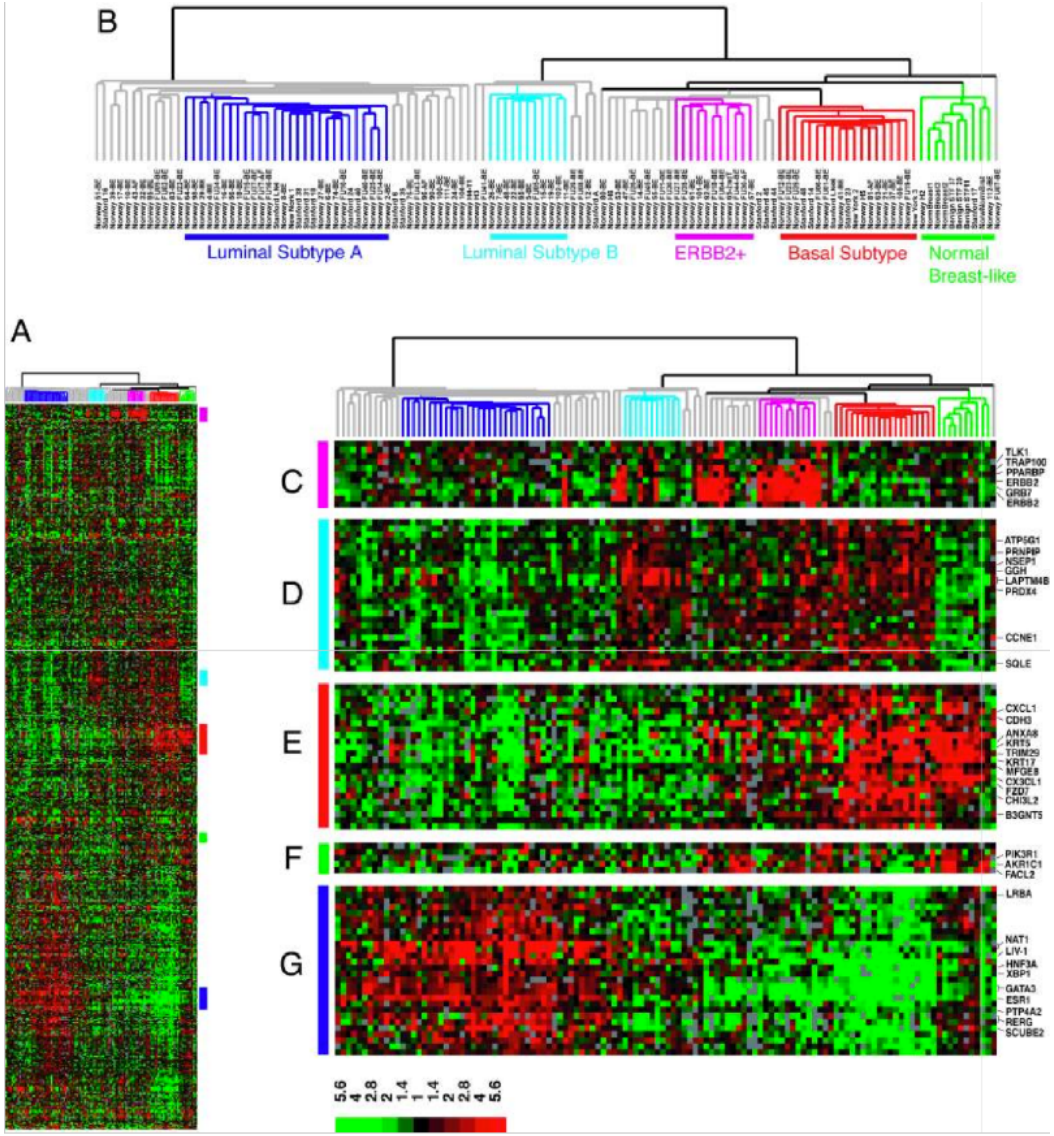
Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between spam *and* email.

IDENTIFY THE NUMBERS

IN A HANDWRITTEN ZIP CODE.

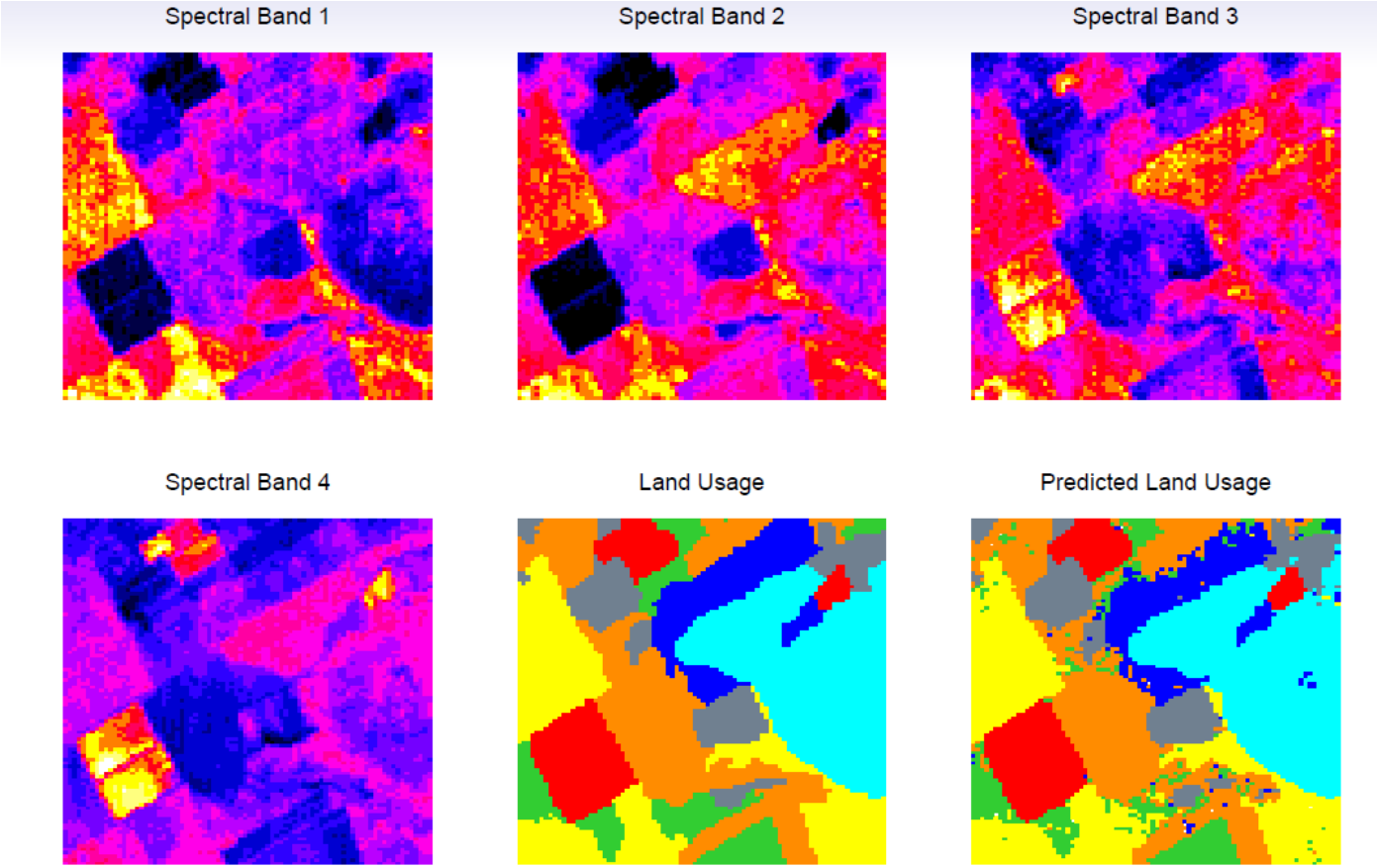


CLASSIFY A TISSUE SAMPLE INTO ONE OF SEVERAL CANCER CLASSES, BASED ON A GENE EXPRESSION PROFILE.



MARQUETTE UNIVERSITY
Be The Difference.

Classify the pixels in a LANDSAT image, by usage.



$Usage \in \{red\ soil,\ cotton,\ vegetation\ stubble,\ mixture,\ gray\ soil,\ damp\ gray\ soil\}$

What is Statistical Learning?

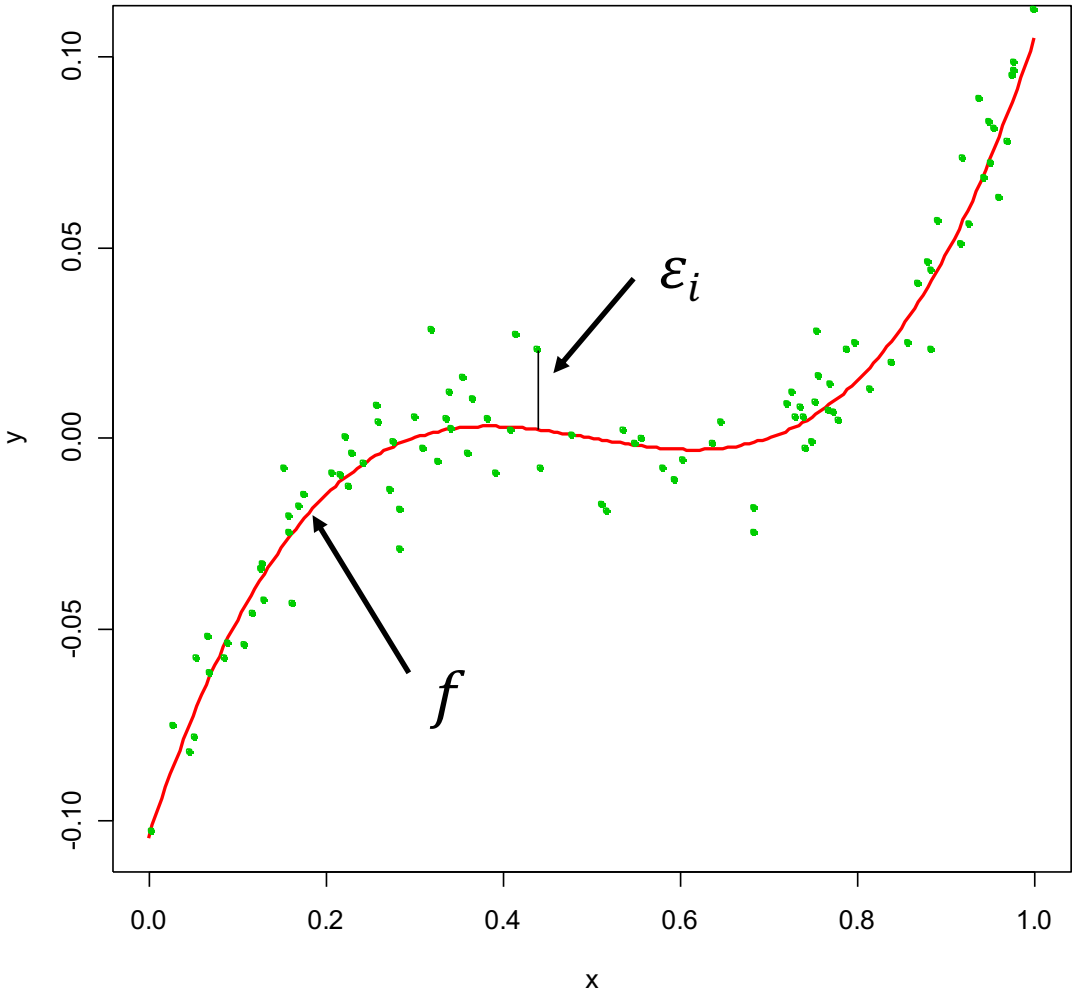
- Suppose we observe Y_i and $\boldsymbol{X}_i = \left(X_{i1}, \cdots, X_{ip}\right)$ for $i = 1, \cdots, n$
- We believe that there is a relationship between Y and at least one of the X's.

- We can model the relationship as

$$Y_i = f(\boldsymbol{X}_i) + \varepsilon_i$$

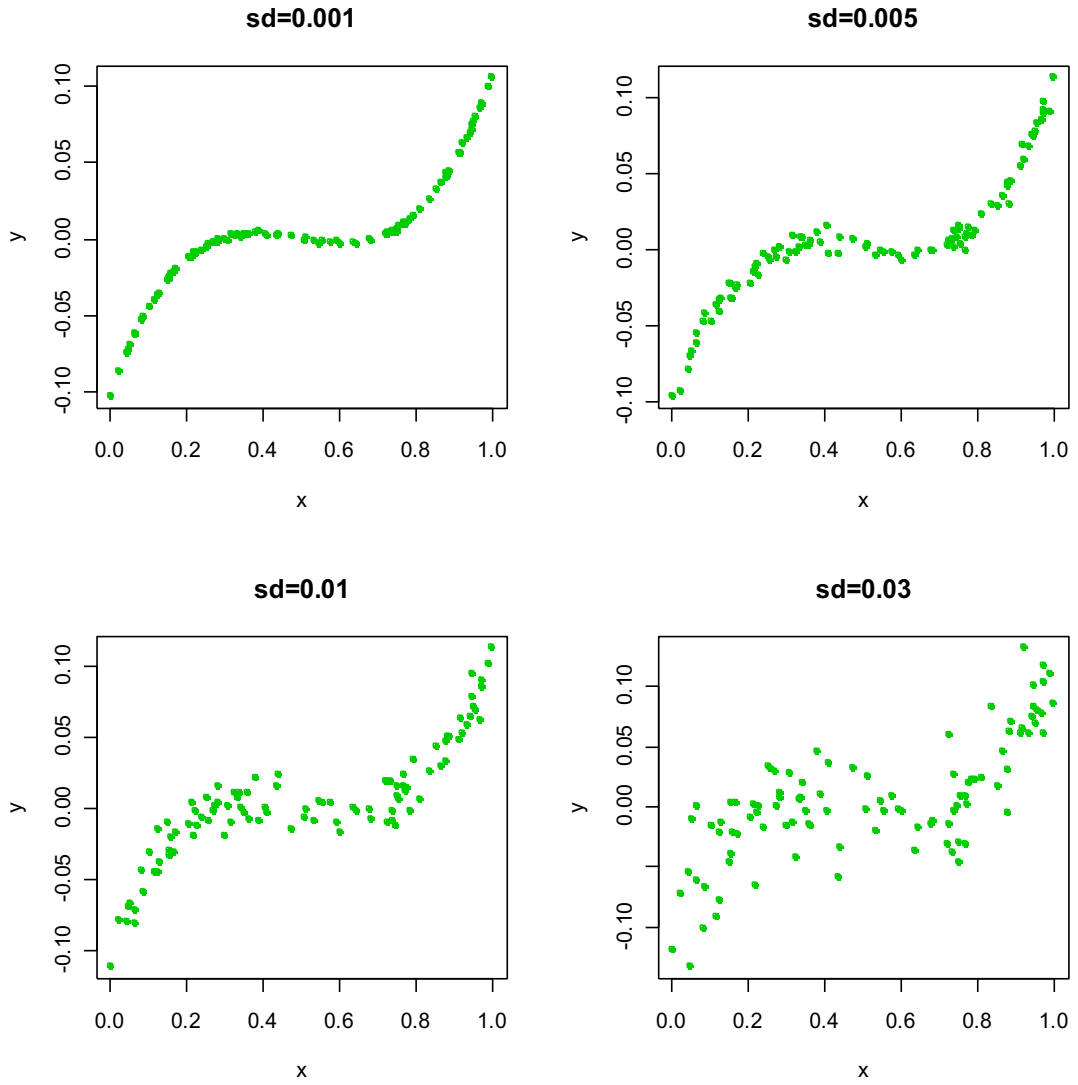
- Where f is an unknown function and ε is a random error with mean zero.

A Simple Example

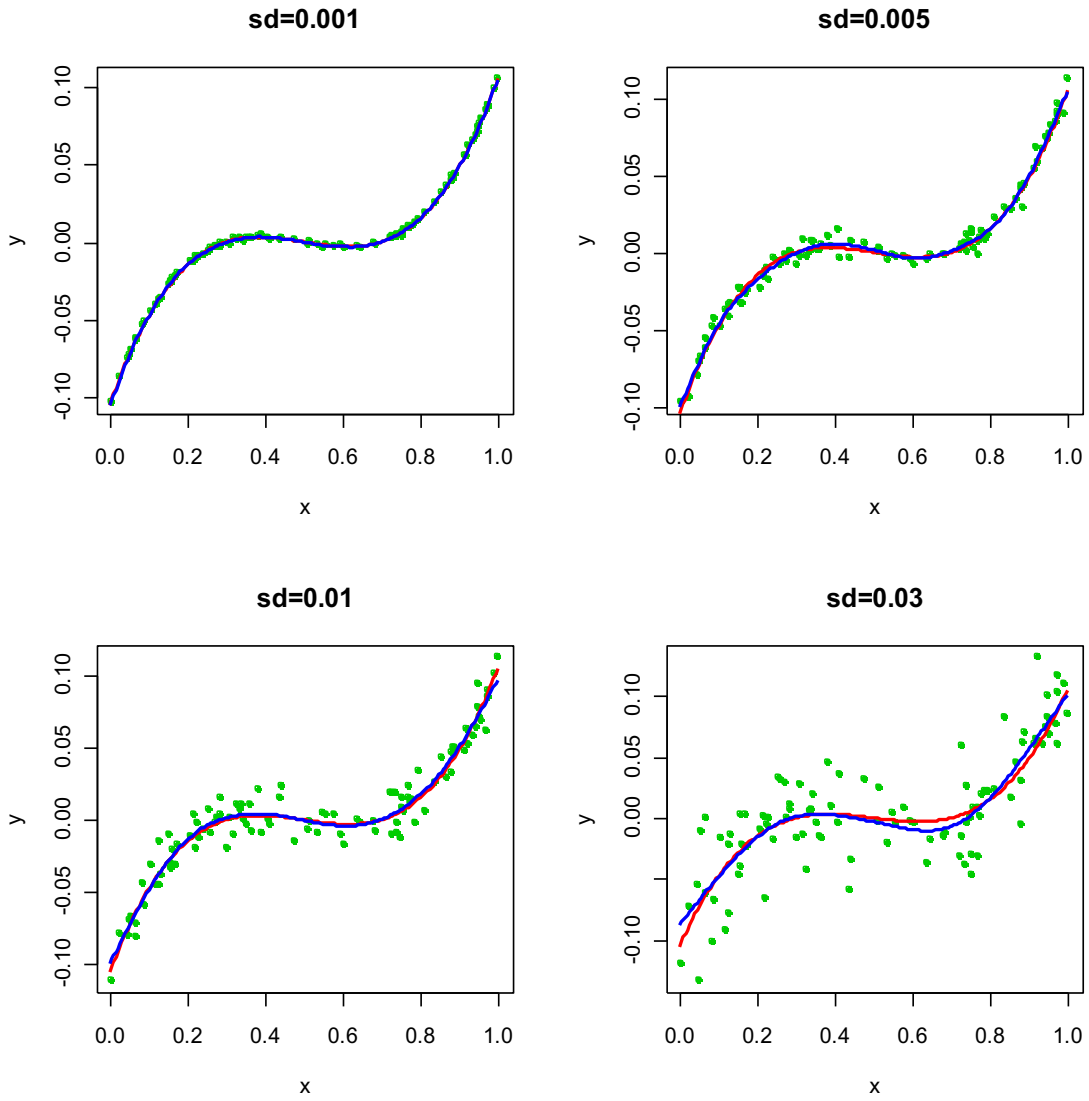


Different Standard Deviations

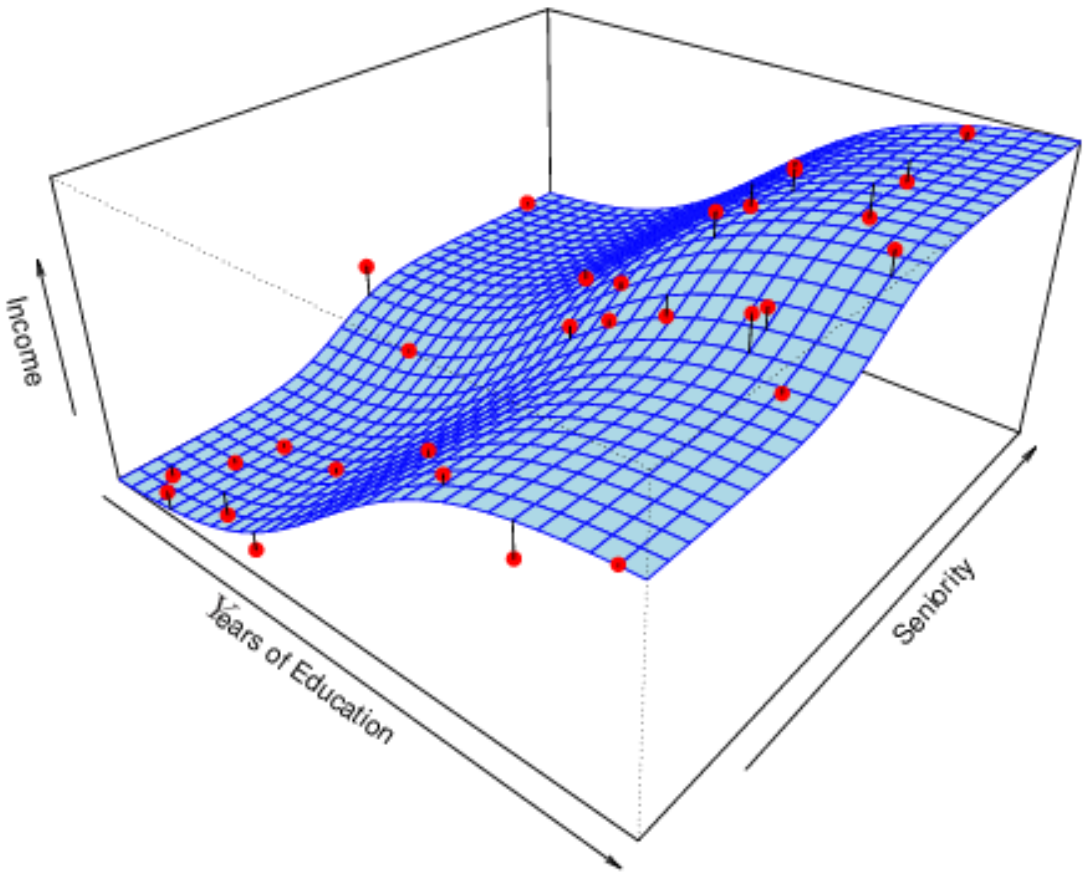
- The difficulty of estimating f will depend on the standard deviation of the ε's.



Different Estimates For f



INCOME VS. EDUCATION SENIORITY



Why Do We Estimate f?

- Statistical Learning, and this course, are all about how to estimate f.
- The term statistical learning refers to using the data to "learn" f.
- Why do we care about estimating f?
- There are 2 reasons for estimating f,
  - **Prediction** and
  - **Inference.**

1. PREDICTION

- If we can produce a good estimate for f (and the variance of ε is not too large) we can make accurate predictions for the response, Y, based on a new value of $\boldsymbol{X}$.

- Example: Direct Mailing Prediction
  - ✓ Interested in predicting how much money an individual will donate based on observations from 90,000 people on which we have recorded over 400 different characteristics.
  - ✓ Don't care too much about each individual characteristic.
  - ✓ Just want to know: For a given individual should I send out a mailing?

2. INFERENCE

- Alternatively, we may also be interested in the type of relationship between Y and the X's.
- For example,
  - Which particular predictors actually affect the response?
  - Is the relationship positive or negative?
  - Is the relationship a simple linear one or is it more complicated etc.?

- Example: Housing Inference
  - ✓ Wish to predict median house price based on 14 variables.
  - ✓ Probably want to understand which factors have the biggest effect on the response and how big the effect is.
  - ✓ For example how much impact does a river view have on the house value etc.

How Do We Estimate f?

- We will assume we have observed a set of **training data**

$$\{(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), \cdots, (\boldsymbol{X}_n, Y_n)\}$$

- We must then use the training data and a statistical method to estimate f.

- Statistical Learning Methods:
  - Parametric Methods
  - Non-parametric Methods

PARAMETRIC METHODS

- It reduces the problem of estimating f down to one of estimating a set of parameters.
- They involve a two-step model based approach

- STEP 1:

  Make some assumption about the functional form of f, i.e. come up with a model. The most common example is a linear model i.e.

$$f(\boldsymbol{X}_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

  However, in this course we will examine far more complicated, and flexible, models for f. In a sense the more flexible the model the more realistic it is.

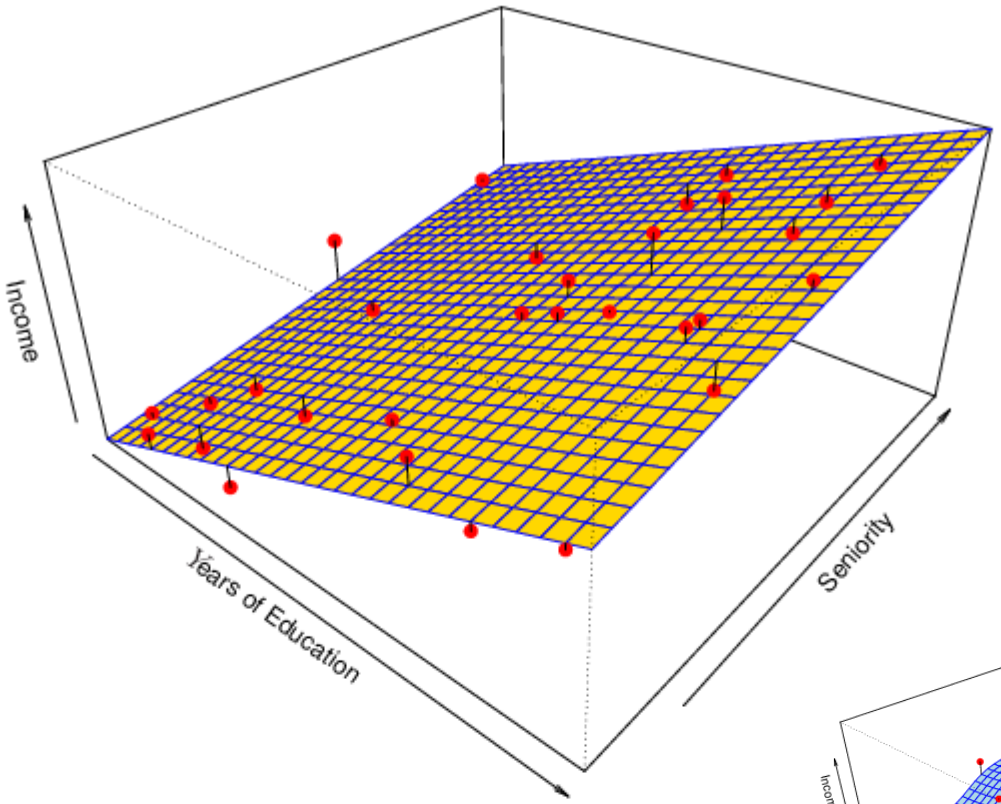
- STEP 2:

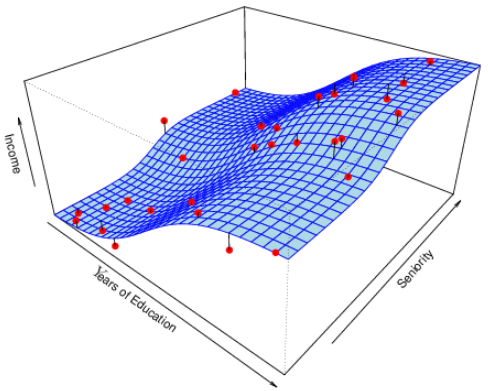
  Use the training data to fit the model i.e. estimate f or equivalently the unknown parameters such as $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

  - The most common approach for estimating the parameters in a linear model is ordinary least squares (OLS).
  - However, this is only one way.
  - We will see in the course that there are often superior approaches.

Example: A Linear Regression Estimate

- **Even if the standard deviation is low we will still get a bad answer if we use the wrong model.**





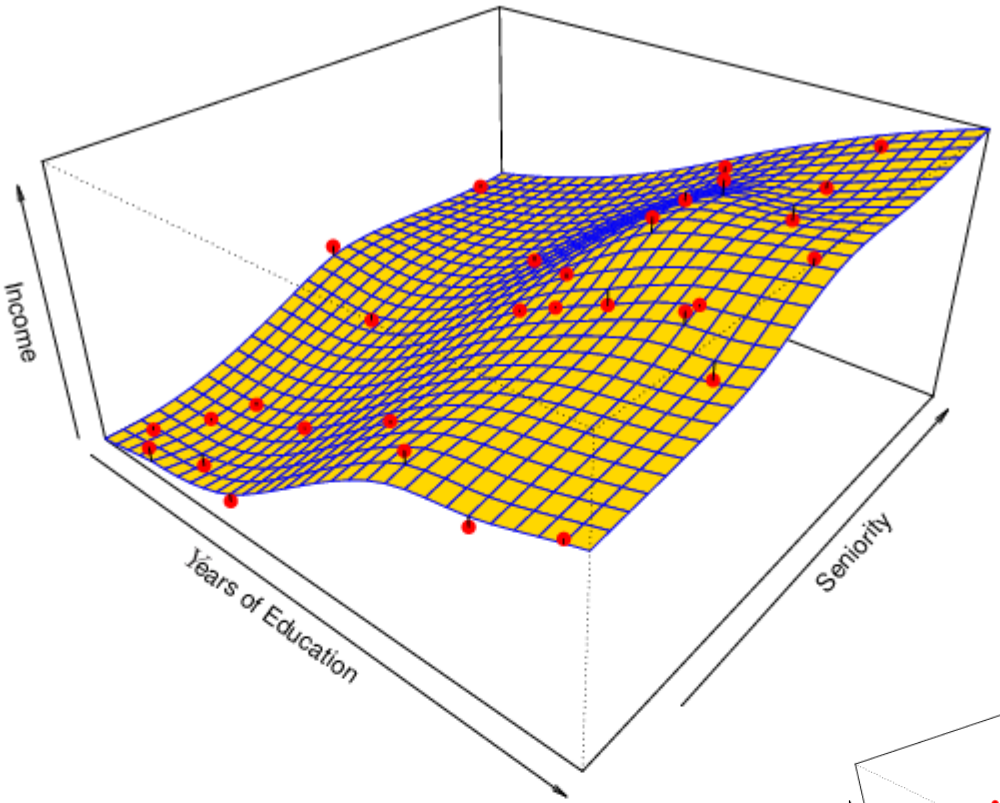
$$f = \beta_0 + \beta_1 \times Education + \beta_2 \times Seniority$$

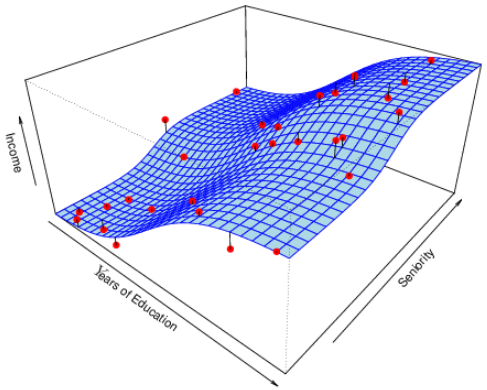
Non-parametric Methods

- They do not make explicit assumptions about the functional form of f.
- Advantages: They accurately fit a wider range of possible shapes of f.
- Disadvantages: A very large number of observations is required to obtain an accurate estimate of f

Example: A Thin-Plate Spline Estimate

- **Non-linear regression methods are more flexible and can potentially provide more accurate estimates.**





http://sctc.mscs.mu.edu:3838/sample-apps/FDA/

Tradeoff Between Prediction Accuracy and Model Interpretability

- Why not just use a more flexible method if it is more realistic?

Reason 1:

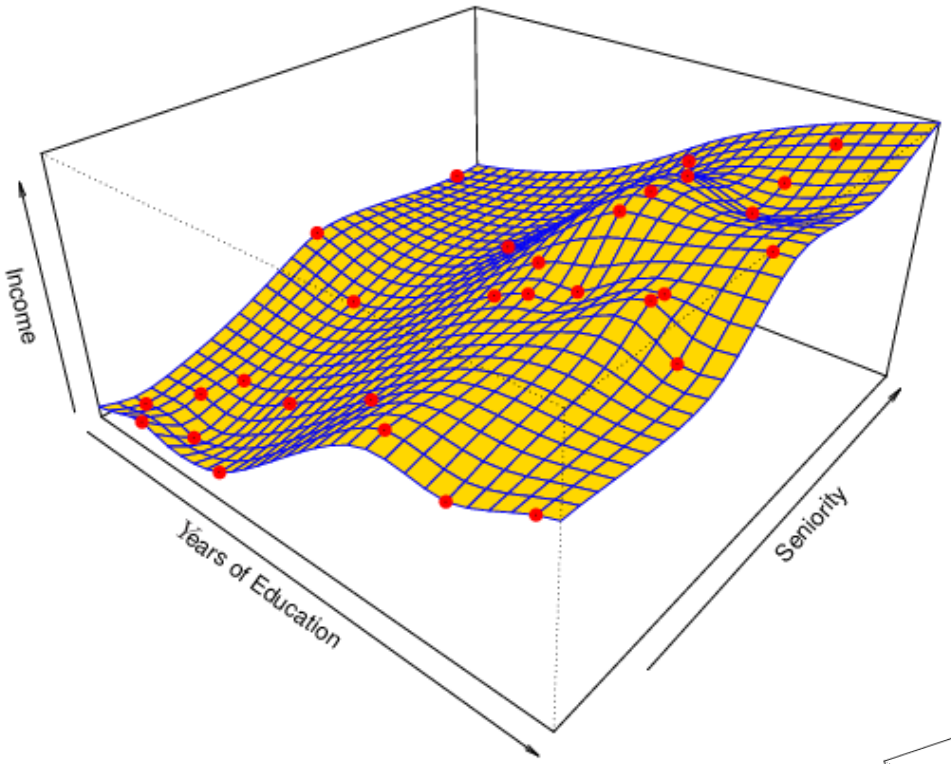
- A simple method such as linear regression produces a model which is much easier to interpret (the Inference part is better). For example, in a linear model, β_j is the average increase in Y for a one unit increase in X_j holding all other variables constant.

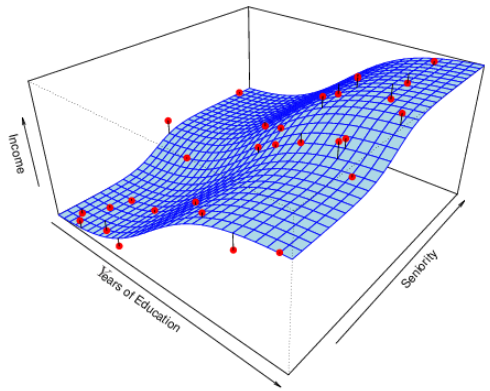
Reason 2:

- Even if you are only interested in prediction, so the first reason is not relevant, it is often possible to get more accurate predictions with a simple, instead of a complicated, model. This seems counter intuitive but has to do with the fact that it is harder to fit a more flexible model.

A Poor Estimate

- **Non-linear regression methods can also be too flexible and produce poor estimates for *f*.**

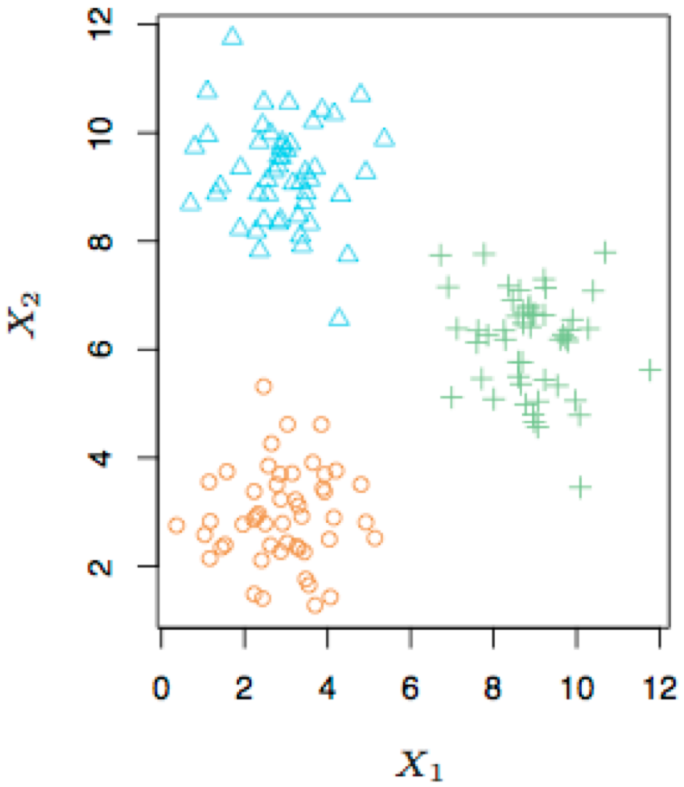


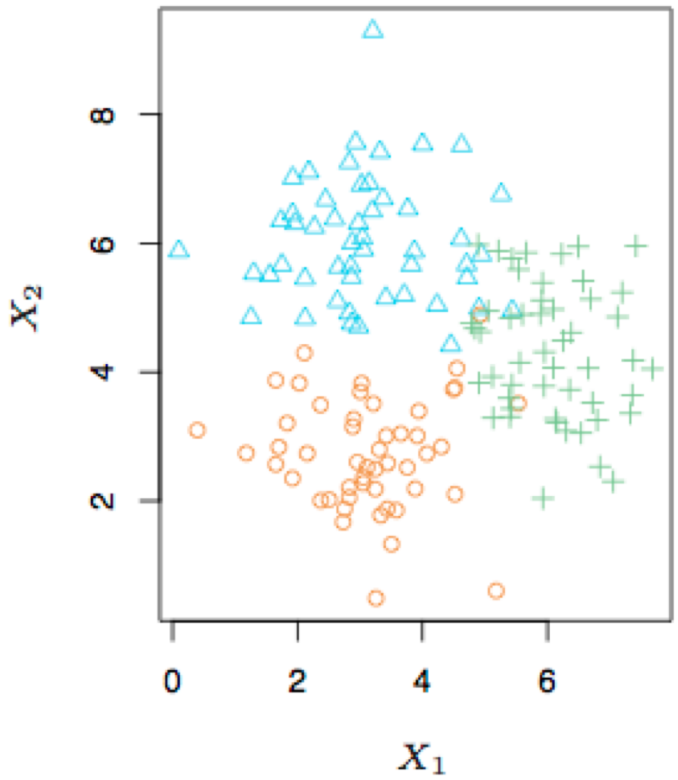


Supervised vs. Unsupervised Learning

- We can divide all learning problems into Supervised and Unsupervised situations
- Supervised Learning:
  - Supervised Learning is where both the predictors, $\boldsymbol{X}_i$, and the response, Y_i, are observed.
  - This is the situation you deal with in Linear Regression classes (e.g. MSCS 5780).
  - Most of this course will also deal with supervised learning.
- Unsupervised Learning:
  - In this situation only the $\boldsymbol{X}_i$'s are observed.
  - We need to use the $\boldsymbol{X}_i$'s to guess what Y would have been and build a model from there.
  - A common example is market segmentation where we try to divide potential customers into groups based on their characteristics.
  - A common approach is clustering.
  - We will consider unsupervised learning at the end of this course.

A Simple Clustering Example





Regression vs. Classification

- Supervised learning problems can be further divided into regression and classification problems.

- Regression covers situations where Y is continuous/numerical. e.g.
  - Predicting the value of the Dow in 6 months.
  - Predicting the value of a given house based on various inputs.

- Classification covers situations where Y is categorical e.g.
  - Will the Dow be up (U) or down (D) in 6 months?
  - Is this email a SPAM or not?

Different Approaches

- We will deal with both types of problems in this course.
- Some methods work well on both types of problem e.g. Neural Networks
- Other methods work best on Regression, e.g. Linear Regression, or on Classification, e.g. k-Nearest Neighbors.

MSSC 6250 / Statistical Machine Learning

Instructor: Mehdi Maadooliat

ASSESSING MODEL ACCURACY

- Chapter 02 – Part II

Department of Mathematics, Statistics and Computer Science

OUTLINE

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

Measuring Quality of Fit

- Suppose we have a regression problem.
- One common measure of accuracy is the mean squared error (MSE) i.e.

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

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

A Problem

- In either case our method has generally been designed to make MSE small on the training data we are looking at e.g. with linear regression we choose the line such that MSE is minimized.

- What we really care about is how well the method works on new data. We call this new data "**Test Data**".

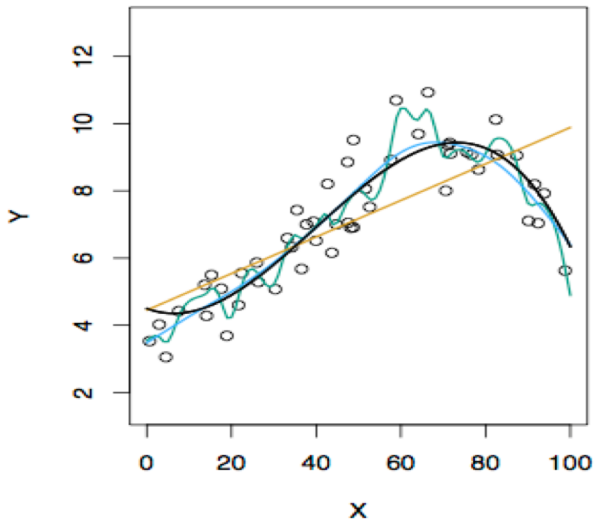
- There is no guarantee that the method with the smallest training MSE will have the smallest test (i.e. new data) MSE.

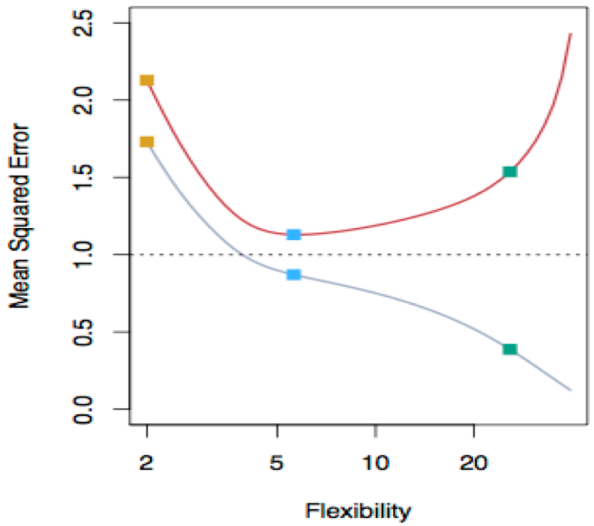
Training vs. Test MSE's

- In general the more flexible a method is the lower its training MSE will be i.e. it will "fit" or explain the training data very well.
  - Side Note: More Flexible methods (such as splines) can generate a wider range of possible shapes to estimate f as compared to less flexible and more restrictive methods (such as linear regression). The less flexible the method, the easier to interpret the model. Thus, there is a trade-off between flexibility and model interpretability.

- However, the test MSE may in fact be higher for a more flexible method than for a simple approach like linear regression.

Examples with Different Levels of Flexibility: Example 1





LEFT

Black: Truth

Orange: Linear Estimate

Blue: smoothing spline

Green: smoothing spline (more flexible)

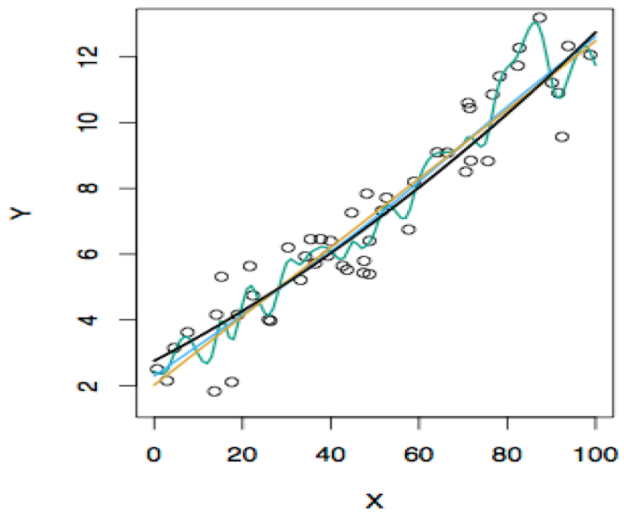
RIGHT

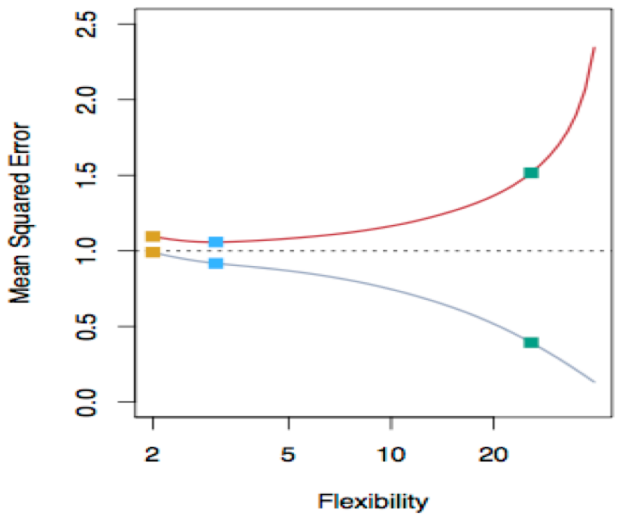
RED: Test MES

Grey: Training MSE

Dashed: Minimum possible test MSE (irreducible error)

Examples with Different Levels of Flexibility: Example 2





LEFT

Black: Truth

Orange: Linear Estimate

Blue: smoothing spline

Green: smoothing spline (more flexible)

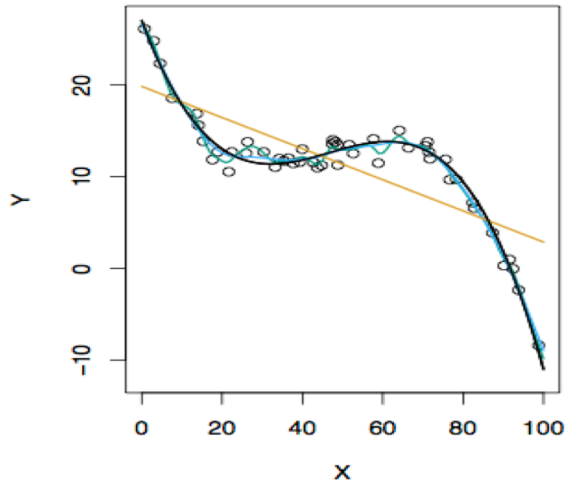
RIGHT

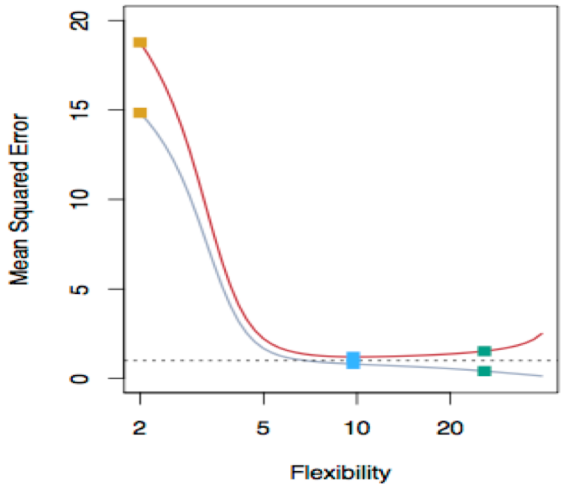
RED: Test MES

Grey: Training MSE

Dashed: Minimum possible test MSE (irreducible error)

Examples with Different Levels of Flexibility: Example 3





LEFT

Black: Truth

Orange: Linear Estimate

Blue: smoothing spline

Green: smoothing spline (more flexible)

RIGHT

RED: Test MES

Grey: Training MSE

Dashed: Minimum possible test MSE (irreducible error)

Bias/ Variance Tradeoff

- The previous graphs of test versus training MSE's illustrates 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 i.e. bias and variance.

Bias of Learning Methods

- Bias refers to the error that is introduced by modeling a real life problem (that is usually extremely complicated) by a much simpler problem.

- 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 present.

- The more flexible/complex a method is the less bias it will generally have.

Variance of Learning Methods

- Variance refers to how much your estimate for f would change by if you had a different training data set.

- Generally, the more flexible a method is the more variance it has.

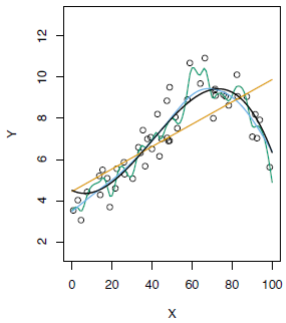
THE TRADE-OFF

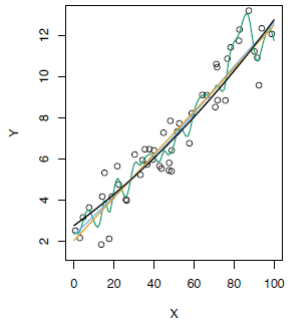
- It can be shown that for any given, $X=x_0$, the expected test MSE for a new Y at x_0 will be equal to

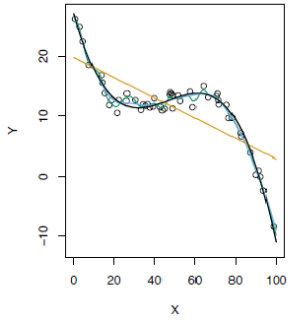
$$\underbrace{E\left(y_0 - \hat{f}(x_0)\right)^2}_{\text{Expected Test } MSE} = \underbrace{\text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2}_{\text{reducible error}} + \underbrace{\text{Var}(\epsilon)}_{\text{irreducible error}}.$$

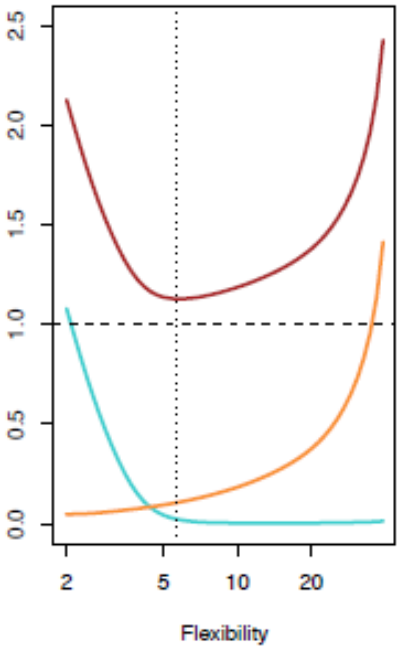
- What this means is that as a method gets more complex the bias will decrease and the variance will increase but expected test MSE may go up or down!

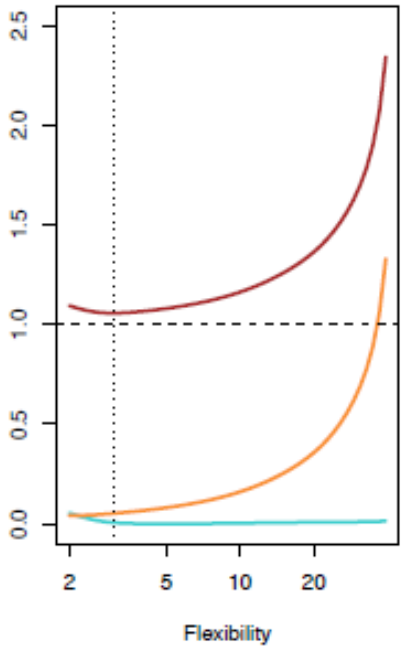
TEST MSE, BIAS AND VARIANCE

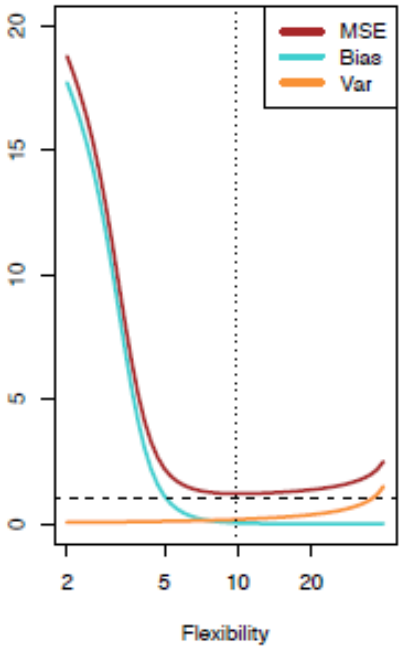












The Classification Setting

- 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.

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

- $I(y_i \neq \hat{y}_i)$ is an indicator function, which will give 1 if the condition $(y_i \neq \hat{y}_i)$ is correct, otherwise it gives a 0.
- Thus the error rate represents the fraction of incorrect classifications, or misclassifications

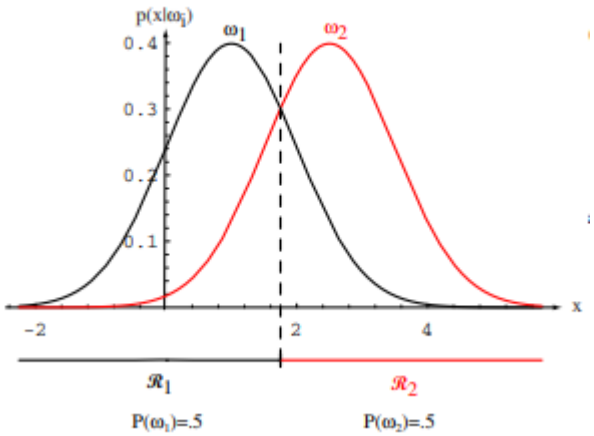
Bayes Error Rate

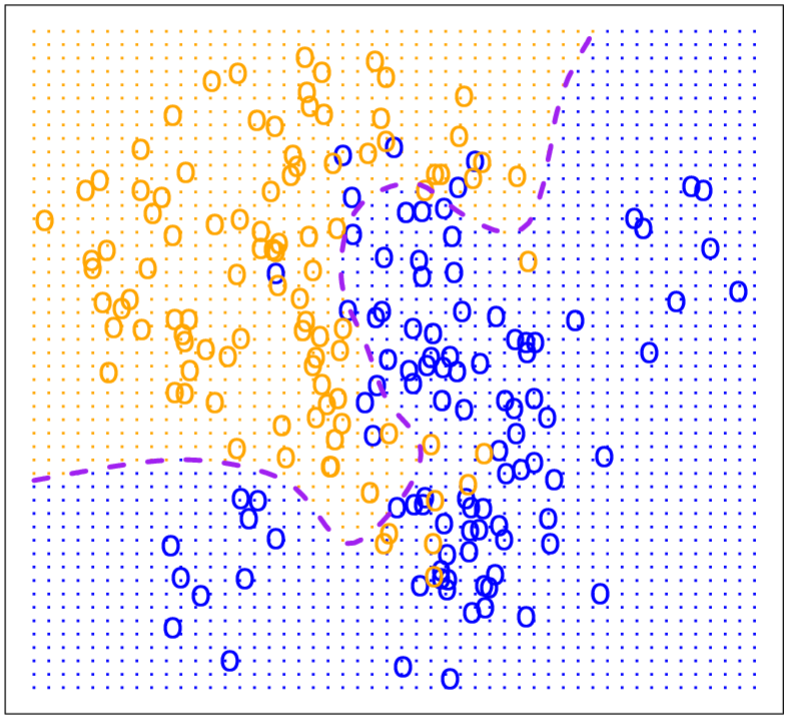
- The Bayes error rate refers to the lowest possible error rate that could be achieved if somehow we knew exactly what the "true" probability distribution of the data looked like.

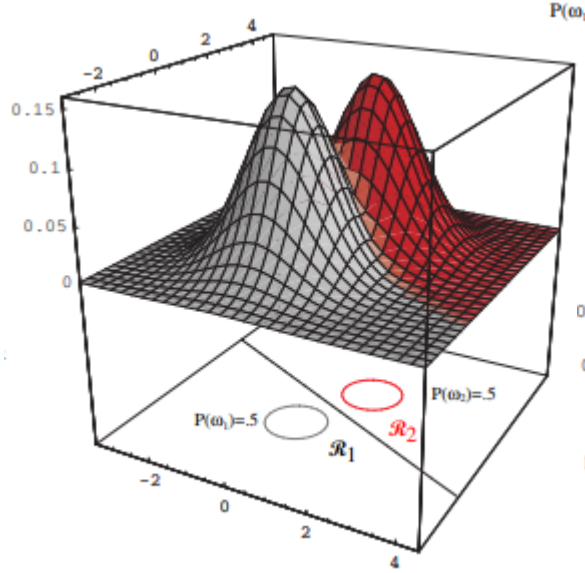
- On test data, no classifier (or stat. learning method) can get lower error rates than the Bayes error rate.

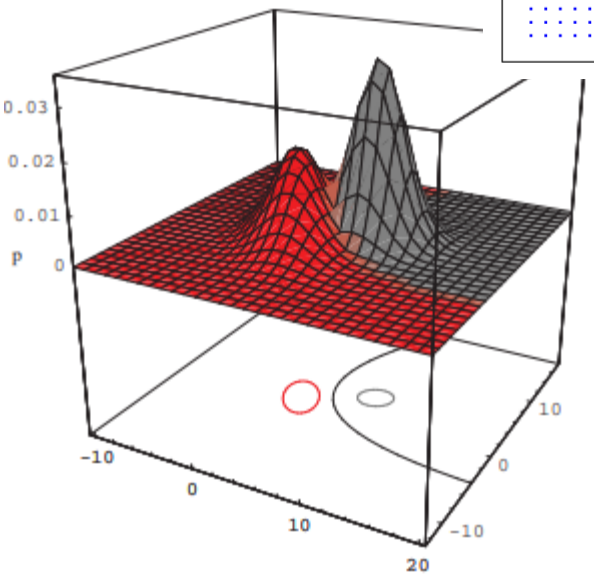
- Of course in real life problems the Bayes error rate can't be calculated exactly.

BAYES OPTIMAL CLASSIFIER







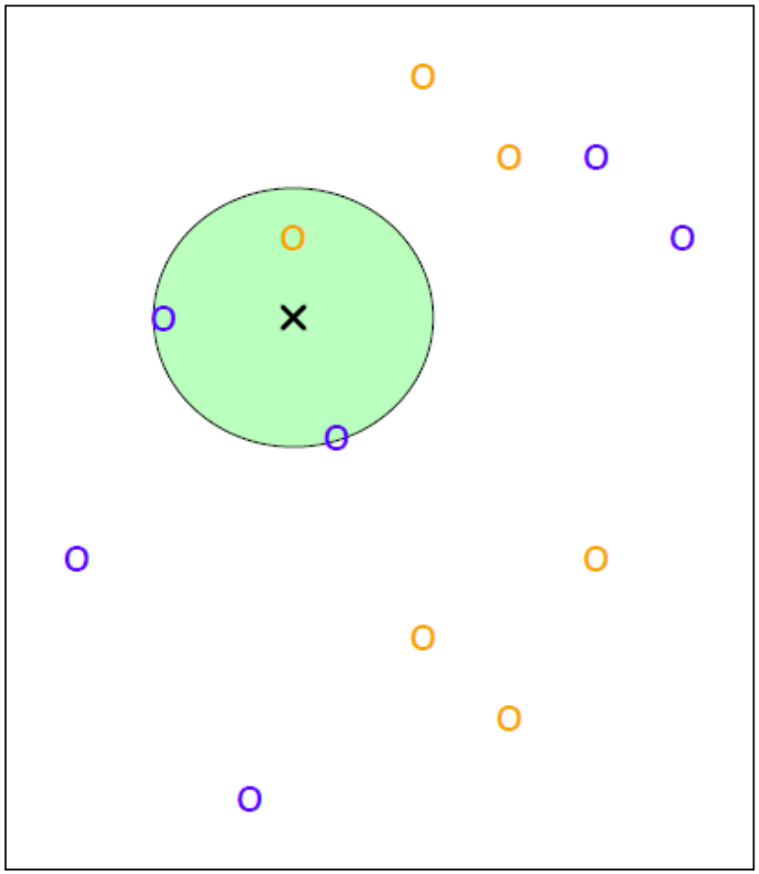


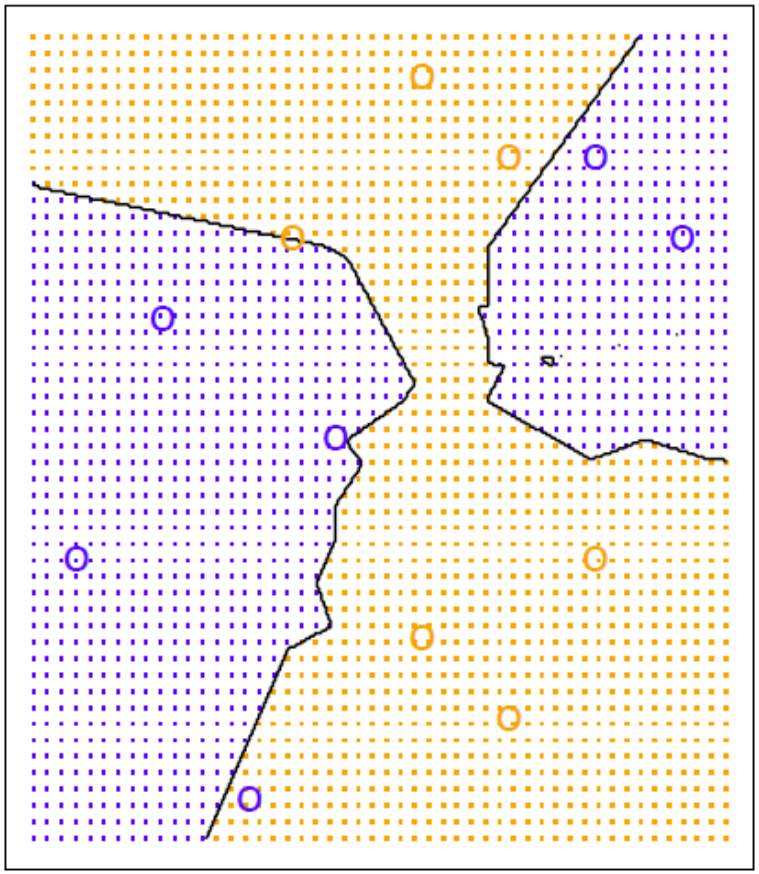
MARQUETTE UNIVERSITY
Be The Difference.

K-Nearest Neighbors (KNN)

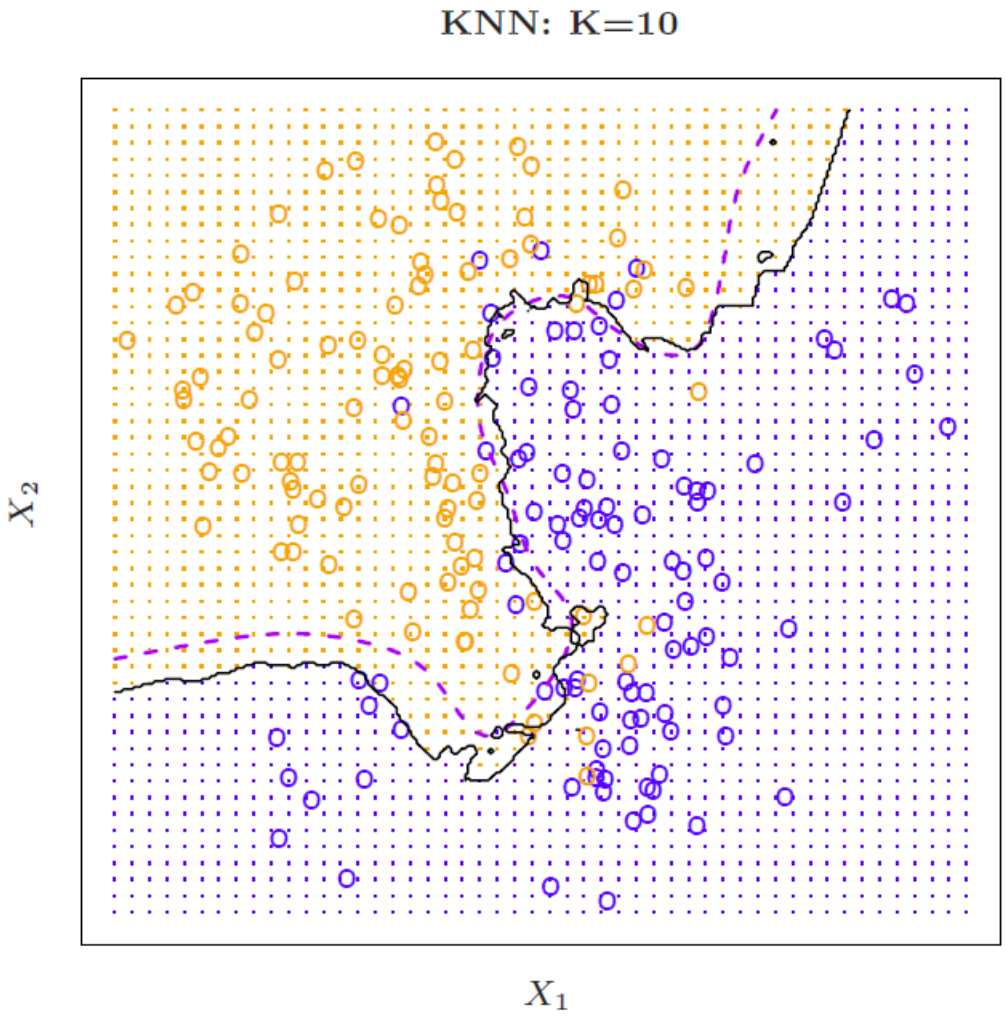
- k Nearest Neighbors is a flexible approach to estimate the Bayes Classifier.
- For any given X we find the k closest neighbors to X in the training data, and examine their corresponding Y.
- If the majority of the Y's are orange we predict orange otherwise guess blue.
- The smaller that k is the more flexible the method will be.

KNN EXAMPLE WITH K = 3

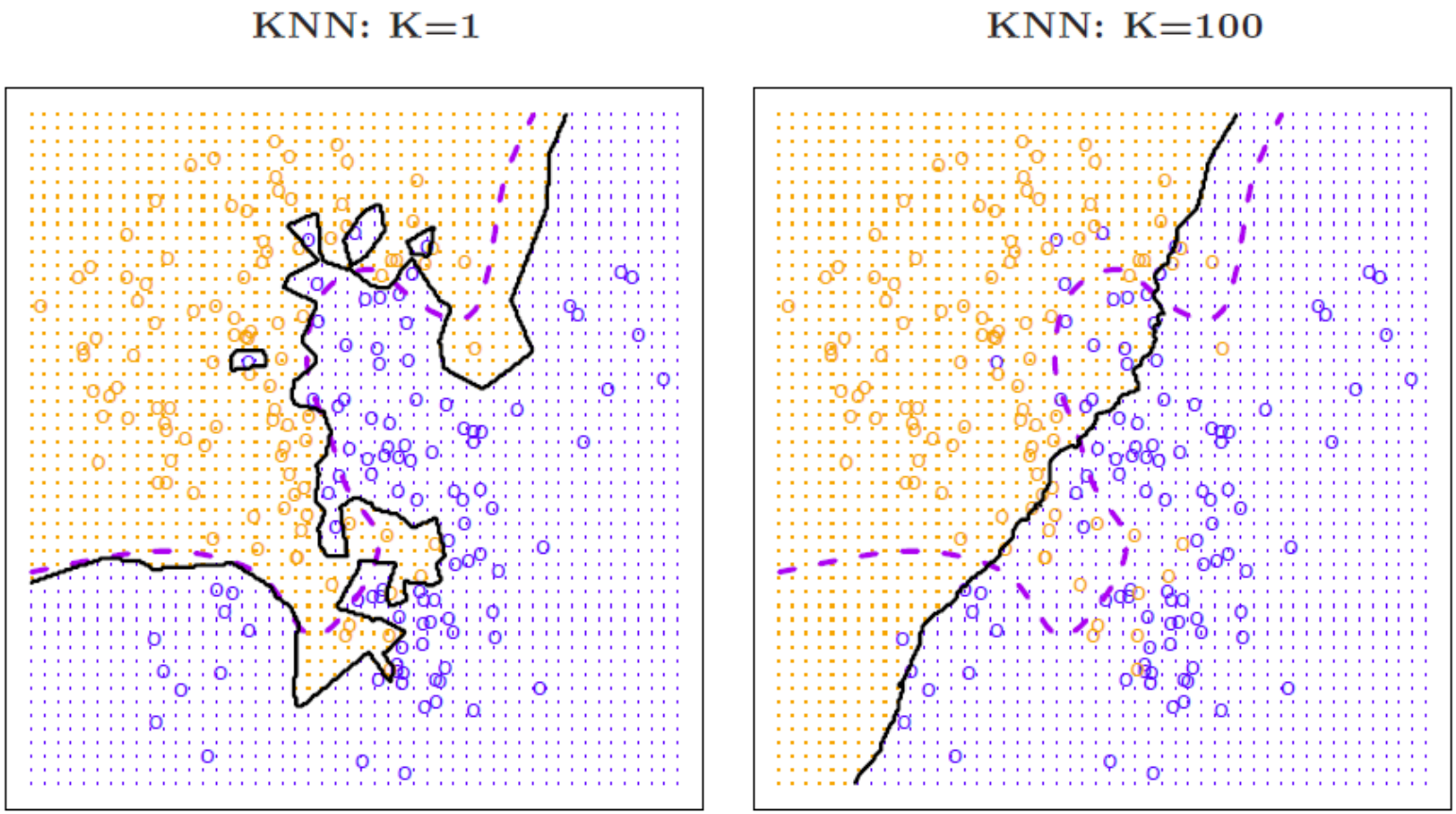




Simulated Data: K = 10

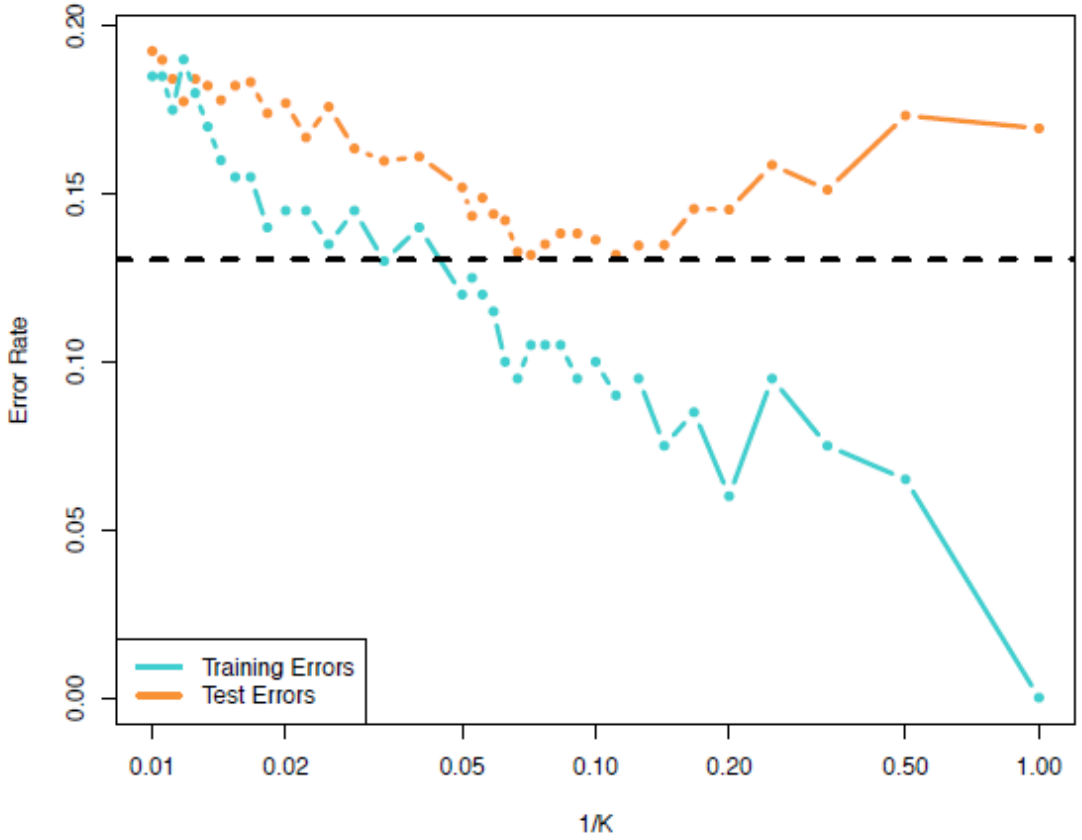


K = 1 AND K = 100



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.

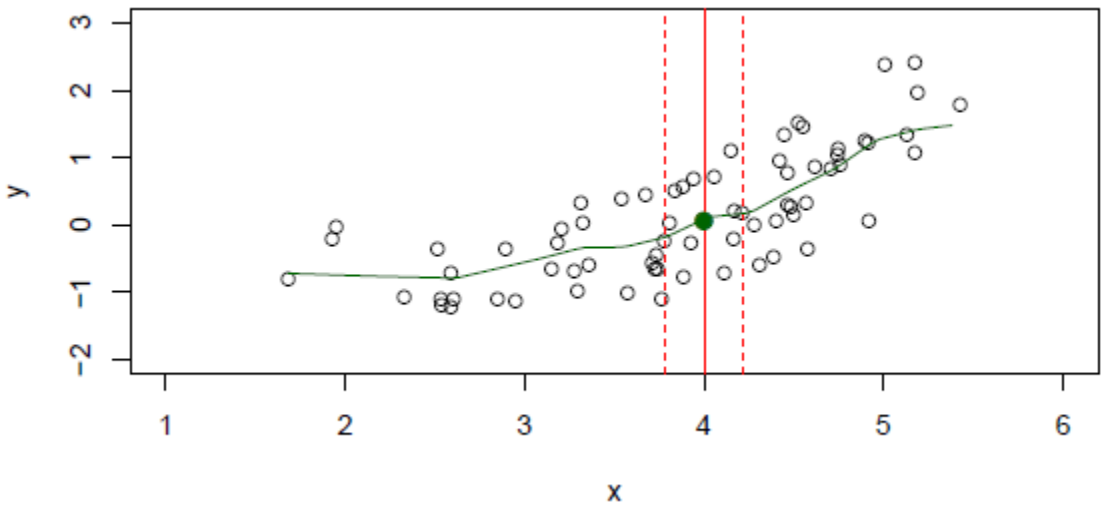


KNN IN REGRESSION

- Typically we have few if any data points with $X = 4$ exactly.
- So we cannot compute $E(Y|X = x)$!
- Relax the denfition and let

$$\hat{f}(x) = \text{Ave}(Y|X \in \mathcal{N}(x))$$

where $N(x)$ is some neighborhood of x.

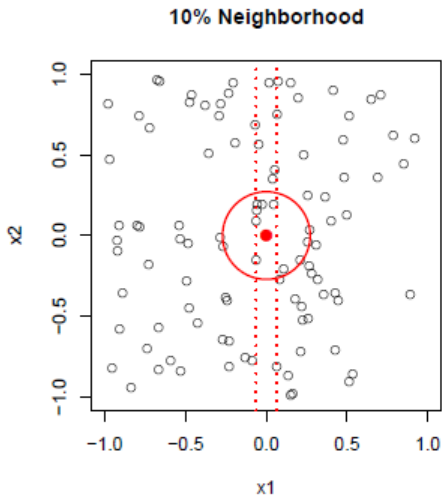


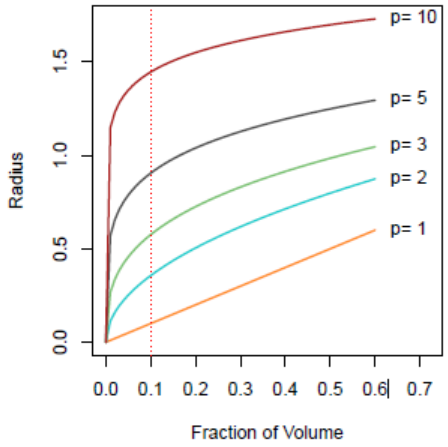
THE CURSE OF DIMENSIONALITY

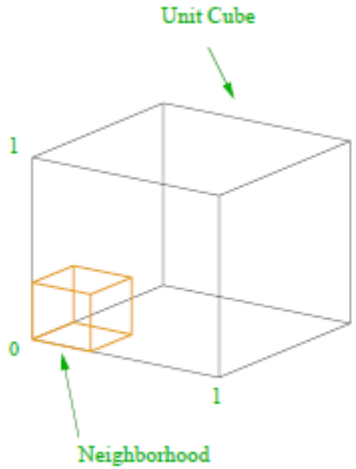
- Nearest neighbor averaging can be pretty good for small p – i.e. $p \leq 4$ and large-ish N.
  - Nearest neighbor methods can be lousy when p is large.

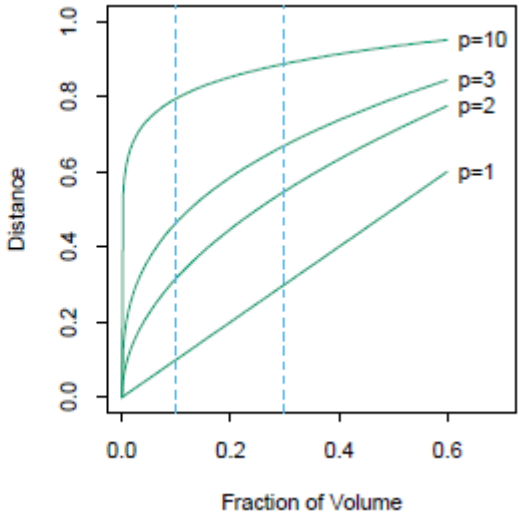
- Reason: the ***curse of dimensionality***. Nearest neighbors tend to be far away in high dimensions.
  - We need to get a reasonable fraction of the N values of y_i to average to bring the variance down – e.g. 10%.
  - A 10% neighborhood in high dimensions need no longer be local, so we lose the spirit of estimating $E(Y|X = x)$ by local averaging.

THE CURSE OF DIMENSIONALITY



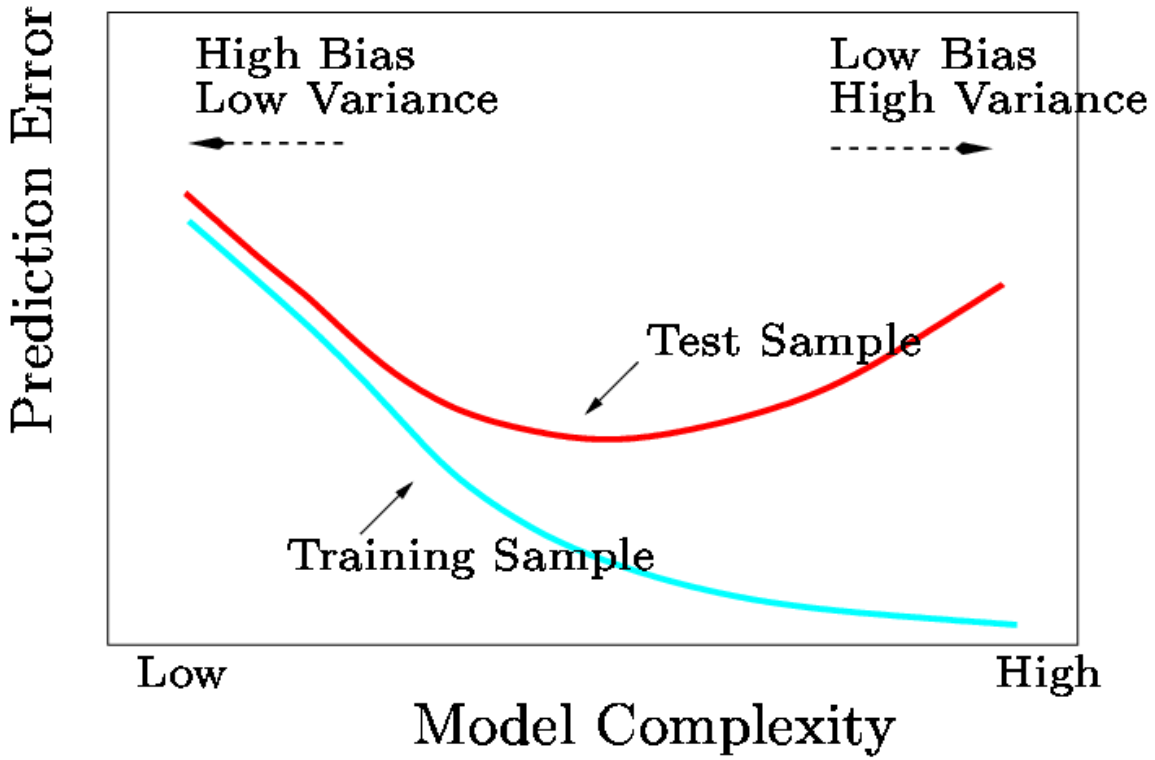






A Fundamental Picture

- In general training errors will always decline.
- However, test errors will decline at first (as reductions in bias dominate) but will then start to increase again (as increases in variance dominate).



We must always keep this picture in mind when choosing a learning method. More flexible/complicated is not always better!

QUESTIONS?

- ANY QUESTION?