IST 5535: Machine Learning Algorithms and Applications

Langtao Chen, Spring 2021

1. Introduction to Machine Learning

Reading

- Book Chapters 1, 2 (sections 2.1, 2.2)

- Online Article: Statistics – Understanding the Levels of Measurement
  - http://www.kdnuggets.com/2015/08/statistics-understanding-levels-measurement.html

Learning Objectives

- Explain important concepts related to machine learning
- Understand dataset and be able to distinguish among different scales of measurement
- Explain methods used to assess model accuracy
- Explain bias-variance trade-off

OUTLINE

- (I) Overview of machine learning (ML)
  1. What is learning?
  2. Practical definition of ML
  3. ML model estimation methods: parametric, nonparametric
  4. Types of ML

- (II) Scale of measurement

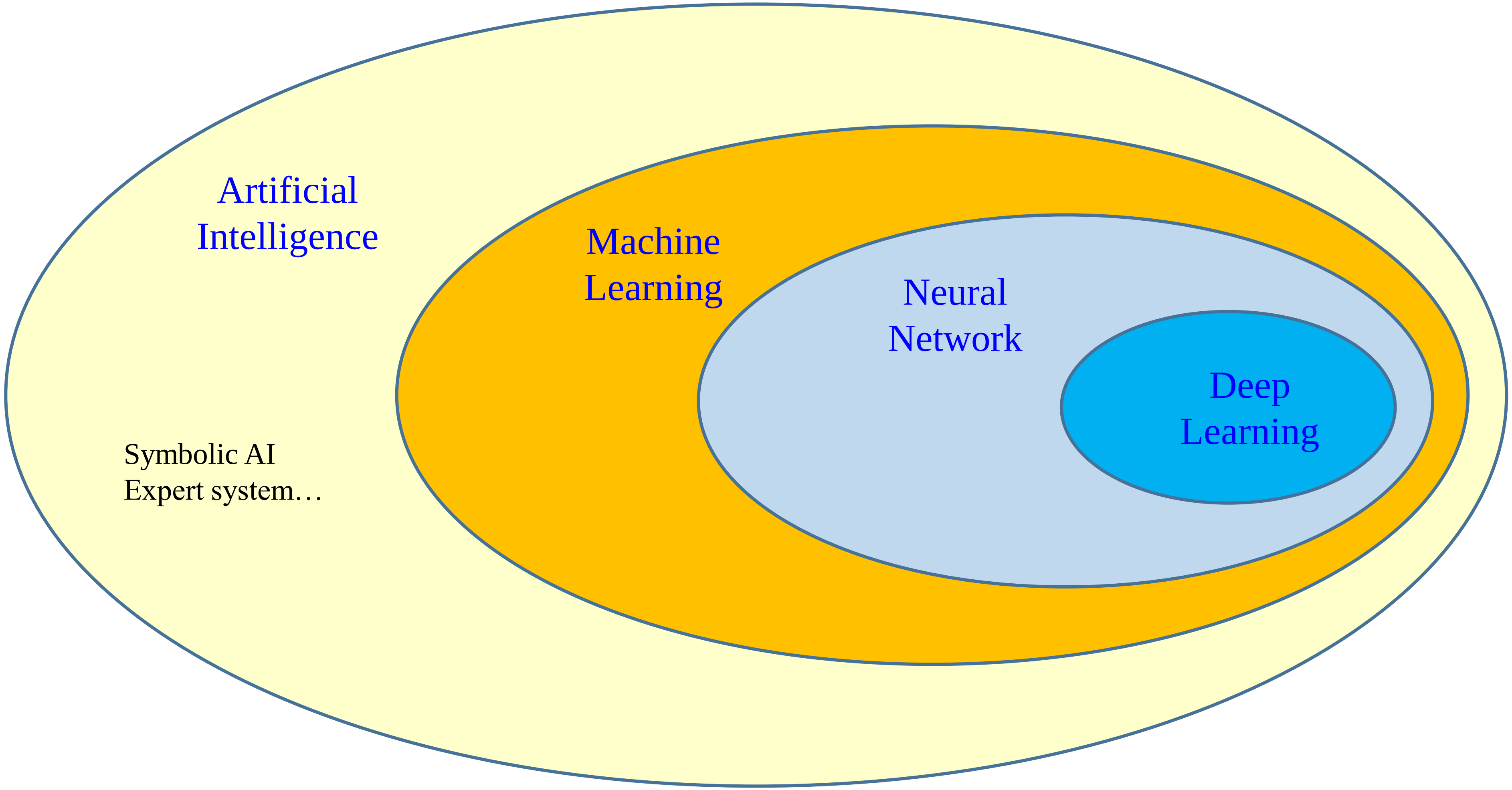
  Nominal, ordinal, interval, ratio

- (III) Model accuracy
  1. Regression setting: MSE, training MSE, test MSE
  2. Classification setting: Error rate
  3. Bias variance tradeoff

AGENDA

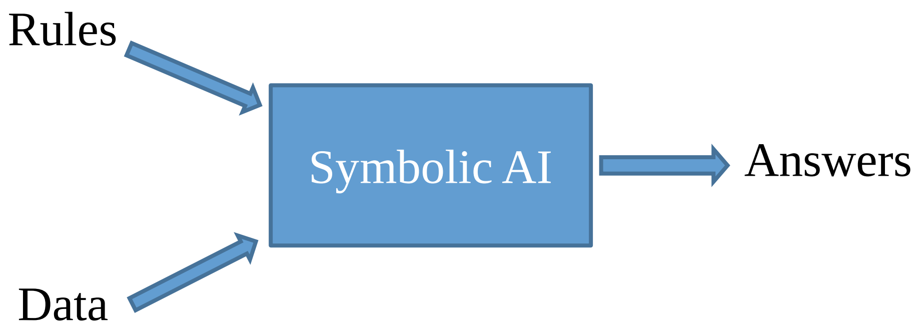
- Overview of Machine Learning
- Dataset and Scales of Measurement
- Assessing Model Accuracy

AI, Machine Learning, Neural Network, and Deep Learning

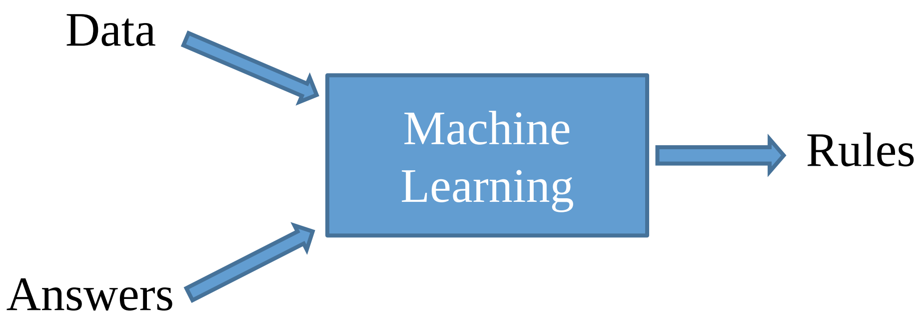


What is a Learning?

- Symbolic AI



- Machine Learning



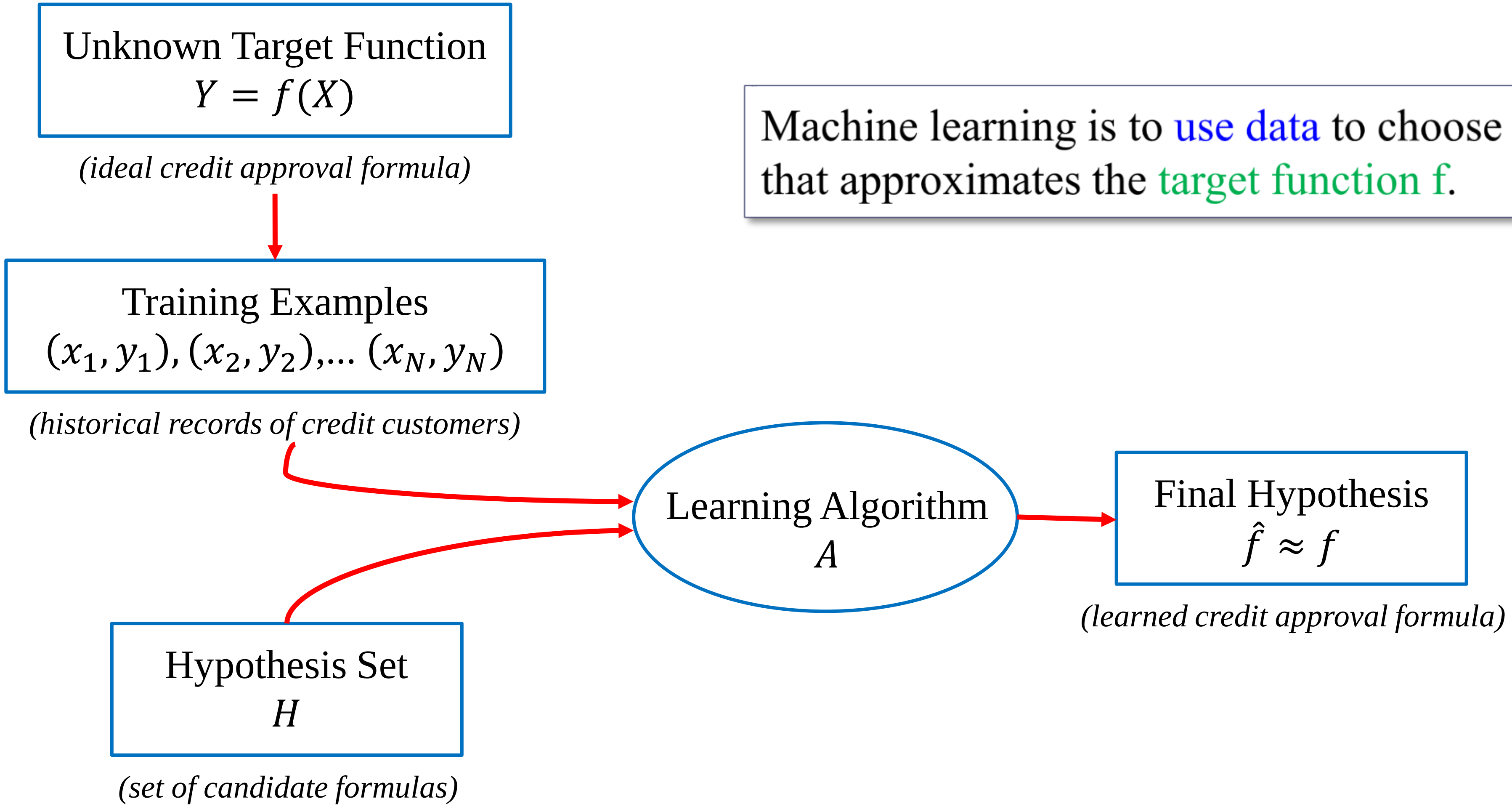
Central Research Questions of Machine Learning

- How can we build computer systems that automatically improve with experience?
- What are the fundamental laws that govern all learning processes?

> "Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed".
>
> ----Wikipedia

Practical Definition of Machine Learning

Basic Setup of the learning problem (adapted from Abu-Mostafa et al 2012)



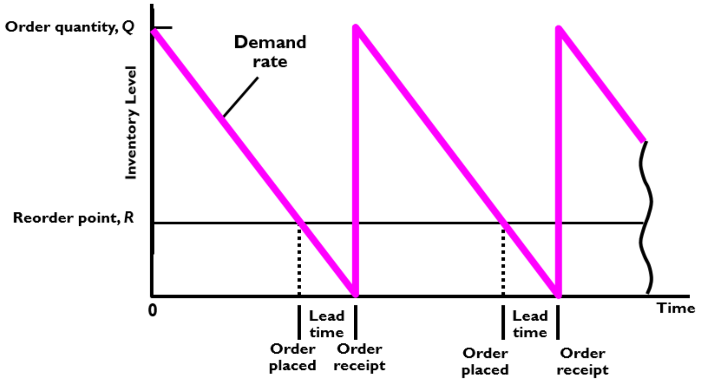
Machine learning is to use data to choose a hypothesis $\hat{f}$ that approximates the target function f.

When do We Need Machine Learning?

Some problems have analytic solutions

- What is the optimal ordering quantity in order to minimize the total inventory cost?

EOQ model: $Q_{opt} = \sqrt{\dfrac{2C_oD}{C_c}}$



Only empirical solutions are feasible

- How can we classify an email as either spam or ham?

When there is no analytic solution but we do have a lot of data, we can use machine learning methods to construct an empirical solution from the data.

Learning = Representation + Evaluation + Optimization

- Representation: Formal language used to represent a learning algorithm
- Evaluation: Assess the performance of algorithms
- Optimization: Search the optimal solutions

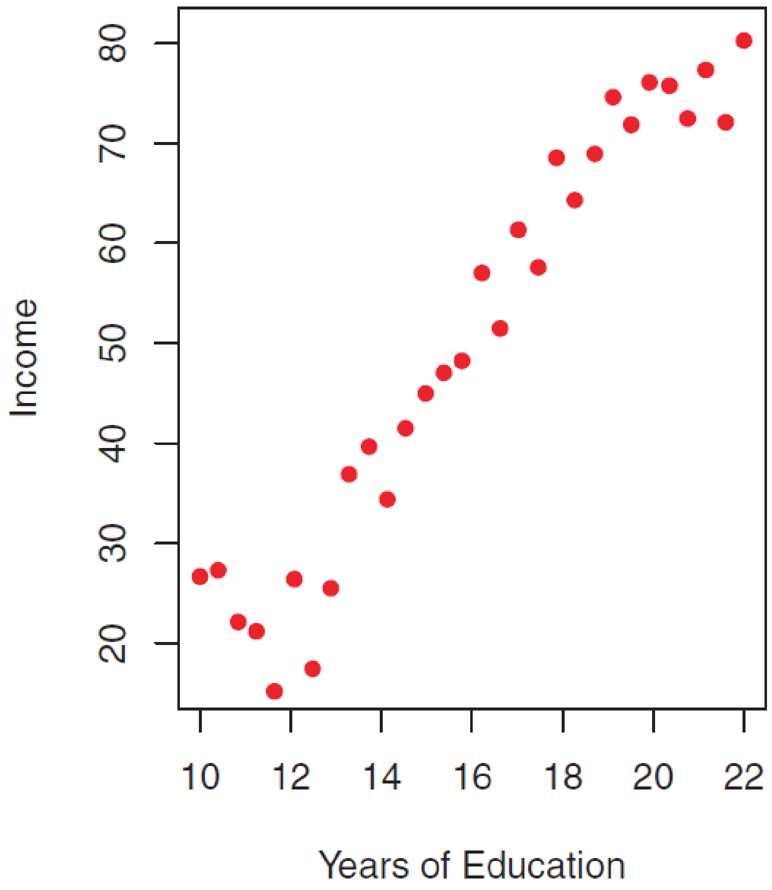
Table 1: The three components of learning algorithms.

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

Source: Pedro Domingos, “A Few Useful Things to Know about Machine Learning”

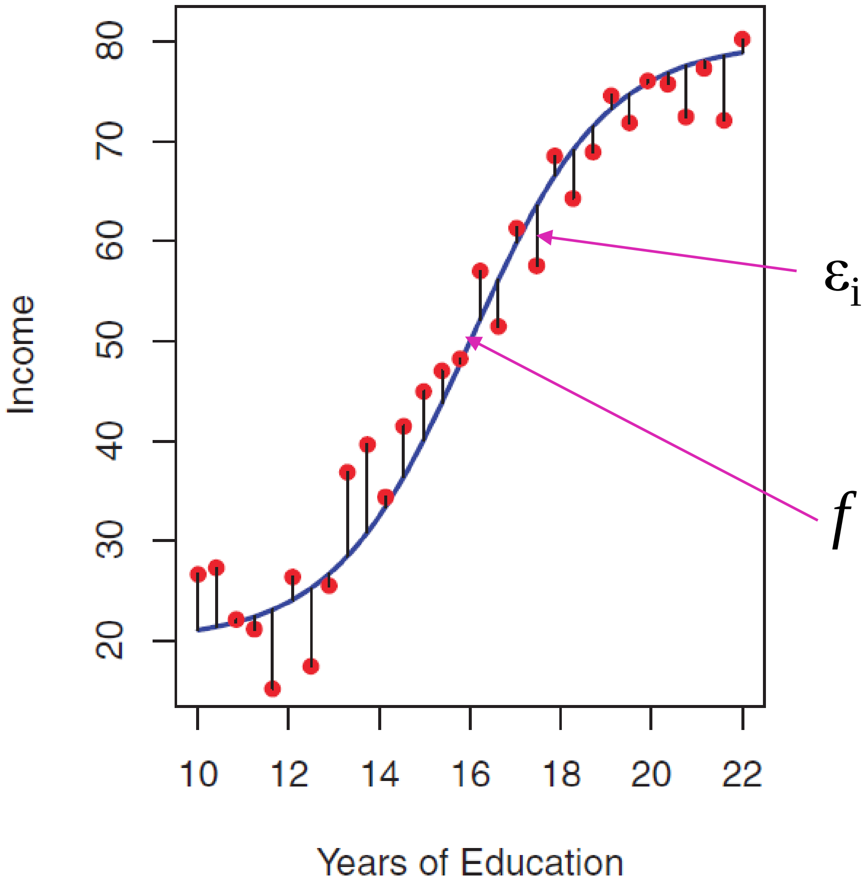
An Example of Machine Learning

- Observed pattern
- True underlying relationship



Machine Learning

$\hat{f} \approx f$



Why Do We Estimate f?

- Machine learning is all about estimating the unknown function f.

- Two major reasons for estimating f: $\hat{Y} = \hat{f}(X)$
  - Prediction
    - If $\hat{f}$ approximates f well, we can accurately predict Y based on new value of X.
    - $\hat{f}$ is often treated as a black box.
  - Inference
    - We are interested in understanding the relationship between X and Y.
    - $\hat{f}$ should be a white box. We need to know its exact form.

How Do We Estimate *f* ?

- Parametric methods: Reduce the problem of estimating f down to one of estimating a set of parameters.
- A two-step model-based approach
  - Step 1: Make assumption about the functional form, or shape, of *f*

    For example, assume linear relationships (linear model)

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

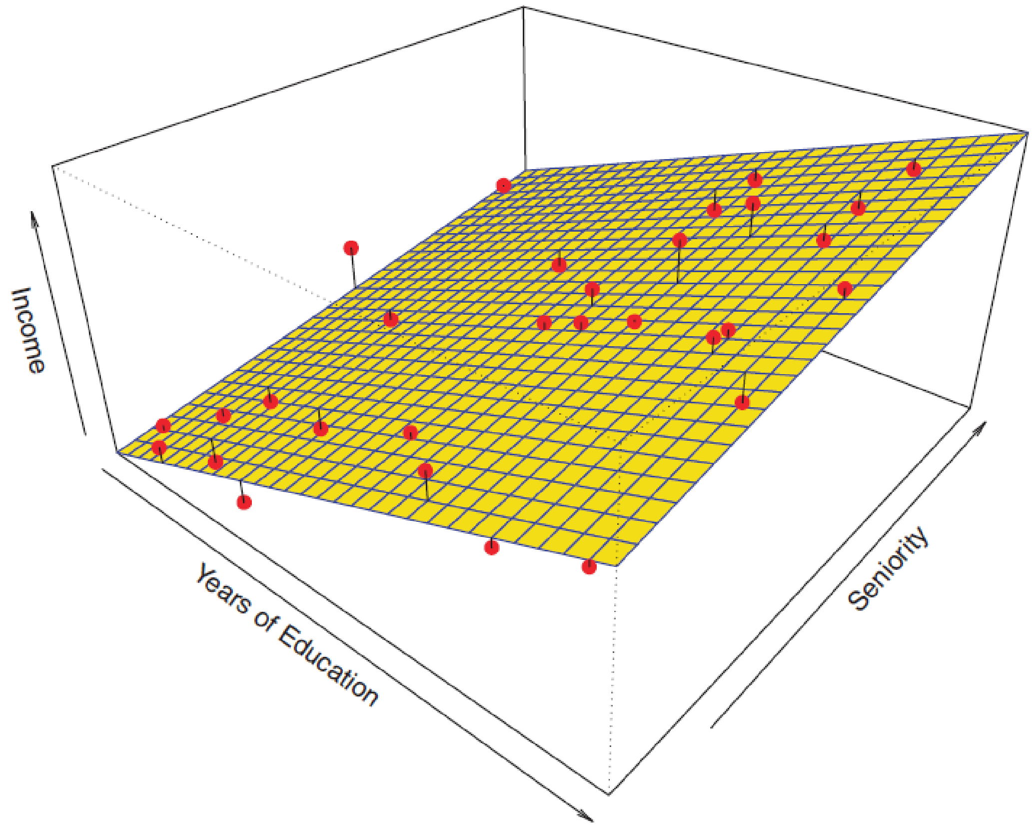
  - Step 2: Use a procedure that uses the training data to *fit* or *train* the model

    For example, use ordinary least square (OLS) or maximum likelihood (ML) to estimate the parameters $\beta_0, \beta_1, \beta_2, \ldots \beta_p$

An Example of Parametric Method

- A linear model fit by OLS to the income data

$$\color{red}{Income} \approx \beta_0 + \beta_1 \times \color{red}{education}$$
$$+\beta_2 \times \color{red}{seniority}$$



The true f has some curvature that is not captured in the linear fit

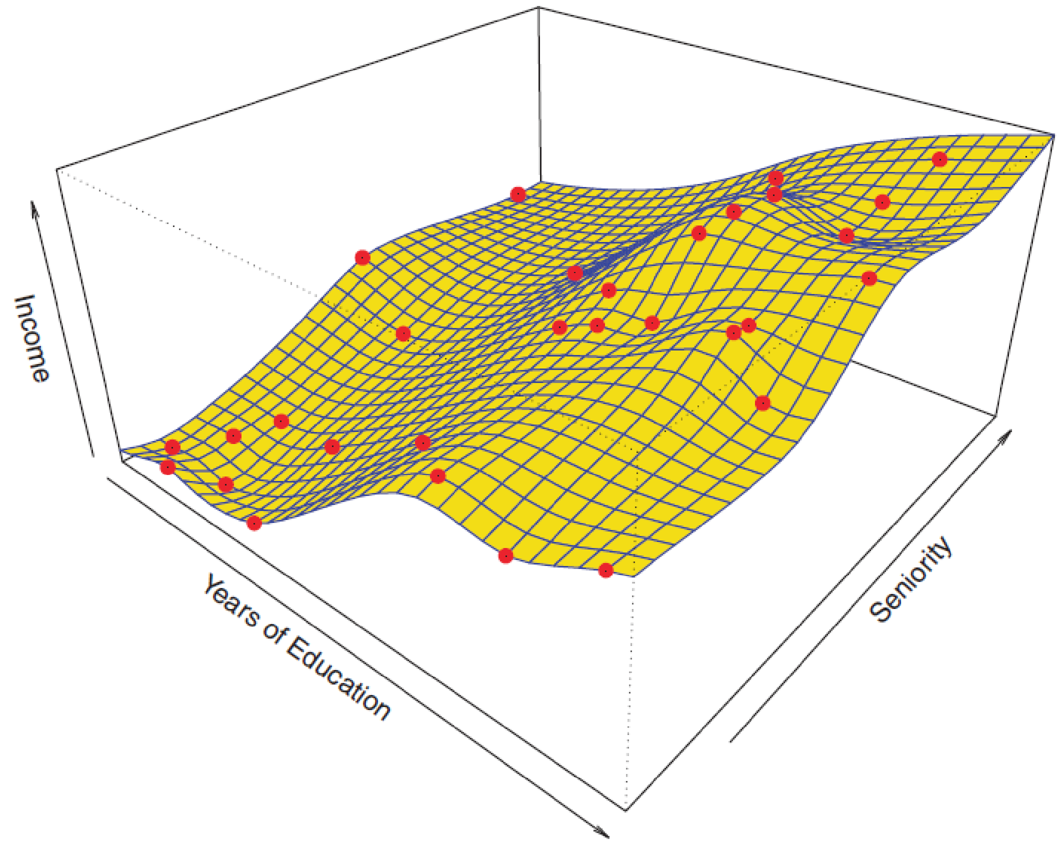
How Do We Estimate *f* ?

- Non-parametric methods
  - No explicit assumptions about the functional form of *f*
  - Advantage: Have the potential to accurately fit a wide range of possible shapes of *f*
  - Disadvantage: A large number of observations is required in order to obtain an accurate estimate of *f*

An Example of Non-parametric Method

- A rough thin-plate spline fit to the income data

This fit makes zero errors in the training data



However, there is likely an overfitting issue: The fitted model will not yield accurate estimates of the response on new data.

Why are there so many machine learning algorithms?

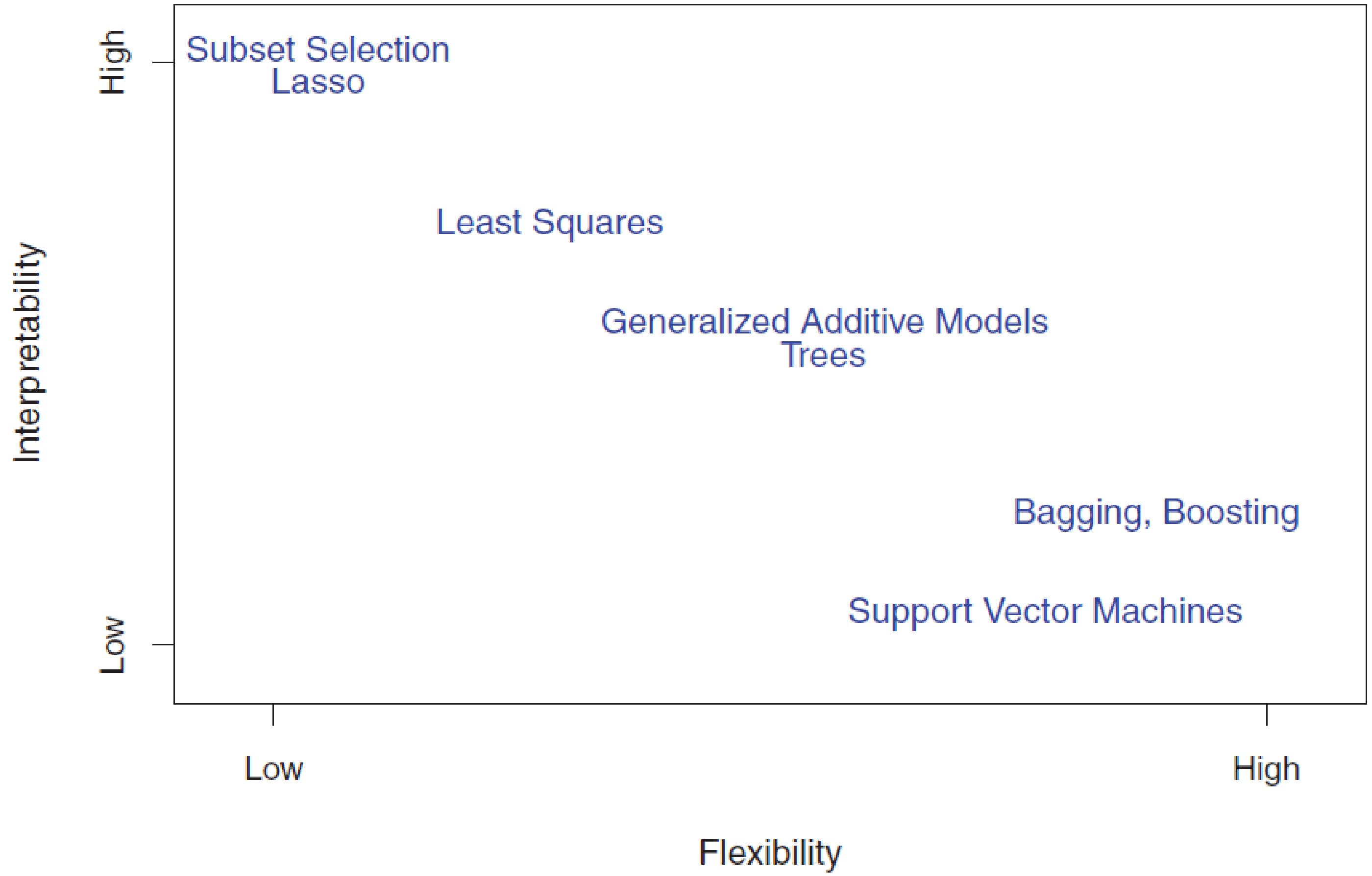
- "No free lunch theorem"

  There is no such a single algorithm that is uniquely better for all problems.

- So we'll learn a couple of important machine learning algorithms in this class.

Tradeoff between Prediction Accuracy and Model Interpretability

- In general, as the flexibility of a method increases, its interpretability decreases



Tradeoff between Prediction Accuracy and Model Interpretability

- Why would we prefer a more restrictive model over a very flexible model?
  - If we are mainly interested in inference, restrictive models are much more interpretable;
  - Even when inference is NOT the goal, less flexible models are not likely to overfit the data, thus often providing more accurate estimate.

Types of Learning

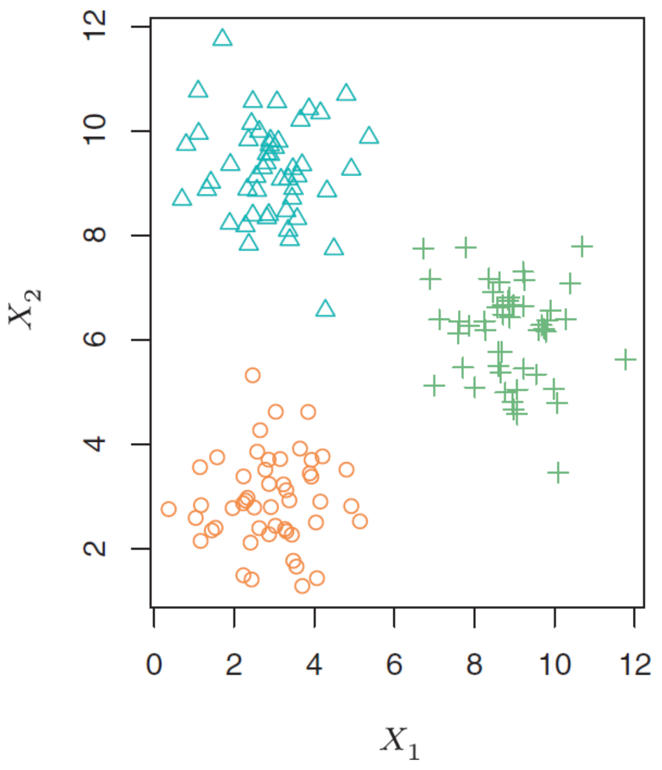
- Supervised Learning
  - Supervised learning algorithms are used for prediction and classification.
  - We need to supervise the learning of the algorithm by using training data to train the algorithm.
  - Data are labeled with correct output: (X_i, y_i), $i=1...N$
  - The most studied type of learning

Types of Learning

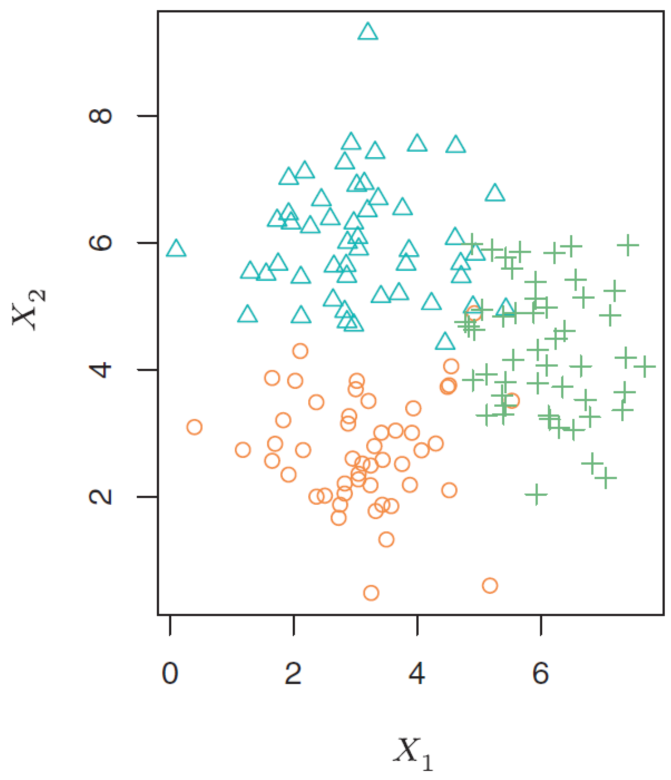
- Unsupervised Learning
  - Unsupervised learning algorithms are used when there is no outcome variable to predict or classify. We simply learn something from the inputs by themselves.
  - Data are unlabeled: only $X_i (i=1...N)$ are observed
  - There is no training-testing partition of the dataset.
  - Popular scenarios include association rules and clustering.

Types of Learning

- Unsupervised Learning Example



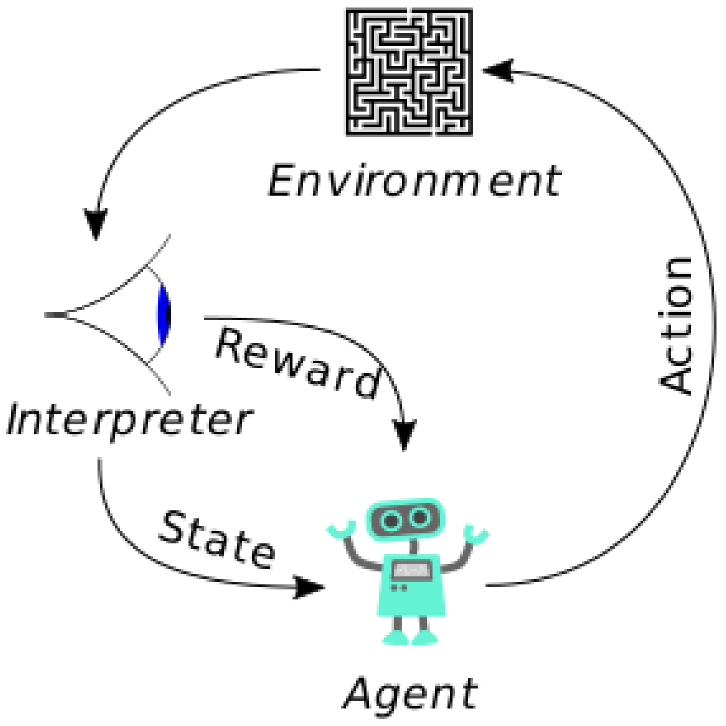
3 groups are well-separated



Overlap among 3 groups.
Clustering is more challenging.

Types of Learning

- Reinforcement Learning
  - To let computer agent learn like people, without godlike “supervisor” providing correct output.
  - Data is unlabeled, but contain some possible outputs with their goodness scores.
  - The agent learns from experience through trial-and-error.
  - The goal is to maximize long-term reward.



An agent takes actions in an environment, which is interpreted into a reward and a representation of the state, which are fed back into the agent.

Applications of Reinforcement Learning

- Game play: AlphaGo trumped a human Go campion in 2016

Photo: Google

Applications of Reinforcement Learning

- Self-driving cars

Model-Based Learning Vs. Instance-Based Learning

Model-Based Learning

- Model-based learning tries to build a model $y=f(x)$ from training data and then use the model to generalize to new problem.
- Usually model training is computationally intensive, while prediction is easy and simple.
- Examples: regression, neural network, hidden Markov model…

Instance-Based Learning

- Instance-based learning compares new problem instances with the instances seen in the training data.
- Classification or prediction is postponed when the new instance needs to be evaluated. Usually prediction stage is computationally intensive. Sometime called as lazy learning.
- Examples: k-nearest neighbors, support vector machines…

Regression Versus Classification Problems

- Supervised machine learning problems can be categorized as regression or classification problems.

- Regression problems: when the response variable is quantitative.
  - What is the customer demand of product ABC in the next month?

- Classification problems: response variable is qualitative or categorical.
  - Will a customer churn her service? (yes or no, binary response)
  - Will a customer default on a debit? (yes or no, binary response)

AGENDA

- Overview of Machine Learning
- Dataset and Scales of Measurement
- Assessing Model Accuracy

Data and Data Set

- Data are the facts collected, analyzed, and interpreted.

- The data collected in a particular data science project are commonly referred to as a data set.

Data Set: Elements, Variables, and Observations

- Elements/subjects: entities of interest
- Variables: characteristic of elements
- Observation: the set of measurements obtained for an element

Variables

car	mpg	cyl	hp	wt
Mazda RX4	21	6	110	2.62
Mazda RX4 Wag	21	6	110	2.875
Datsun 710	22.8	4	93	2.32
Hornet 4 Drive	21.4	6	110	3.215
Hornet Sportabout	18.7	8	175	3.44
Valiant	18.1	6	105	3.46

Element Names

Observations

Scales of Measurement

- Scale/level of measurement determines:
  - the amount of information contained in data
  - data summarization and analysis methods that are appropriate

- Four types of scales
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Qualitative / categorical data

Quantitative data

Nominal Scale

- Numerical values are just names or labels of the attribute
  - Ordering of these values is meaningless
  - No mathematical calculation (+, -, *, /) applicable

- For example:
  - Gender (1 = “Male”, 0 = “Female”)
  - Student ID (1,2,3...)
  - Department (1 = “BIT”, 2 = “CS”...)
  - Zip code (65401, 65402...)

Ordinal Scale

- Attributes can be ranked/ordered.

- For example:
  - Football team rank (1^{st}, 2^{nd}, 3^{rd}...)
  - Customer rating (1 = “Bad”, 2 = “OK”, 3 = “Excellent”)

Interval Scale

- Have all characteristics of ordinal scale
- Distance between attributes does have meaning.
- Ratios are not meaningful.

- For example:
  - Temperature
    - The distance from 40 – 60 is same as the distance from 60 – 80
    - 80 cannot be said as twice hot as 40
  - SAT Score
  - GMAT Score

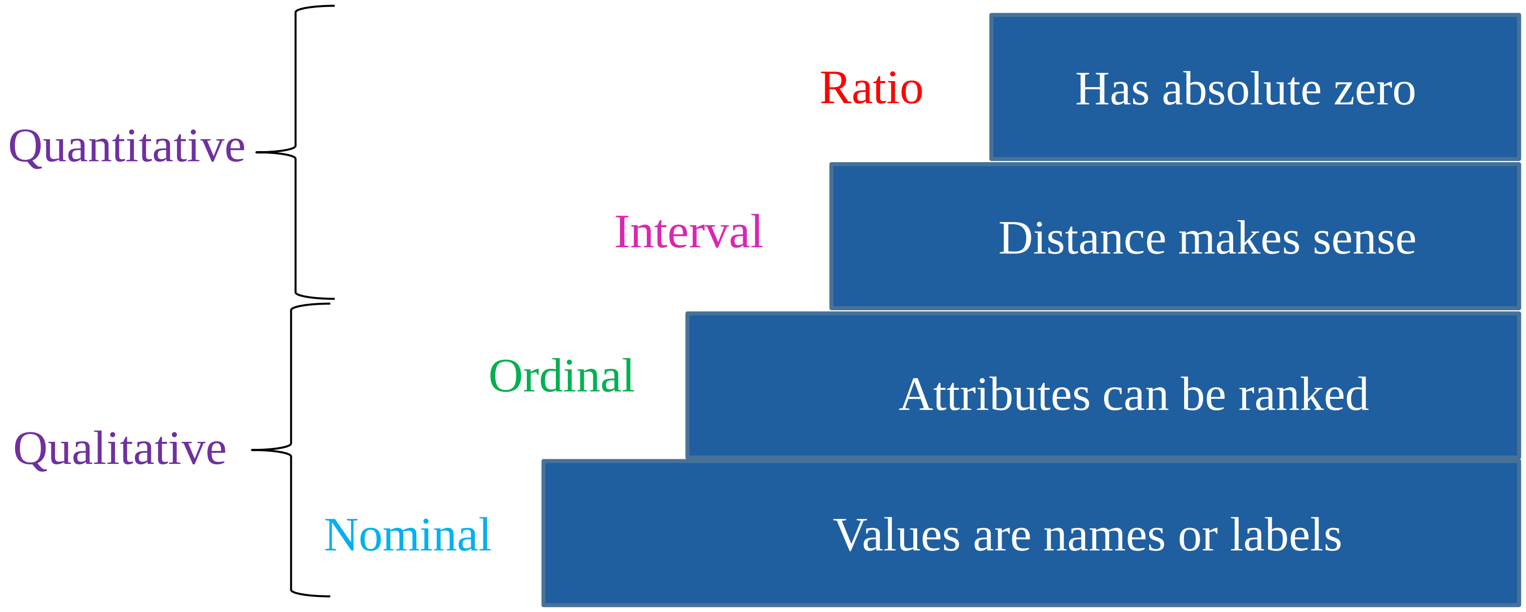
Ratio Scale

- Have all characteristics of interval scale
- A ratio of two values is meaningful.
- An absolute zero is meaningful.

- For example:
  - Weight
  - Height
  - Distance
  - Number of visits
  - Credit hours earned

Hierarchy of Measurement Scales

- A higher level scale contains all properties of its lower scale.
- From lower to higher levels, analysis tends to be more comprehensive. Improper use of lower level scales suffers information loss in the data
- In general, we prefer a higher scale of measurement than a lower one.



Exercise

- Decide the scales of measurement for the following columns:

A sales summary of two stores by operating hour

Store#	City	Hour	Sale
101	Rolla, MO	9	$1,000
101	Rolla, MO	10	$1,100
101	Rolla, MO	11	$1,200
102	St. Louis, MO	9	$3,000
102	St. Louis, MO	10	$3,300
102	St. Louis, MO	11	$4,000

Note: In the hour column, 9 means time between 9AM and 10AM.

Things Could be Tricky: Measurement Scale for Year

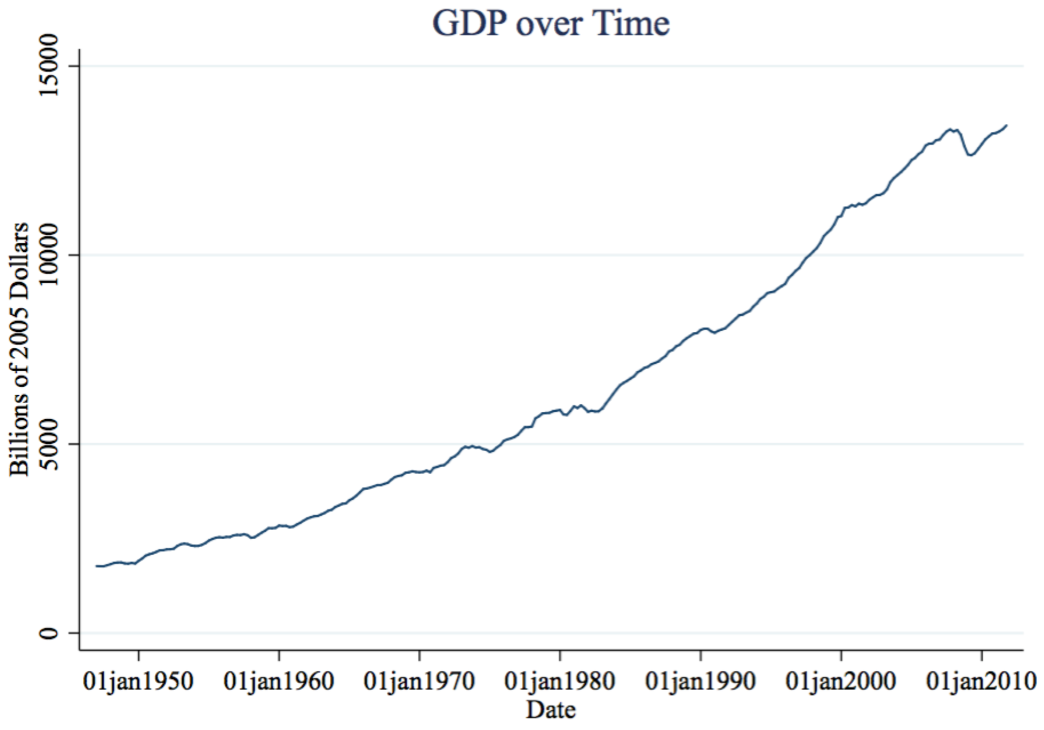
- GDP has an increasing trend over time

- We want to predict world GDP in 2020 using the following regression model:

$$GDP_{year} = \beta_0 + \beta_1 * year$$

What is the scale of measurement for *year*?

Ratio



Things Could be Tricky: Measurement Scale for Year

- It seems a consumer's spending is partly determined by his/her income.

- We collected a dataset of annual spending and revenue from 100 consumers across a 5-year period from 2011 to 2015.

- We want to estimate the effect of annual income on annual spending by controlling for possible time effect.

$$Spend_{i,t} = \beta_0 + \beta_1 * Income_{i,t} + \theta * year\_dummies$$

=> A panel regression

where t=1,2,...5, i = 1, 2,..., 100

What is the scale of measurement for *year*?

Nominal

AGENDA

- Overview of Machine Learning
- Dataset and Scales of Measurement
- Assessing Model Accuracy

Why Do We Need to Assess Model Accuracy?

- There are so many different statistical learning approaches.

- *There is no free lunch in statistics*: no one method dominates all others overall all possible dataset.

- On a particular dataset, one specific method may work best.

- Selecting the best approach can be a challenge in practice

Assessing Model Accuracy

- Measuring the quality of fit
- The classification setting
- The bias-variance trade-off

Measuring the Quality of Fit

- In regression setting, one commonly used measure is the *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 that a learning method gives for the *i*th observation in the training data

- More accurately, this is the <u>training MSE</u>.

- MSE is small if the predicted responses are very close to the true responses

Training MSE Vs Test MSE

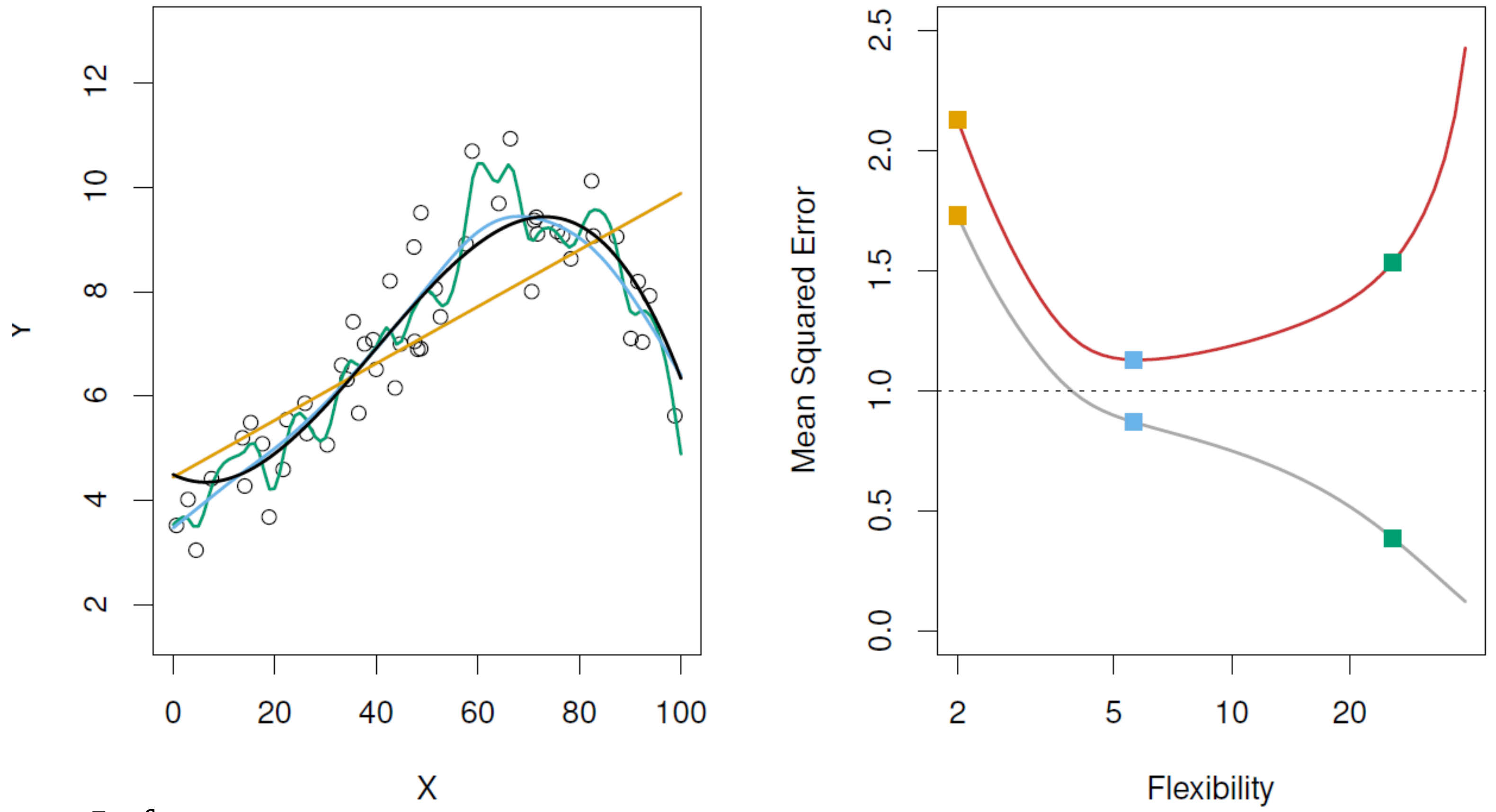
- Statistical learning methods are trying to minimize MSE on the training data.
  - For example, OLS (ordinary least squares) minimizes the MSE. But this does not guarantee OLS to be the best method for prediction.

- What we really care is how well the method performs on previously unseen test data.
  - Stock price prediction: We don't really care how well our method predicts last week's stock price. Instead, we care about how it will predict tomorrow's price.

- Smallest training MSE does not guarantee smallest test MSE.

- Thus, we should use test MSE to select models. The best model is the one that has the smallest test MSE.

Example of Training and Test MSEs



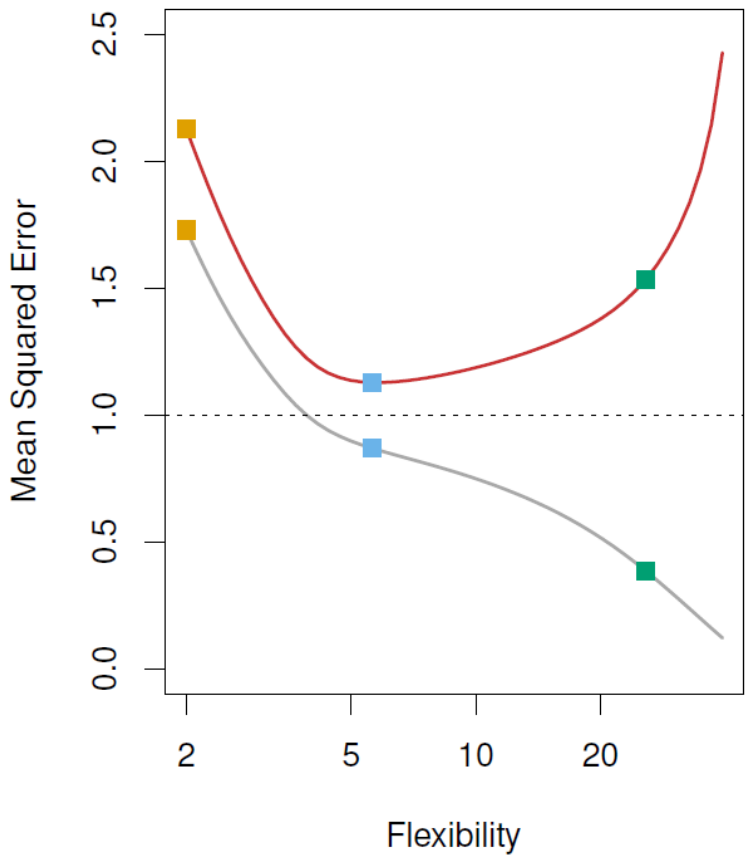
Left:

- Black: truth f(x)
- Orange: linear regression
- Blue: smoothing spline (less flexible)
- Green: smoothing spline (more flexible)

Right:

- Red: Test MSE
- Gray: Training MSE
- Dashed line: minimum possible test MSE

Example of Training and Test MSEs



- Red: Test MSE
- Gray: Training MSE
- Dashed line: minimum possible test MSE

- As flexibility (degrees of freedom) increases:
  - Training MSE declines monotonically;
  - Test MSE follows a U-shape.
- Overfitting: When a method yields a small training MSE but a large test MSE.
- Underfitting: When a method yields both a large training MSE and a large test MSE.
- Thus, our objective is to find a method with proper flexibility that fits the data just right.

The Classification Setting

- For classification problems, we can use *error rate* to assess the model accuracy

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

where $I(y_i \neq \hat{y}_i)$ is an indicator function

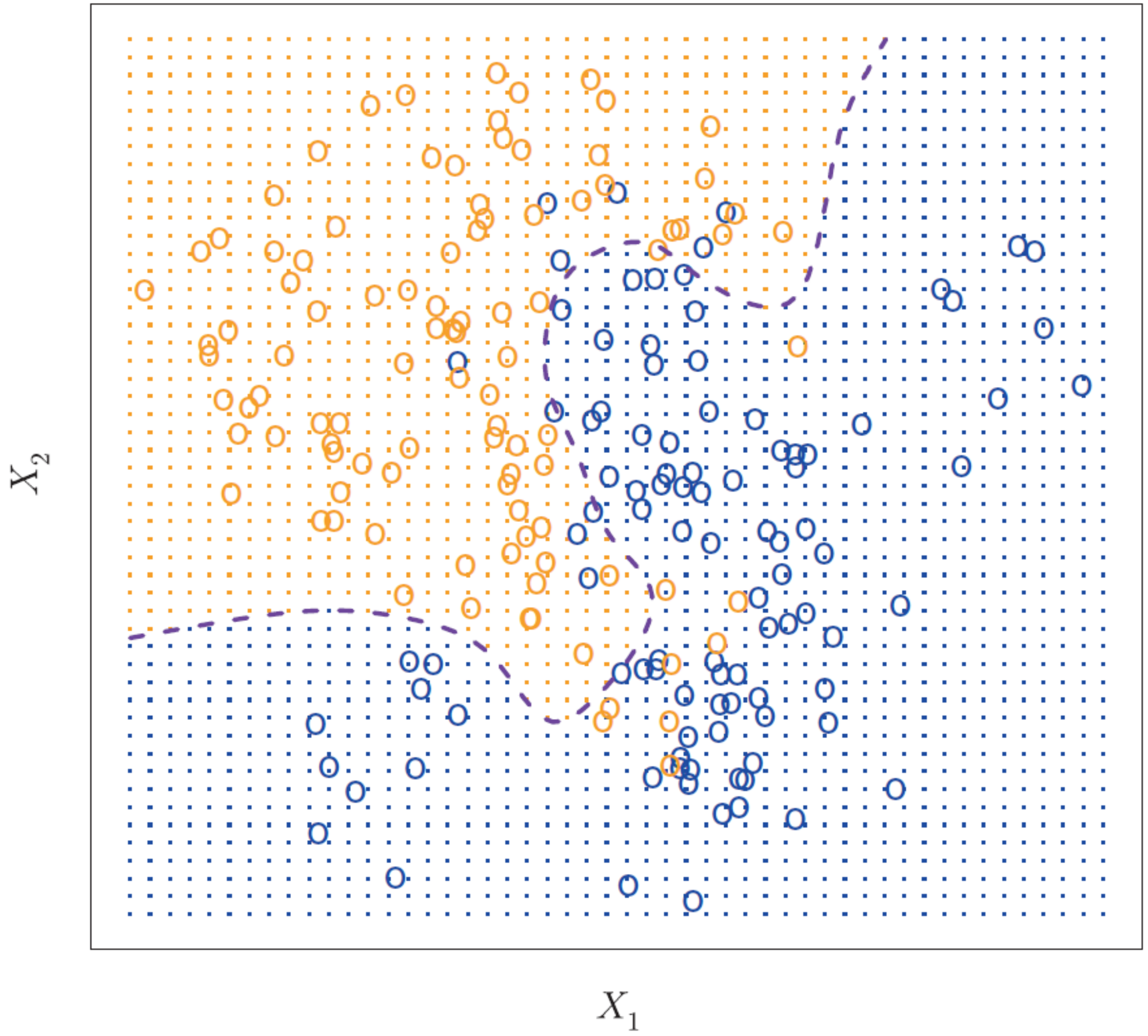
$$I(y_i \neq \hat{y}_i) = \begin{cases} 1, if\ the\ condition(y_i \neq \hat{y}_i)\ is\ true \\ 0, if\ the\ condition(y_i \neq \hat{y}_i)\ is\ false \end{cases}$$

The Bayes Classifier

- In order to minimize test error rate, on average, a classifier can assign each observation to the most likely class given its predictor values.

- Bayes classifier works in a simple way:
  - First, calculates conditional probability $\Pr(Y = j|X = x_0)$;
  - Then, assign the class j for which the conditional probability is largest.

Bayes Optimal Classifier – Unattainable Gold Standard

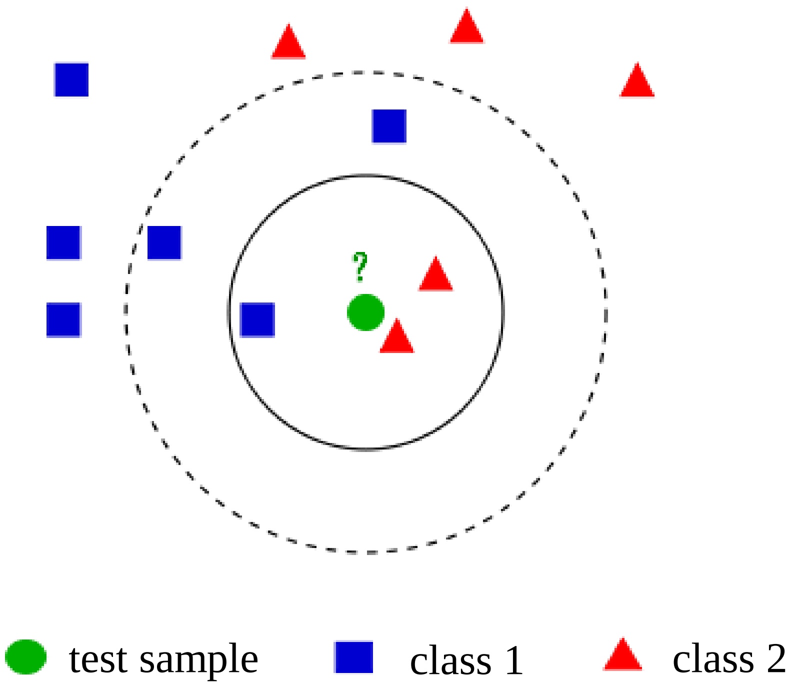


- Two predictors X_1 and X_2
- Two classes: Orange and Blue
- Dashed line: Bayes decision boundary

For real data, conditional probability is unknown, so that it's impossible to implement Bayes classifier.

K-Nearest Neighbors (KNN)

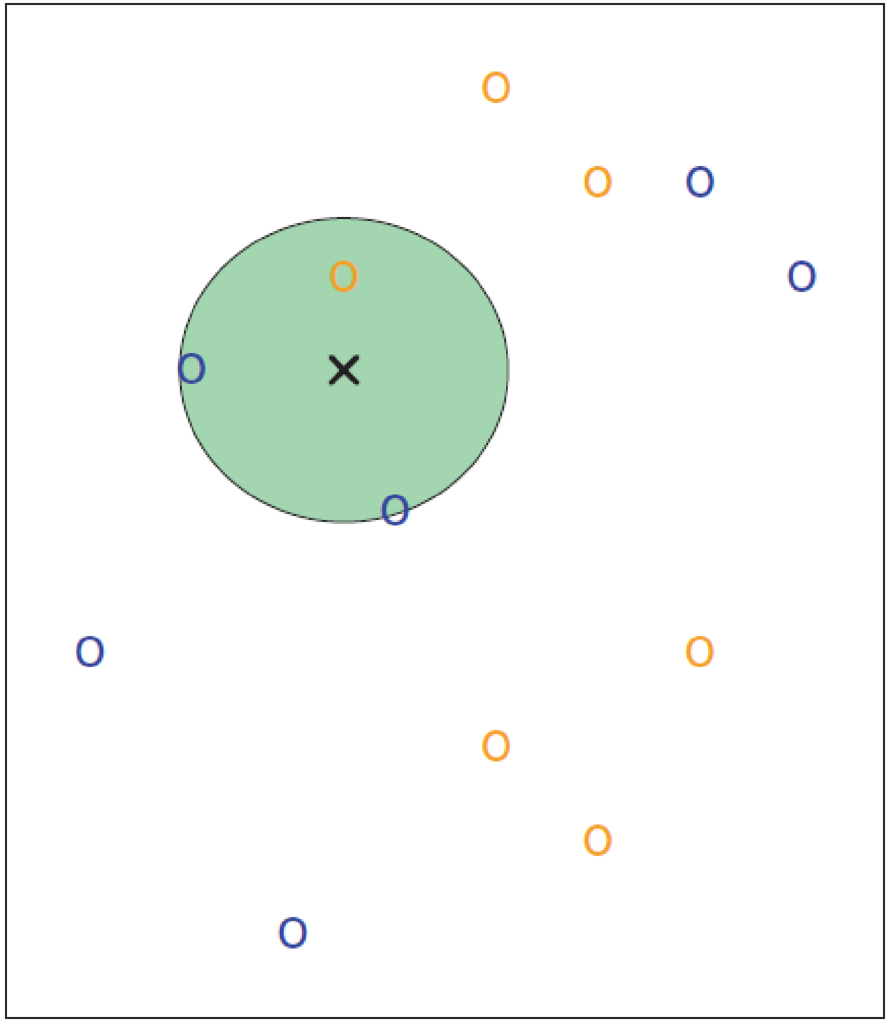
- Many approaches including KNN tries to estimate the conditional distribution of Y given X.
- KNN contains a parameter k
  - Large values of k reduce the effect of noise, but make boundaries between classes less distinct.

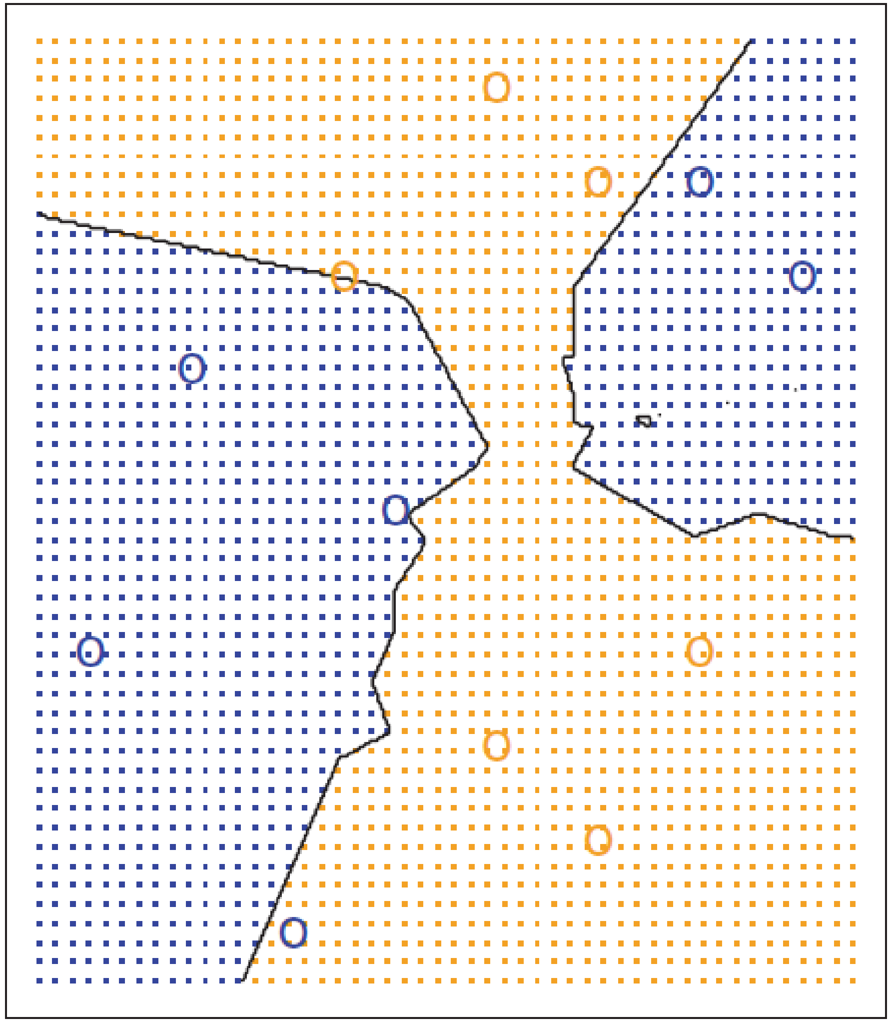


If k = 3, the test sample is classified as class 2;

If k = 5, class 1 is assigned.

KNN Example: k=3

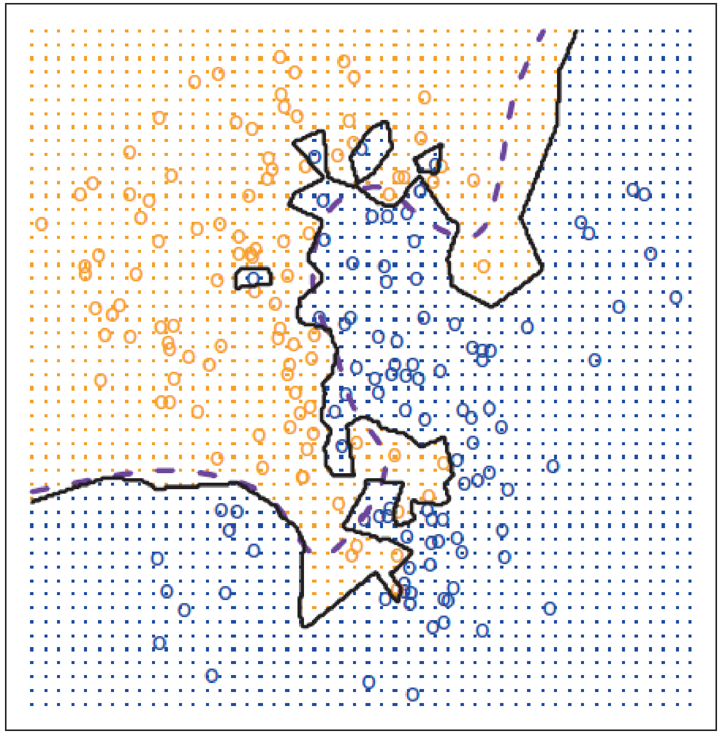




KNN Simulated Data

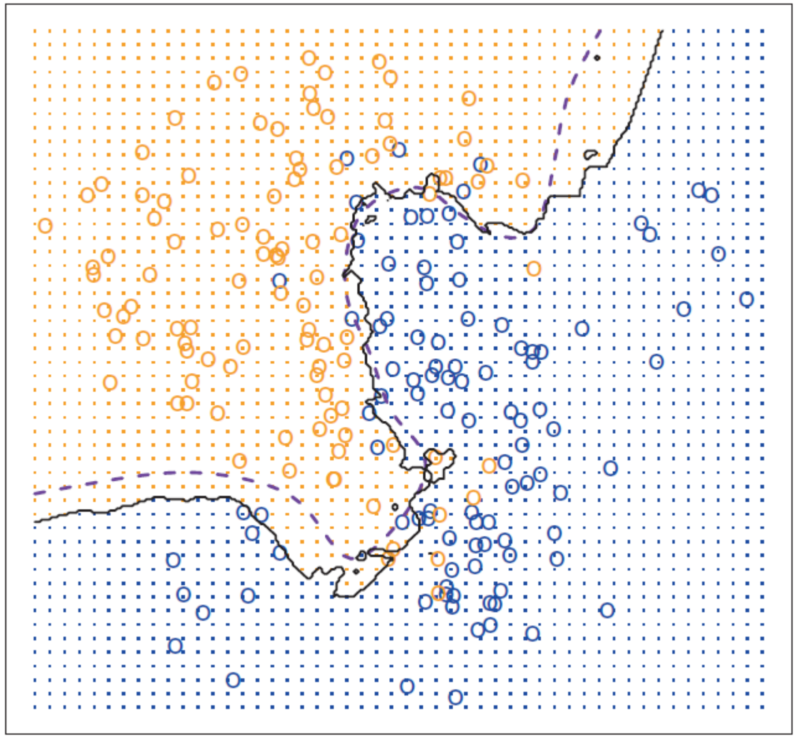
Overly flexible

KNN: K=1



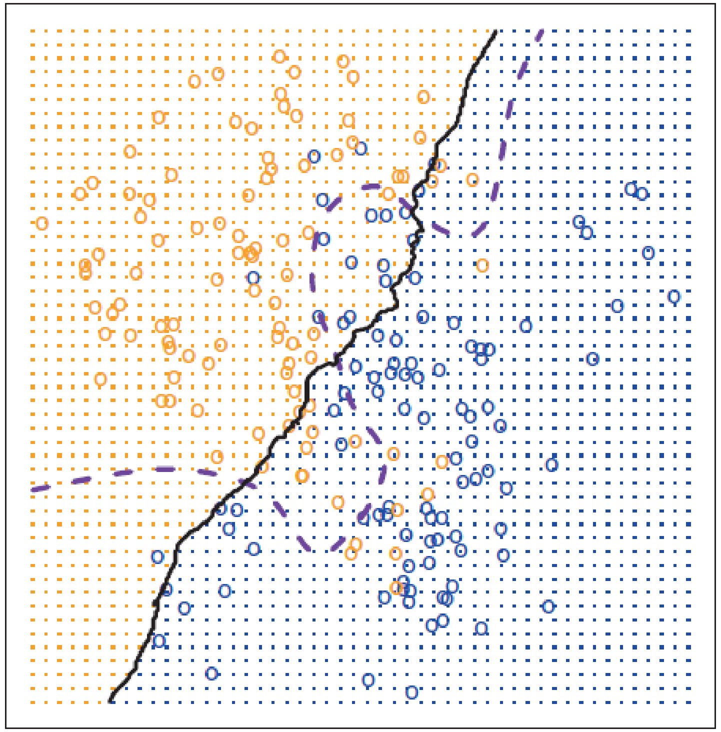
Looks right flexible

KNN: K=10



Insufficiently flexible

KNN: K=100

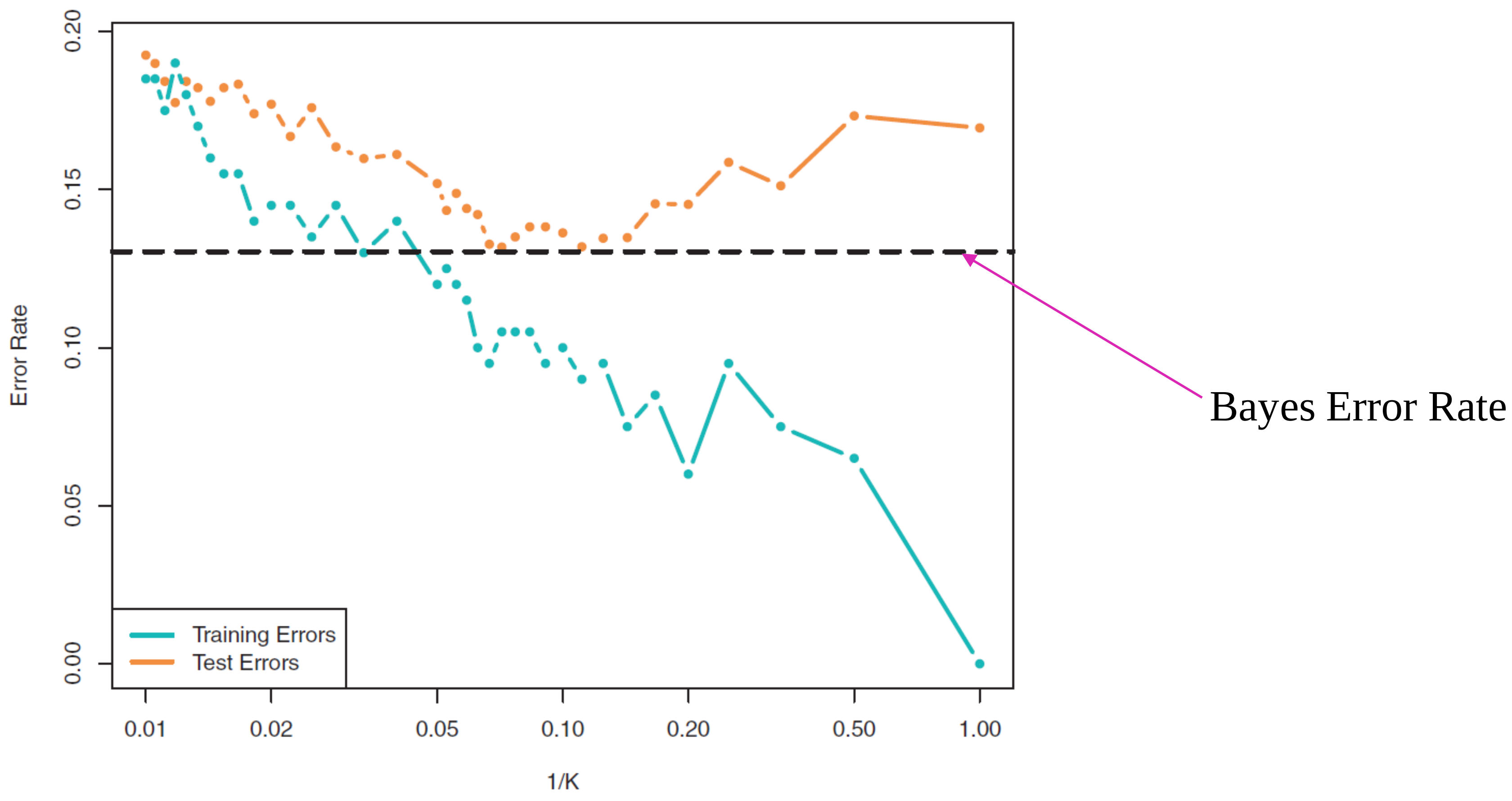


X_2

X_1

- Purple dashed line: Bayes decision boundary
- Black curve: KNN decision boundary

KNN Training and Test Error Rates



Choosing the correct level of flexibility is critical to the success of any statistical learning method.

The bias-variance tradeoff, and the resulting U-shape in the test error, can make this a difficult task.

Select the "Optimal" Model: Bias-Variance Tradeoff

- Bias is an error from improper assumptions in the learning algorithm.

$$Bias = E[\hat{f}(x)] - f(x)$$

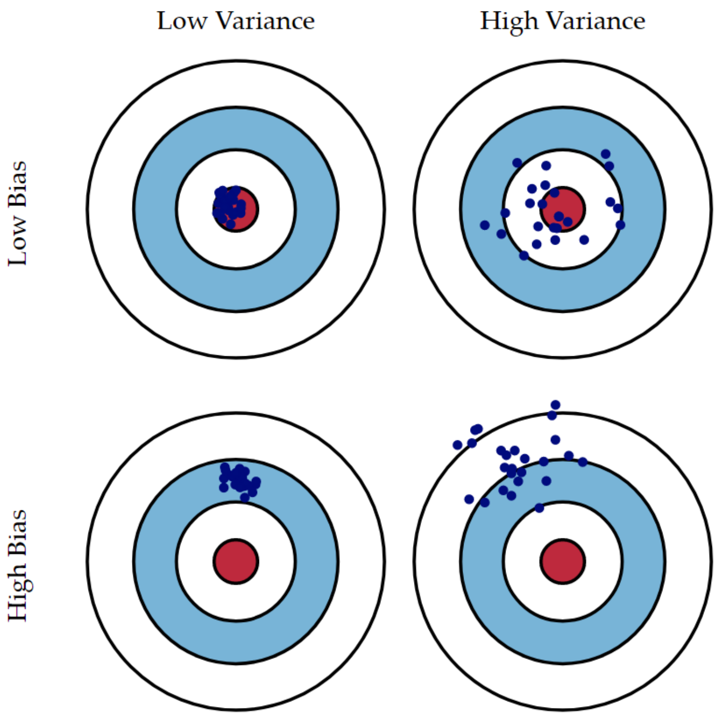
- Variance is an error from sensitivity to small fluctuations in the training set.

$$Variance = E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right]$$

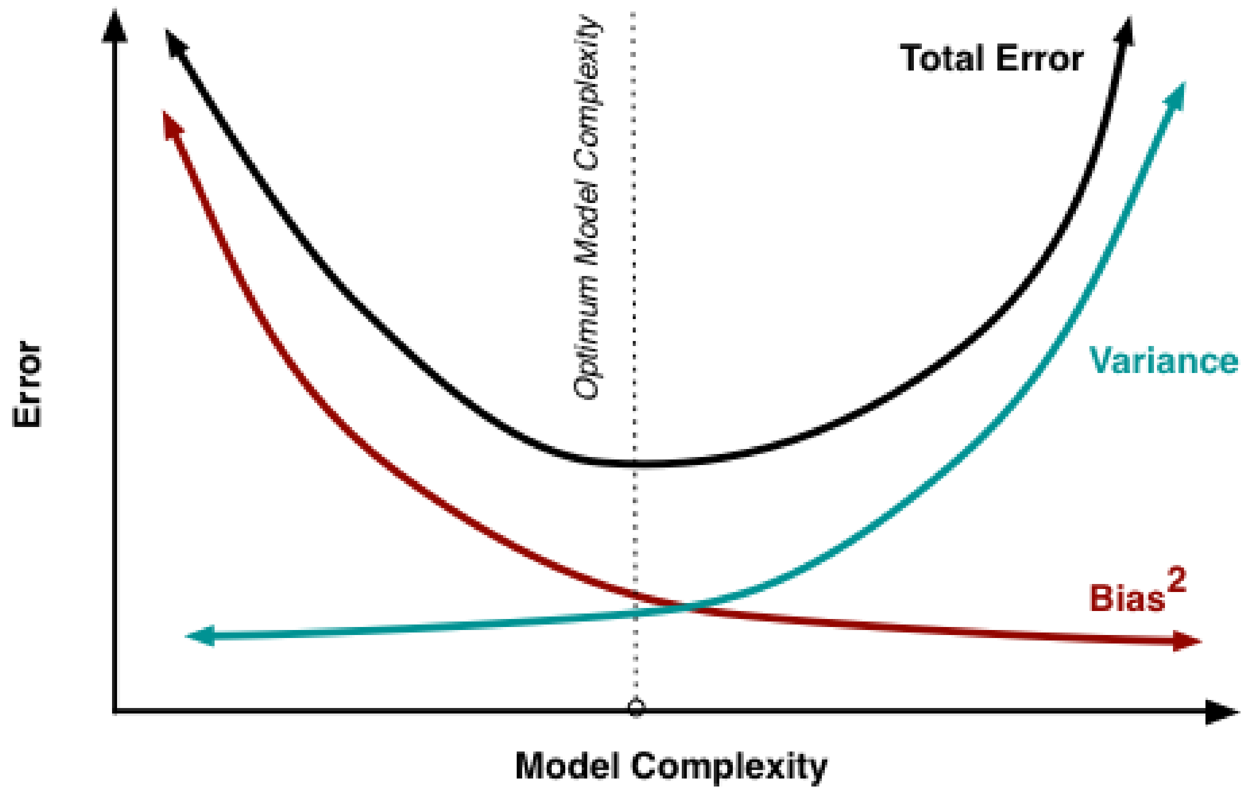
- Squared estimation error can be decomposed as:

$$\begin{aligned}
Error(x) &= E\left[\left(Y - \hat{f}(x)\right)^2\right] \\
&= E\left[\left(f(x) + \varepsilon - \hat{f}(x)\right)^2\right] = E\left[\left(f(x) - \hat{f}(x)\right)^2\right] + E(\varepsilon^2) \\
&= E\left[f(x)^2 - 2f(x)\hat{f}(x) + \hat{f}(x)^2\right] + \sigma_\varepsilon^2 \\
&= f(x)^2 - 2f(x)E[\hat{f}(x)] + E[\hat{f}(x)^2] + \sigma_\varepsilon^2 \\
&= \left(f(x) - E[\hat{f}(x)]\right)^2 + \left(E[\hat{f}(x)^2] - E[\hat{f}(x)]^2\right) + \sigma_\varepsilon^2 \\
&= \left(f(x) - E[\hat{f}(x)]\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \sigma_\varepsilon^2
\end{aligned}$$

$$Error(x) = \underbrace{Bias^2 + Variance}_{Reducible\ Error} + Irreducible\ Error$$



Select the "Optimal" Model: Bias-Variance Tradeoff



Under-fitting: high bias, low variance

Over-fitting: low bias, high variance

Given imperfect models and finite data, there is a tradeoff between minimizing bias and minimizing variance.

More Flexible Model vs. Less Flexible Model

- A more flexible model can better fit non-linear relationship, thus decreasing bias;
- But a more flexible model may also fit the noise (rather than signal) too closely, thus increasing variance;
- Also the results of a more flexible model are more difficulty to explain.

- A more flexible model tends to be better when:
  - *n* is very large, *p* is small;
  - Non-linear relationship between predictors and response;
  - Emphasis on prediction rather than interpretation.

RECAP: OUTLINE

- (I) Overview of machine learning (ML)
  1. What is learning?
  2. Practical definition of ML
  3. ML model estimation methods: parametric, nonparametric
  4. Types of ML

- (II) Scale of measurement

  Nominal, ordinal, interval, ratio

- (III) Model accuracy
  1. Regression setting: MSE, training MSE, test MSE
  2. Classification setting: Error rate
  3. Bias variance tradeoff

Q & A

Assignments

- Homework 1 (Due Jan 24)

- Reading (Due Jan 25)
  - Book Chapter 2 Section 2.3 and Try the Code
  - "An Introduction to R" Chapters 1, 2, 3, 4, 5, 6, 9,10; pg 2-29, 40-50

- Install R and RStudio to your PC (Due Jan 25)