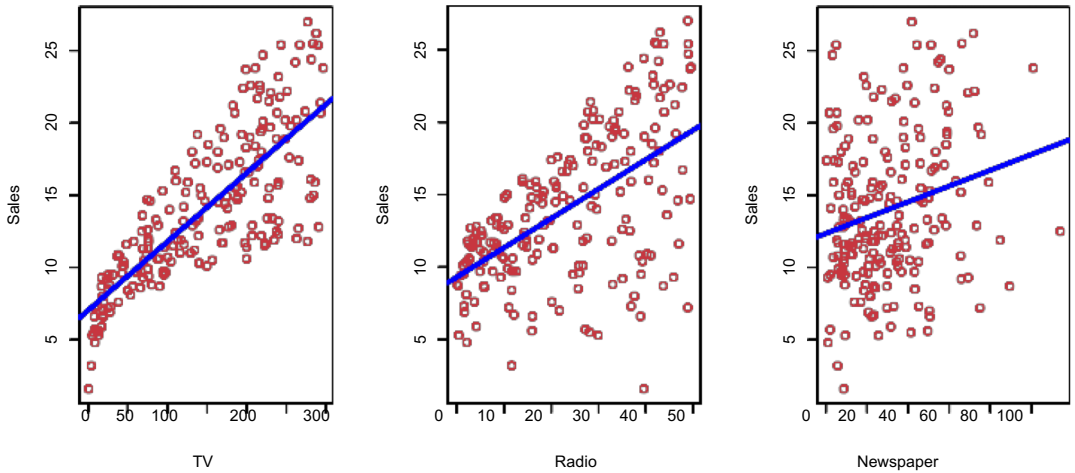
slides originally by
Dr. Richard Burns,
modified by
Dr. Stephanie Schwartz

STATISTICAL LEARNING

CSCI 406: Data Mining

What is Statistical Learning?



Shown are Sales vs TV, Radio and Newspaper, with a blue linear-regression line fit separately to each.

Can we predict Sales using these three? Perhaps we can do better using a model $\texttt{Sales} \approx f(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper})$

Statistical Learning – General Form

- In general, assuming we have
  - Observation of quantitative (numerical) response Y
  - Observation of p different predictors $\{X_1, X_2, \ldots, X_p\}$
  - A relationship between Y and X
  - We can write this in the very general form:

$$Y = f(X) + \varepsilon$$

Statistical Learning – General Form

$$Y = f(X) + \varepsilon$$

- Y is the *target* or *response* (in previous example: Sales)
- f is unknown function of $X = \{X_1, X_2, \ldots, X_p\}$
  - f may involve more than one input variable (in previous example: Radio, TV, Newspaper)
- ε is a random error term
  - Independent of X
  - Has mean equal zero
- f represents information that X provides about Y
- Statistical learning refers to a set of approaches for estimating f

Why estimate *f*?

- Two usual objectives:
  1. Prediction:
     - With a good f we can make predictions of Y at new points $X = x$
  2. Inference / Descriptive:
     - We can understand which components of $X = (X1, X2, \ldots, Xp)$ are important in explaining Y , and which are irrelevant. e.g. Seniority and Years of Education have a big impact on Income, but Marital Status typically does not.

Estimating *f* - Prediction

- In many situations, a set of *X* inputs are readily available, but *Y* is not easily obtained.

$$Y = f(X) + \varepsilon$$

- Since error term averages to zero, we can predict *Y* using,

$$\hat{Y} = \hat{f}(X)$$

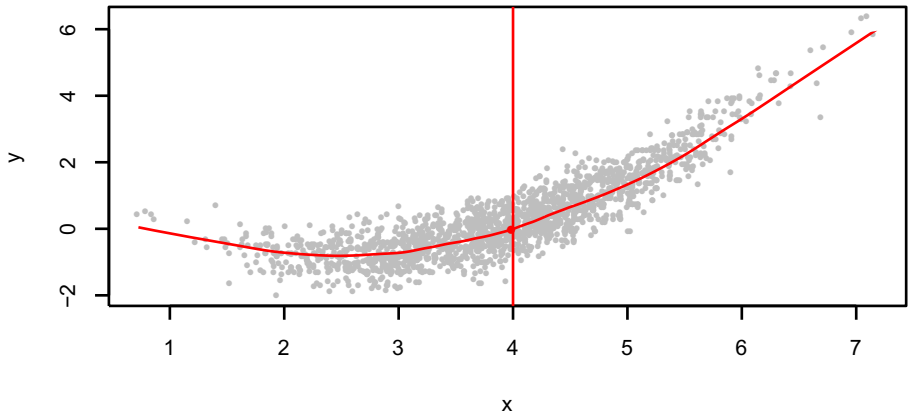
$\hat{f}$ represents estimate for f

$\hat{Y}$ represents prediction for Y

Estimating *f* - Prediction

$$\hat{Y} = \hat{f}(X)$$

- $\hat{f}$ often treated as a *black box*
  - Not typically concerned with the exact form of *f*
    - linear, quadratic, etc.
  - We only care that our predictions are "near accurate"



Is there an ideal $f(X)$? In particular, what is a good value for $f(X)$ at any selected value of X, say $X = 4$? There can be many Y values at $X = 4$. A good value is

$$f(4) = E(Y \mid X = 4)$$

$E(Y \mid X = 4)$ means *expected value* (average) of Y given $X = 4$.

This ideal $f(x) = E(Y \mid X = x)$ is called the *regression function*.

Estimating *f* – Types of Error

- The accuracy of $\hat{Y}$ as a prediction for *Y* depends on:
  1. Reducible error
  2. Irreducible error
- $\hat{f}$ will not be perfect estimate for *f*
  - *reducible*, because we can use more appropriate data mining techniques

Estimating *f* – Irreducible Error

- The accuracy of $\hat{Y}$ as a prediction for Y also depends on:
  - Irreducible error

$$Y = f(X) + \varepsilon$$

- $\varepsilon = Y - f(x)$
  - Even if we knew f(x), we would still make errors in prediction since at each X=x there is a distribution of possible Y values
  - Thus, variability associated with ε also affects prediction accuracy
- Cannot reduce error introduced by ε no matter how well we estimate *f*

Estimating *f* – Irreducible Error

- Why is irreducible error larger than zero?
- Quantity ε may contain unmeasured variables that are useful in predicting *Y*
  - If we don't measure them, *f* can't use them for its prediction
- Quantity ε may also contain unmeasureable variation

Estimating *f* –

- Focus in this course is on techniques for estimating *f* with the aim of minimizing reducible error.
- The irreducible error will always provide an upper bound on the accuracy of our predictions.
  - In practice, the upper bound because of irreducible error is almost always unknown.

Estimating f - Inference

- Rather than predicting Y based on observations of X,
- Goal is to understand the way that Y is affected as $X = \{X_1, X_2, \ldots, X_p\}$ changes
  - Understand the relationship between X and Y
- $\hat{f}$ not treated as "black box" anymore, we need to know its exact form

Estimating *f* - Inference

- May be interested in answering the following questions:
  - "Which predictors are associated with the response?"
    - Often the case that only small fraction of the available predictors are associated with *Y*
    - Identifying the few, important predictors

Estimating *f* - Inference

- May be interested in answering the following questions:
  - "What is the relationship between the response and each predictor?"
    - Some predictors may have a positive relationship with *Y* (or vice versa, a negative relationship)
    - Increasing the predictor is associated with increasing values of *Y*

Estimating *f* - Inference

- May be interested in answering the following questions:
  - "Can $\hat{f}$ be summarized using a linear equation, or is the relationship more complicated?"
    - Historically, most methods for estimating *f* have taken a linear form
    - But often true relationship is more complicated
    - Linear model may not accurately represent relationship between input and output variables

How do we estimate *f*?

- Most statistical learning methods classified as:
  1. Parametric
  2. Non-parametric

Parametric Methods

- Assume that the functional form, or shape, of *f* is linear in *X*

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

- This is a linear model, for *p* predictors $X = \{X_1, X_2, \ldots, X_p\}$
- Model fitting involves estimating the parameters $\beta_0, \beta_1, \ldots, \beta_p$
- Only need to estimate *p+1* coefficients,
  - Rather than an entirely arbitrary *p*-dimensional function *f(X)*
- Parametric: reduces the problem of estimating *f* down to estimating a set of parameters

Non-parametric Methods

- Do not make explicit assumptions about the functional form of *f* (such that it is *linear*)

Parametric Methods

- Assumption of form of model (perhaps linear)
- Possible that functional estimate is very different from the true *f*
  - If so, won't fit data well
- Only need to estimate set of parameters

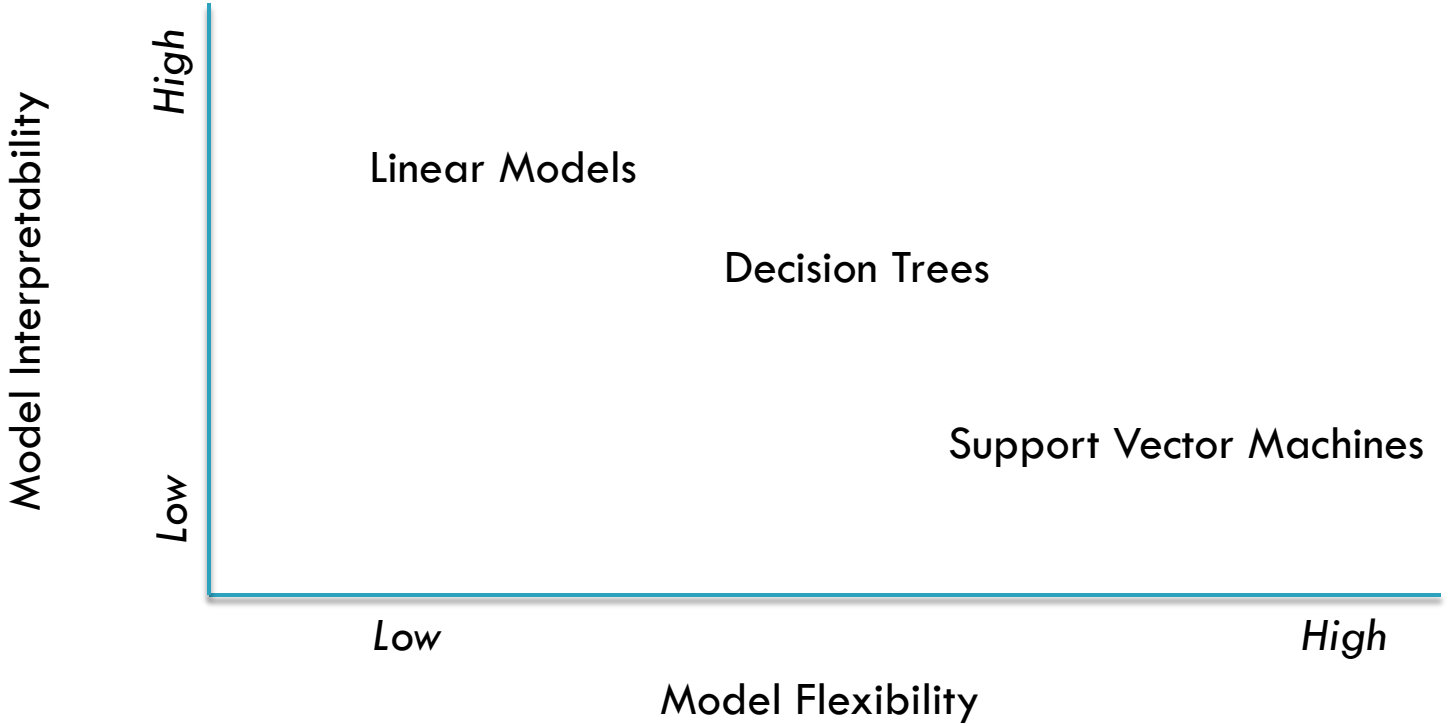
Non-parametric Methods

- Potential to accurately fit a wider range of possible shapes for *f*
- Many, many more observations needed
- Complex models can lead to overfitting

Trade-Off Between Model Flexibility and Model Interpretability

- Some statistical models (e.g. linear models) are less flexible and more restrictive.
- Q: Why would be ever choose to use a more restrictive method instead of a very flexible approach?
- A: When *inference* is the goal, the restrictive models are much more interpretable.
  - In linear model, it is easy to understand relationship between Y and X_1, X_2, …
- For *prediction*, we might only be interested in accuracy and not the interpretability of the model

Trade-Off Between Model Flexibility and Model Interpretability



Trade-Off Between Model Flexibility and Model Interpretability

- Even for *prediction*, where we might only care about accuracy, more accurate predictions are sometimes made from the less flexible methods
  - Reason: overfitting in more complex models

Classification vs. Regression

- Given a dataset: instances with *X* set of predictors/attributes, and single *Y* target attribute
- Classification:
  - *Y* Class label is discrete (usually categorical/nominal or binary) attribute
- Regression:
  - *Y* Class label is continuous
  - Numeric prediction

Supervised Learning Approach to Classification or Regression Problems

- Given a collection of records (training set)
  - Each record contains predictor attributes as well as target attribute
- Learn a model (function *f*) that predicts the class value (category or numeric value) based on the predictor attributes
- Goal: "*previously unseen*" instances should be assigned a class as *accurately* as possible
  - A test set is used to evaluate the model's accuracy.

Training Set vs. Test Set

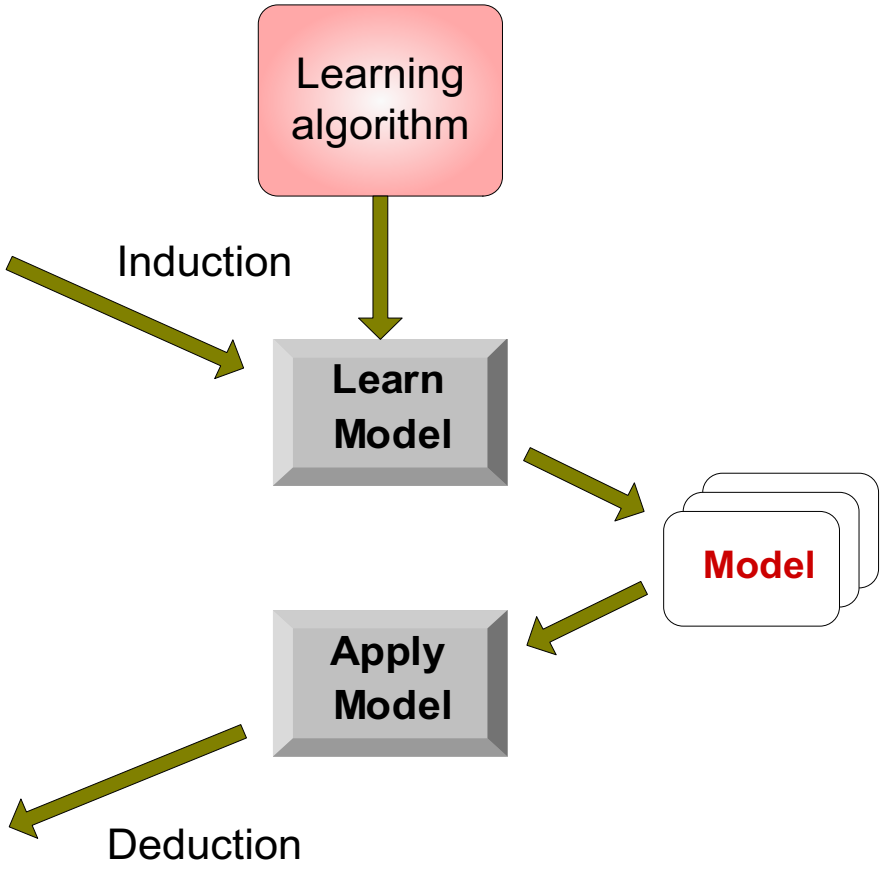
- Overall dataset can be divided into:
  1. Training set – used to build model
  2. Test set – evaluates model

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Model Evaluation on Test Set (Classification) – Error Rate

- Error Rate: proportion of mistakes that are made by applying our $\hat{f}$ model to the testing observations:

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

Observations in test set: $\{(x_1,y_1), \ldots, (x_n,y_n)\}$

$\hat{y}_i$ is the predicted class for the ith record

$I(y_i \neq \hat{y}_i)$ is an indicator variable: equals 1 if $y_i \neq \hat{y}_i$ and 0 if $y_i = \hat{y}_i$

Model Evaluation on Test Set (Classification) – Confusion Matrix

- Confusion Matrix: tabulation of counts of test records correctly and incorrectly predicted by model

		Predicted Class	
		Class = 1	*Class* = 0
Actual Class	*Class* = 1	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

(Confusion matrix for a 2-class problem.)

Model Evaluation on Test Set (Classification) – Confusion Matrix

		Predicted Class	
		Class = 1	Class = 0
Actual Class	Class = 1	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

$$\text{Error rate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

Most classification tasks seek models that attain the highest accuracy when applied to the test set.

Model Evaluation on Test Set (Regression) – Mean Squared Error

- Mean Squared Error: measuring the "quality of fit"
  - will be small if the predicted responses are very close to the true responses

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

Observations in test set: $\{(x_1,y_1), \ldots, (x_n,y_n)\}$

$\hat{f}(x_i)$ is the predicted value for the ith record

A Problem

- We already know that there is no one "best" data mining method or statistical learning method.
  - Depends on the characteristics of the data
- We've introduced evaluation:
  - We can quantify error (classification error, mean squared error) in hopes of comparing accuracy of different models
- We have datasets partitioned:
  - Training set – model learns on this data
  - Test set – model evaluated on this data

How well the model works on new data is what we really care about!

A Problem

- Error rates on training set vs. testing set might be drastically different.
- There is no guarantee that the method with the smallest training error rate will have the smallest testing error rate.
- Why?
  - Statistical methods specifically estimate coefficients so as to minimize the training set error

Overfitting

- Overfitting: occurs when statistical model "memorizes" the training set data
  - very low error rate on training data
  - higher error rate on test data
- Model *does not generalize* to the overall problem
- This is bad! We wish to avoid overfitting.

Learning Method Bias

- Bias: the error introduced by modeling a real-life problem (usually extremely complicated) by a much simpler problem
  - *Example:* linear regression assumes a linear relationship between the target variable *Y* and the predictor variables *X*
  - It's unlikely that the relationship is exactly linear, so some bias will be present in the model.
- The more flexible (complex) a method is, the less bias it will generally have.

Learning Method Variance

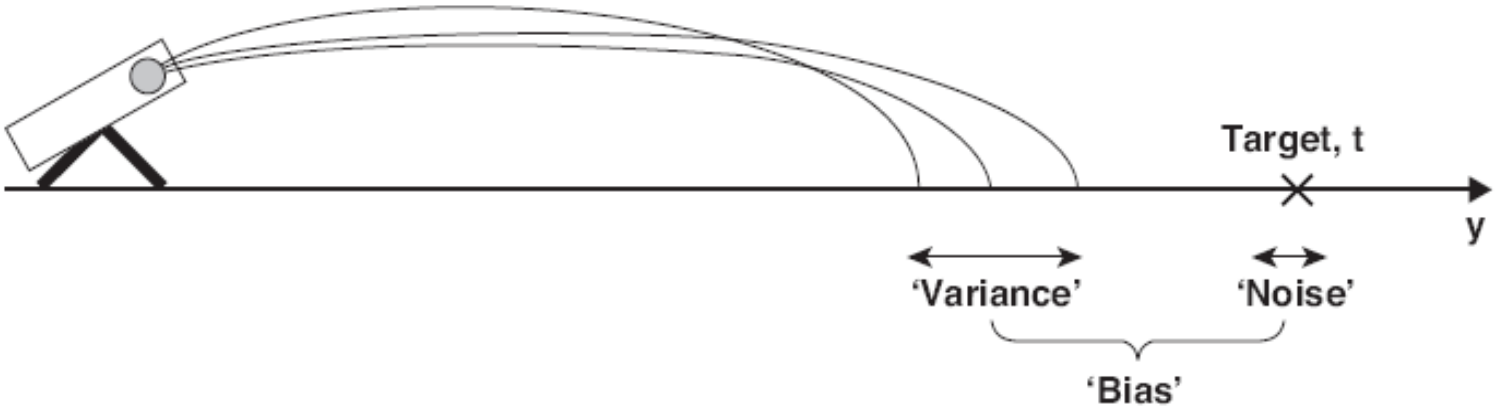
- Variance: how much the learned model would change if the training set was different
  - *Does changing a few observations in the training set, dramatically affect the model?*
    - Ideally, answer is no.
- Generally, the more flexible (complex) a method is, the more variance it has.

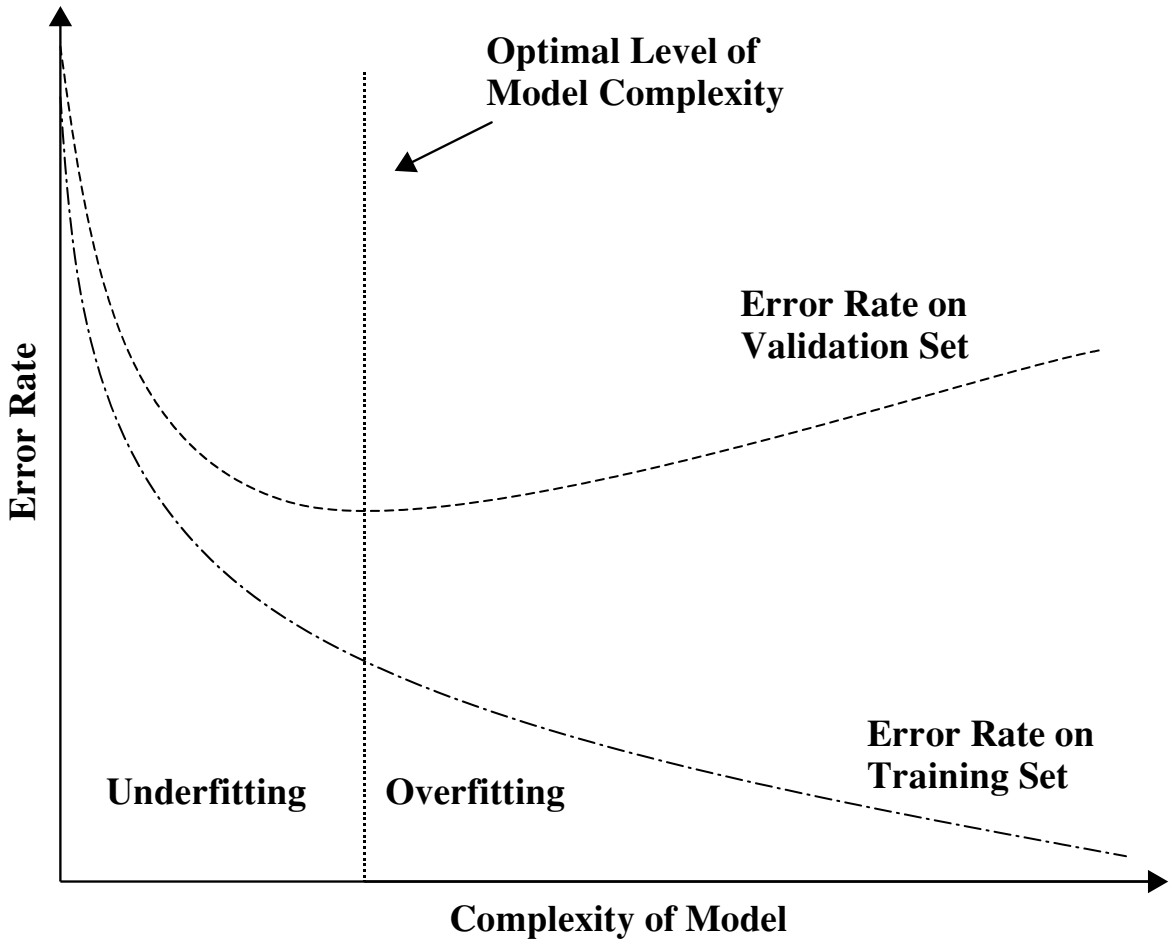
Bias-Variance Trade-Off

- Math proof! *(beyond scope of this course)*
- Expected test set error can be decomposed into the sum of the model's variance, its squared bias, and the variance of its error terms.

$$E(y_0 - \hat{f}(x_0))^2 = Variance(\hat{f}(x_0)) + [Bias(f(x_0))]^2 + Variance(\varepsilon)$$

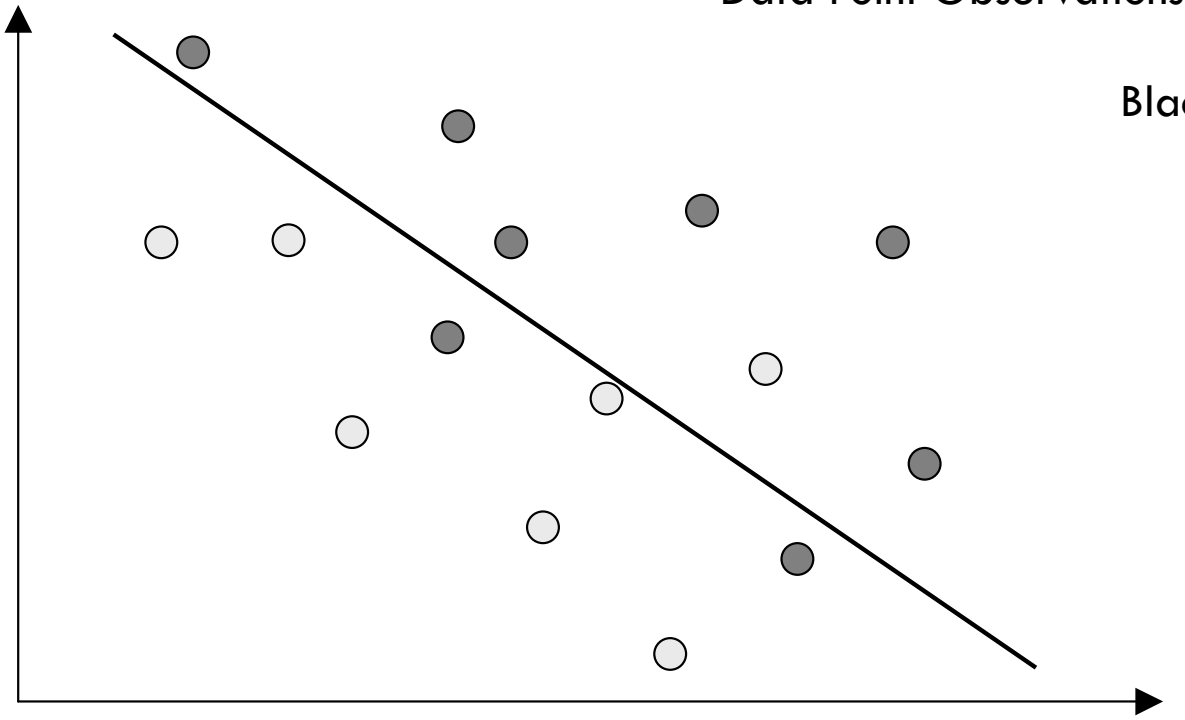
- As a statistical method gets more complex, the bias will decrease and the variance will increase.
- Expected error on the test set may go up or down.


Target, t
y
'Variance'
'Noise'
'Bias'


Optimal Level of
Model Complexity
Error Rate on
Validation Set
Error Rate
Error Rate on
Training Set
Underfitting
Overfitting
Complexity of Model

Example: we wish to build a model that separates the dark-colored points from the light-colored points.

Data Point Observations created by: $Y=f(X)+\varepsilon$

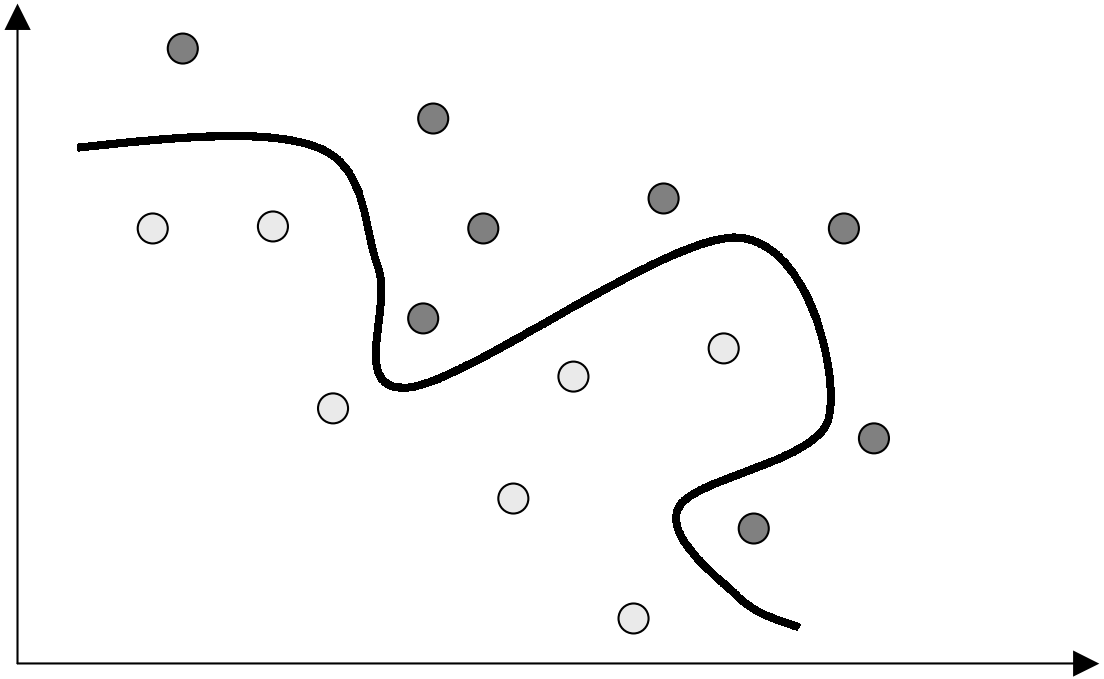


Black line is simple, linear model $\hat{f}$

Currently, some classification error

- Low variance
- Bias present

More complex model (curvy line instead of linear)

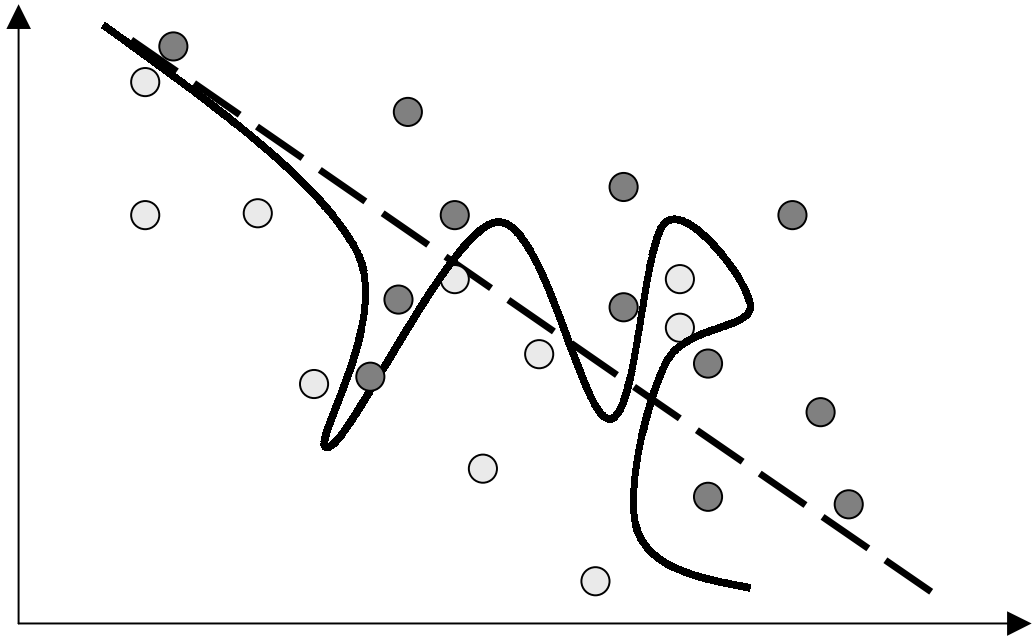


Zero classification error for these data points

- No linear model bias
- Higher Variance?

More data has been added

Re-train both models (linear line, and curvy line) in order to minimize error rate



Variance:

- Linear model doesn't change much
- Curvy line significantly changes

Which model is better?

- Now that we know the definitions of “training set” and “testing set”,
  - A more complete view of the Data Mining process…

Data Mining Process

1. Engage in efficient data storage and data preprocessing
2. Select appropriate response variables
   - Decide on the number of variables that should be investigated
3. Screen data for outliers
   - Address issues of missing values
4. Partition datasets into training and testing sets
   - Sample large datasets that cannot easily be analyzed as a whole

Data Mining Process (cont.)

5. Visualize data
   - Box plots, histograms, etc.
6. Summarize data
   - Mean, median, sd, etc.
7. Apply appropriate data mining methods (decision trees)
8. Evaluate model on test set
9. Analyze, interpret results
   - Act on findings

References

- *Introduction to Data Mining*, 1st edition, Tan et al.
- *Data Mining and Business Analytics in R*, 1st edition, Ledolter
- *An Introduction to Statistical Learning*, 1st edition, James et al.
- *Discovering Knowledge in Data*, 2nd edition, Larose et al.