CSCE 633: Machine Learning

Lecture 1: Introduction

Texas A&M University

8-26-19

Welcome to CSCE 633!

- About this class
  - Who are we?
  - Syllabus
  - Text and other references
- Introduction to Machine Learning

Goals of this lecture

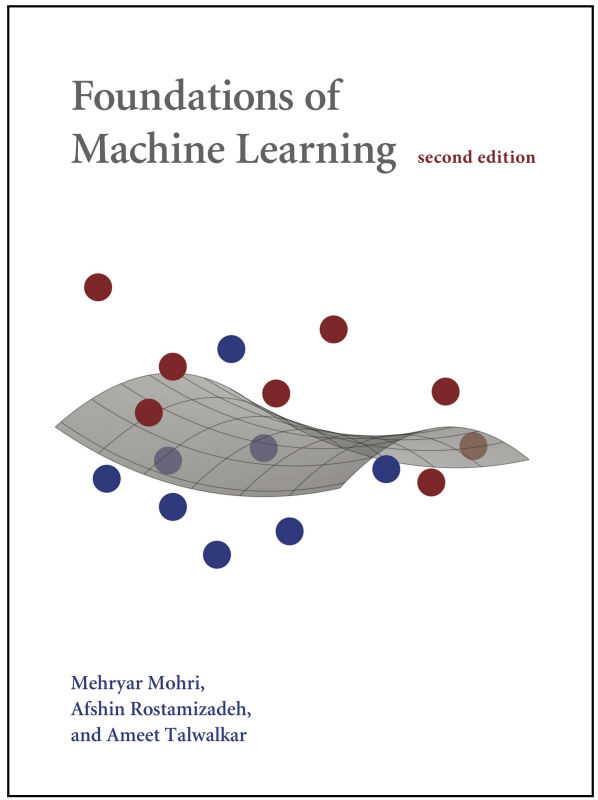
- Course Structure
- Machine Learning Common Tasks
- Terminology
- Learning Scenarios
- Mathematical Framework
- Generalization/Empirical Error
- Bayes Error
- Bias Variance Tradeoff
- No Free Lunch Theorem

Who are we?

- Instructor
  - Bobak Mortazavi
  - bobakm@tamu.edu - Please put [CSCE 633] in subject line
  - Office Hours: W 10-11a R 11-12p 328A HRBB
- TA
  - Arash Pakbin
  - a.pakbin@tamu.edu
  - Office Hours: M 10:15-11:15a W 4:30-5:30p 320 HRBB
- Discussion Board on eCampus

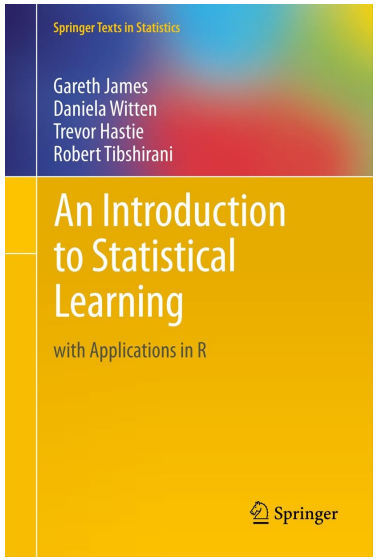
Syllabus, Schedule, and eCampus

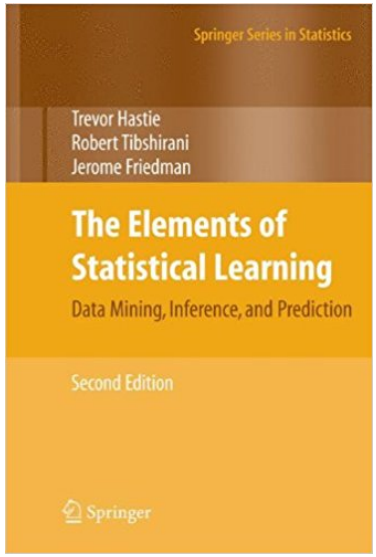
Text Book



Textbook for this course

Other useful references





Prerequisite Topics

The topics you need to know for this class from probability theory and matrix operations are:

- Linear Algebra Review (Appendix A of the book)
- Probability Review (Appendix C of the book)

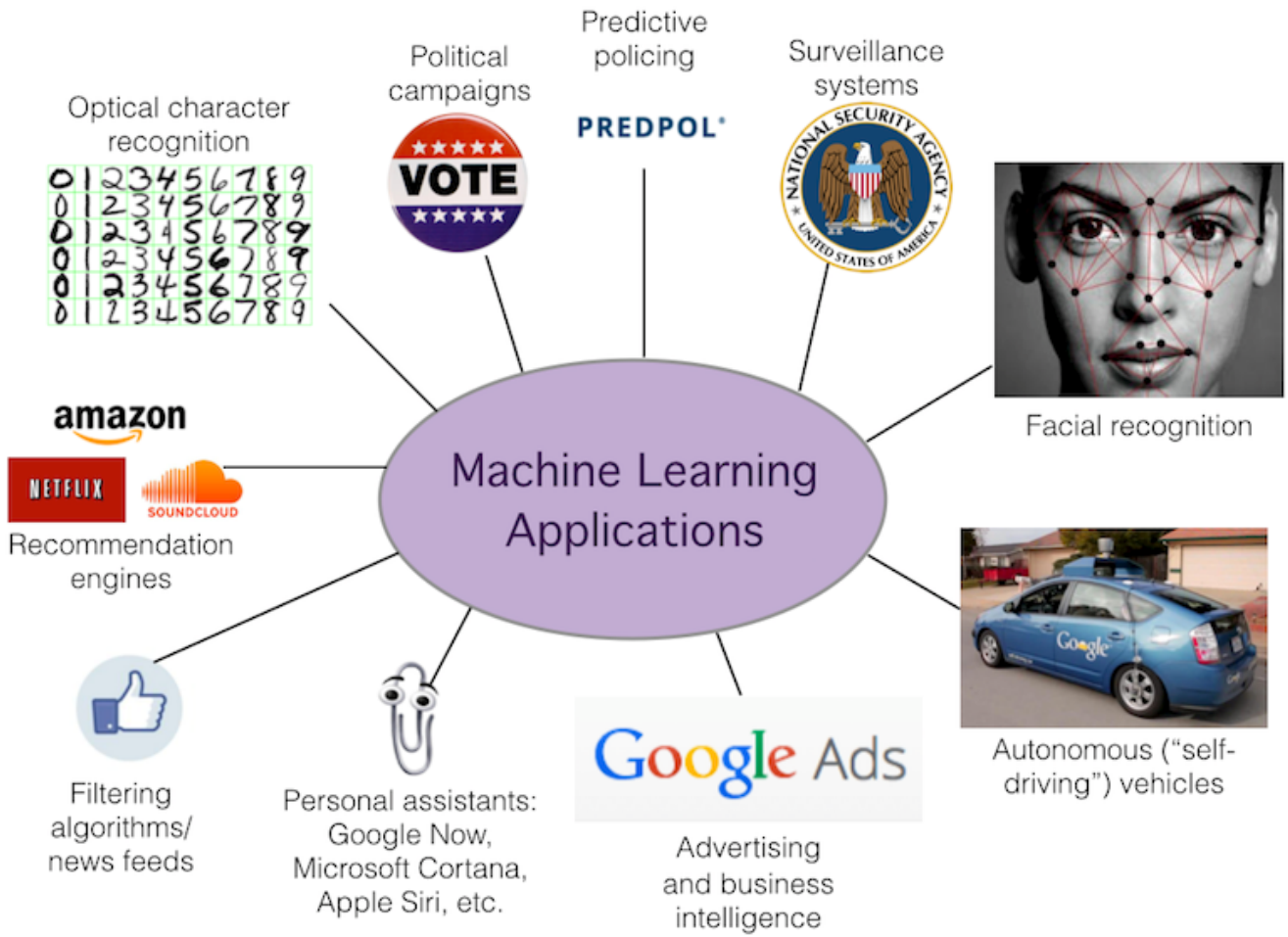
Assignments

- Homework will be compiled in Latex - you will turn in PDFs
- Programming for homework will be in R or Python.
- for Projects - if you need to use a different language please explain why
- Project teams will be 2-3 people per team.
- Project proposals will be a 2 page written document
- Project report will be an 8 page IEEE conference-style report

Topics

- About this class
- Introduction to Machine Learning
  - What is machine learning?
  - Models and accuracy
  - Some general supervised learning

What is machine learning?

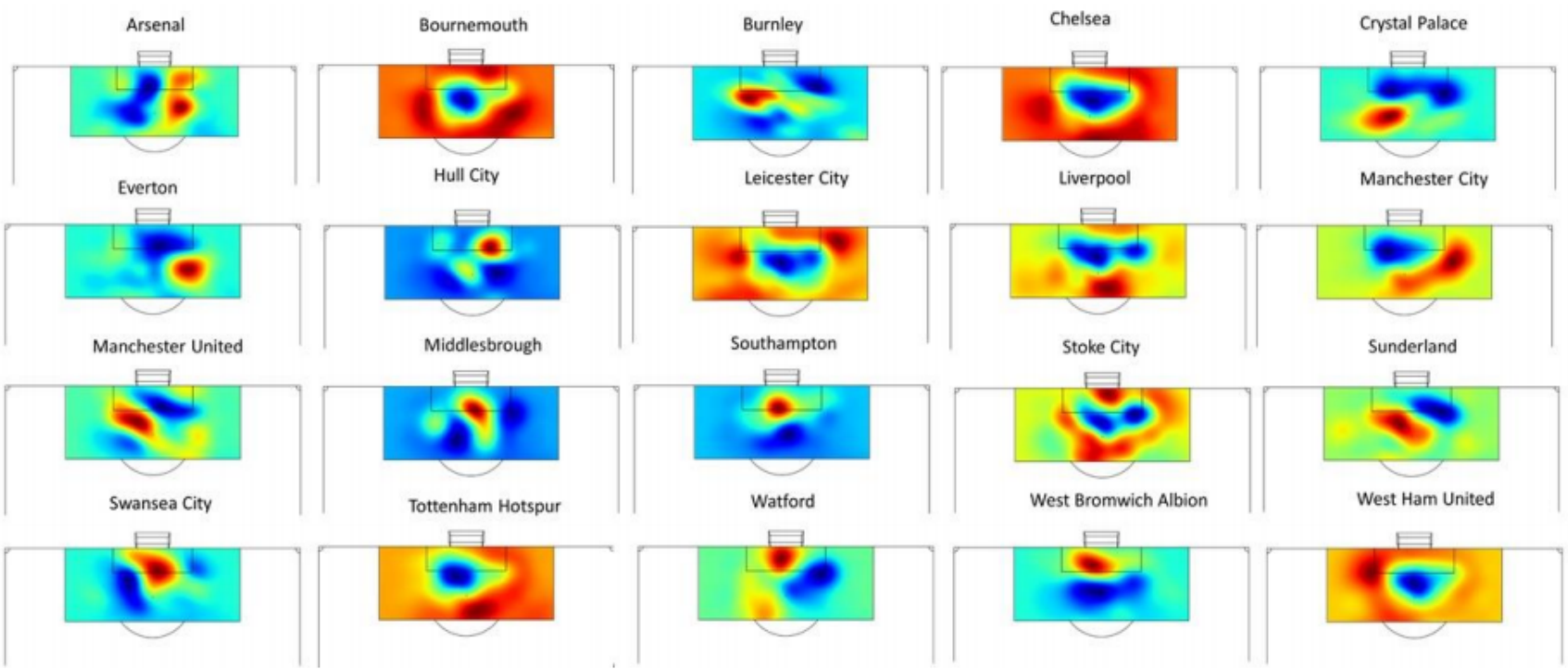


What is machine learning?

Big Data! 40 billion web pages, 100 hours of videos on YouTube every minute, genomes of 1000s of people.

Murphy defines **Machine learning** as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kindso f decision making under uncertainty (such as planning how to collect more data!).

Some cool examples!



Power P, Hobbs J, Ruiz H, Wei X, Lucey P. Mythbusting Set-Pieces in Soccer. MIT Sloan Sports Conference 2018

Some cool examples!

Offensive Efficiency

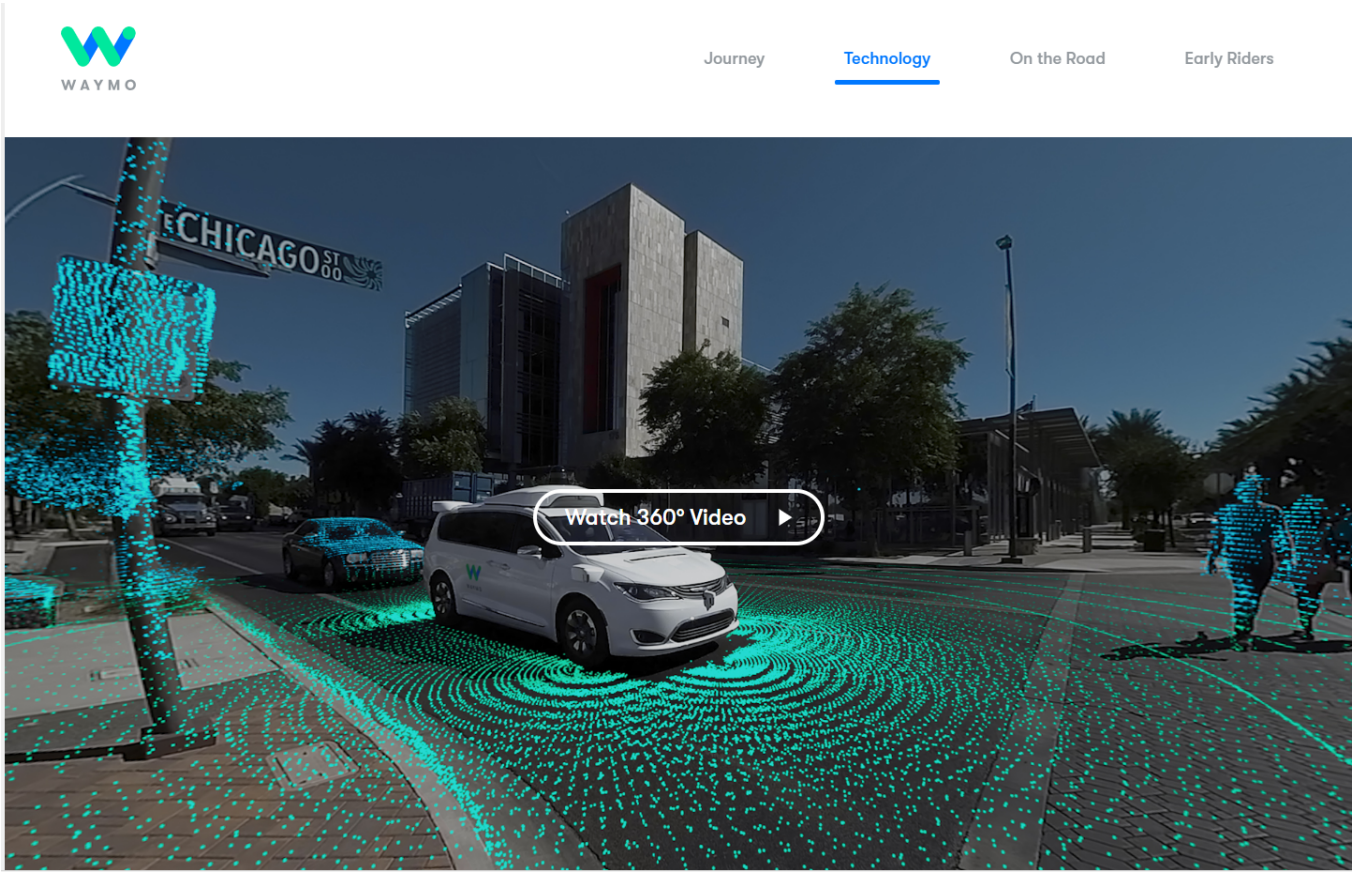
Ranking	Team	Goals Scored	Shots Taken	Efficiency
1	West Bromwich Albion	16	117	13.68%
2	Swansea City	13	119	10.92%
3	Chelsea	15	144	10.42%
4	Hull City	9	94	9.57%
5	Tottenham Hotspur	15	165	9.09%
6	West Ham United	10	120	8.33%
7	Manchester City	12	161	7.45%
8	Bournemouth	9	123	7.32%
9	Middlesbrough	6	85	7.06%
10	Liverpool	11	161	6.83%
11	Crystal Palace	9	133	6.77%
12	Southampton	8	124	6.45%
13	Leicester City	6	111	5.41%
14	Watford	7	130	5.38%
15	Manchester United	7	130	5.38%
16	Stoke City	6	118	5.08%
17	Arsenal	7	142	4.93%
18	Everton	6	125	4.80%
19	Burnley	5	137	3.65%
20	Sunderland	2	83	2.41%

Defensive Efficiency

Ranking	Team	Goals Conceded	Shots Conceded	Efficiency
1	Bournemouth	6	143	4.20%
2	Chelsea	4	92	4.35%
3	Arsenal	5	114	4.39%
4	Liverpool	5	97	5.15%
5	Burnley	10	188	5.32%
6	Tottenham Hotspur	5	92	5.43%
7	West Ham United	6	105	5.71%
8	Manchester City	6	97	6.19%
9	Manchester United	7	107	6.54%
10	West Bromwich Albion	8	119	6.72%
11	Middlesbrough	10	146	6.85%
12	Stoke City	11	155	7.10%
13	Sunderland	14	178	7.87%
14	Everton	10	125	8.00%
15	Watford	12	146	8.22%
16	Swansea City	8	96	8.33%
17	Leicester City	11	128	8.59%
18	Southampton	12	129	9.30%
19	Crystal Palace	12	122	9.84%
20	Hull City	17	143	11.89%

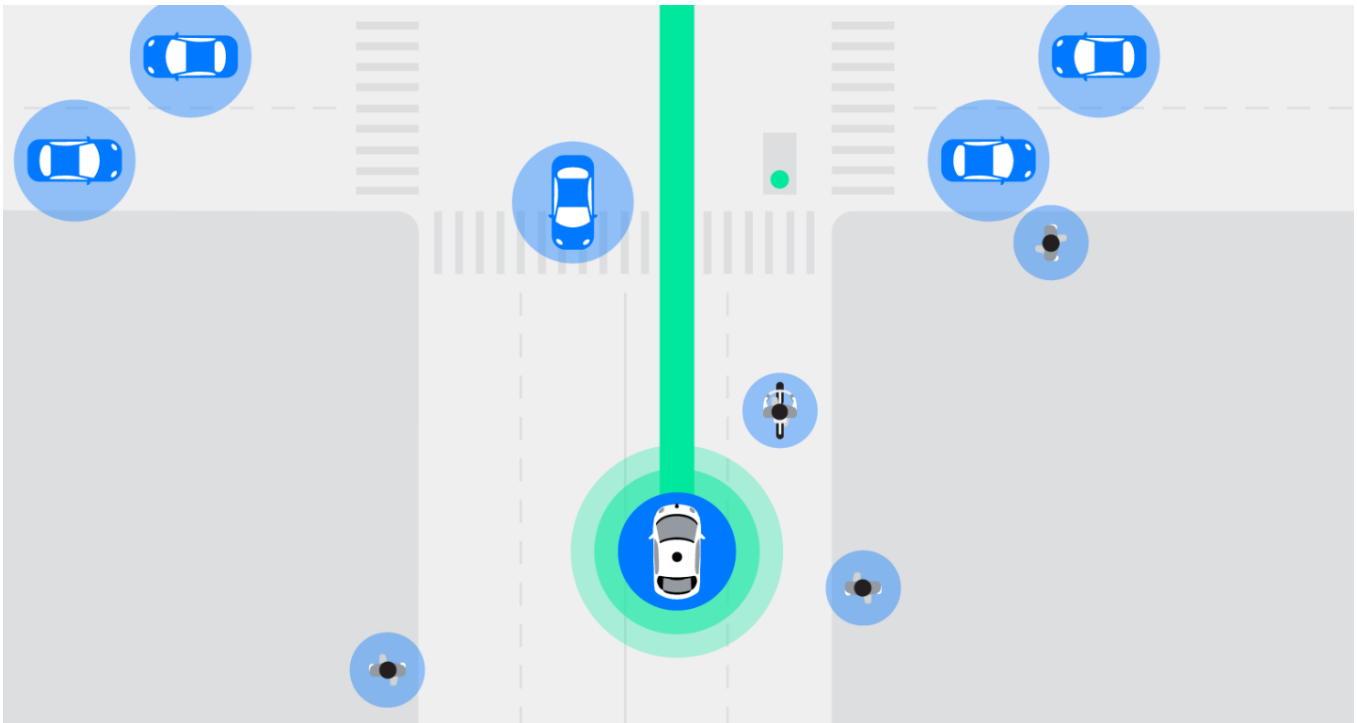
Power P, Hobbs J, Ruiz H, Wei X, Lucey P. Mythbusting Set-Pieces in Soccer. MIT Sloan Sports Conference 2018

Some cool examples!



Waymo.com/tech

Some cool examples!



Waymo.com/tech

Some cool examples!

Waymo.com/tech

Standard Machine Learning Tasks

Machine learning tasks mostly fall into the following tasks:

- Classification - assigning a category to each item, i.e., text classification and speech recognition
- Regression - learning a real value for each item, i.e., stock value prediction
- Clustering - learning how to partition a set of items into homogeneous subsets, i.e., identifying communities within large groups of people
- Dimensionality Reduction - learning a transformation from the initial representation into a lower-dimensional representation, i.e., digital image preprocessing

Machine Learning Terminology (buzzwords!)

The list of definitions and terminology commonly used in ML are as follows:

- Examples - instances of data for either learning or evaluation
- Features - set of attributes of a vector, often a vector
- Labels - values/categories assigned to examples
- Hyperparameters - free parameters not determined in training, but rather specified as inputs
- Training sample - examples used in training
- Validation sample - examples used in parameter tuning
- Test sample - examples used for performance evaluation

Machine Learning Terminology (buzzwords!)

- Loss function - A function measuring the difference between a predicted label and the true label. This function is to be minimized in the optimization stage.
- Hypothesis set - A set of functions which map the feature vectors to the set of labels.

Learning Scenarios

Based on how the training data is available to the learner, several training scenarios can occur:

- Supervised Learning - the learner receives a set of labeled examples, making predictions for all unseen points. Most common scenario. Could be, among others, classification or regression.
- Unsupervised Learning - the training data is unlabeled and the learner makes predictions for unseen points. Examples are clustering and dimensionality reduction.
- Semi-supervised Learning - the training sample contains unlabeled AND labeled samples, making predictions for unseen samples. Common when labels are expensive to obtain.

Learning Scenarios

- On-line Learning - training and testing phases are intermixed. The data is made available to the learner over time. The objective is to minimize the cumulative loss over all rounds.
- Reinforcement Learning - training and testing phases are intermixed. The learner interacts with/affects the environment. Some actions of the learner lead to being rewarded. Objective is to maximize the rewards based on a course of actions. Exploration vs exploitation dilemma.
- Active Learning - the learner interactively collects training examples, querying an oracle for requesting new labels. Used when labels are expensive to obtain.

Mathematical Framework of Learning

A learning framework can be mathematically laid out as:

- $\mathcal{X}$: all possible instances (input space)
- $\mathcal{Y}$: all possible labels/target values
- Concept $c : \mathcal{X} \to \mathcal{Y}$: a mapping from $\mathcal{X}$ to $\mathcal{Y}$
- Concept class: a set of concepts we wish to learn denoted by $\mathcal{C}$
- The learning problem would be:
  - The learner assumes a **fixed** set of concepts $\mathcal{H}$, called *hypothesis set* (not necessarily $\mathcal{C}$)
  - Input samples are $S = (x_1, ..., x_m)$ with labels $S = (c(x_1), ..., c(x_m))$ drawn i.i.d. according to data $\mathcal{D}$, based on a target concept $c \in \mathcal{C}$ to be learned
  - The learning task would be selecting a hypothesis $h_s \in \mathcal{H}$ having a small *generalization error* $R(h)$.

Generalization Error vs Empirical Error

- **Generalization Error**: Given a hypothesis $h \in \mathcal{H}$, a target concept $c \in \mathcal{C}$, and an underlying distribution $\mathcal{D}$, the generalization error of h is defined by:

$$R(h) = \mathop{\mathbb{P}}_{x \sim \mathcal{D}}[h(x) \neq c(x)] = \mathop{\mathbb{E}}_{x \sim \mathcal{D}}[\mathbf{1}_{h(x) \neq c(x)}] \quad (1)$$

- The generalization error cannot be computed which gives rise to:
- **Empirical Error**: Given a hypothesis $h \in \mathcal{H}$, a target concept $c \in \mathcal{C}$, and a sample $S = (x_1, ..., x_m)$, the empirical error of h is defined by:

$$\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{h(x) \neq c(x)} \quad (2)$$

经验误差与泛化误差、偏差与方差、欠拟合与过拟合、交叉验证
https://blog.csdn.net/zhihua_oba/article/details/78684257

Bayes Error

- The data distribution $\mathcal{D}$ is defined over $\mathcal{X} \times \mathcal{Y}$, and the labeled sample $\mathcal{S}$ drawn i.i.d according to distribution $\mathcal{D}$:

$$S = ((x_1, y_1), ..., (x_m, y_m)) \tag{3}$$

- In a general scenario, the label for a specific input is not unique and is a probabilistic function of it.
- In such a *stochastic scenario*, there is a non-zero error for any hypothesis.

Bayes Error

Check Bayes' theorem later

- **Bayes Error**: Given a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, the Bayes error $\mathcal{R}^*$ is defined as the infimum of the errors achieved by measurable functions $h : \mathcal{X} \to \mathcal{Y}$:

infimum

$$\mathcal{R}^* = \inf_h R(h) \tag{4}$$

- A hypothesis h with $R(h) = \mathcal{R}^*$ is called Bayes hypothesis or Bayes classifier.

Advertising Example

Assume we are trying to improve sales of a product through advertising in three different markets, tv, newspapers, and online/mobile. We define the advertising budget as:

- input variables x_i: consist of TV budget $x_{i,1}$, newspaper budget $x_{i,2}$, and mobile budget $x_{i,3}$. (also known as predictors, independent variables, features, covariates, or just variables).
- output variable y_i: sales numbers. (also known as the response variable, or dependent variable)
- Can generalize input predictors to $x_{i,1}, x_{i,2}, \cdots, x_{i,m} = x_i$

Advertising Example

Assume we are trying to improve sales of a product through advertising in three different markets, tv, newspapers, and online/mobile. We define the advertising budget as:

- input variables x_i: consist of TV budget $x_{i,1}$, newspaper budget $x_{i,2}$, and mobile budget $x_{i,3}$. (also known as predictors, independent variables, features, covariates, or just variables).
- output variable y_i: sales numbers. (also known as the response variable, or dependent variable)
- Can generalize input predictors to $x_{i,1}, x_{i,2}, \cdots, x_{i,n} = x_i$
- Model for the relationship is then: $Y = f(X) + \epsilon$
- f is the systematic information, ϵ is the random error

Why estimate f?

- **Prediction**
- **Inference**

Why estimate f?

- **Prediction** - In many situations, x_i is available to us but y_i is not, would like to estimate y_i as accurately as possible.
- **Inference**

Why estimate f?

- **Prediction** - In many situations, x_i is available to us but y_i is not, would like to estimate y_i as accurately as possible.
- **Inference** - We would like to understand how y_i is affected by changes in x_i.

Predicting f?

Create a $\hat{f}(x)$ that generates a $\hat{y}$ that estimates y as closely as possible by attempting to minimize the reducible error, accepting that there is some irreducible error generated by ϵ that is independent of x.

Why is there irreducible error? (hint: remember Bayes error!)

Estimating f

Assume we have an estimate $\hat{f}$ and a set of predictors x, such that $\hat{y} = \hat{f}(x)$. Then we can calculate the expected value of the average error as:

$$\mathbb{E}(y - \hat{y})^2 = \mathbb{E}[f(x) + \epsilon - \hat{f}(x)]^2 = [f(x) - \hat{f}(x)]^2 + Var(\epsilon)$$

The focus of this class is to learn techniques to estimate f to minimize reducible error.

How do we estimate f

- Need: Training Data to learn from

How do we estimate f

- Need: Training Data to learn from
- Training samples $S = (x_1, y_1), ..., (x_m, y_m)$, where $x_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,n})^T$ for n subjects and p dimensions.
- Notations could be different across different books
- y can be:
  - categorical or nominal $\rightarrow$ the problem we solve is classification or pattern recognition
  - continuous $\rightarrow$ the problem we solve is regression
  - categorical with order (grades A, B, C, D, F) $\rightarrow$ the problem we solve is ordinal regression

统计学中的 DATA:Nominal,Ordinal, Interval and Ratio怎么区别
https://blog.csdn.net/qq_35286745/article/details/77896061

Types of Data: Nominal, Ordinal, Interval/Ratio - Statistics Help
https://www.youtube.com/watch?v=hZxnzfnt5v8

Parametric Methods

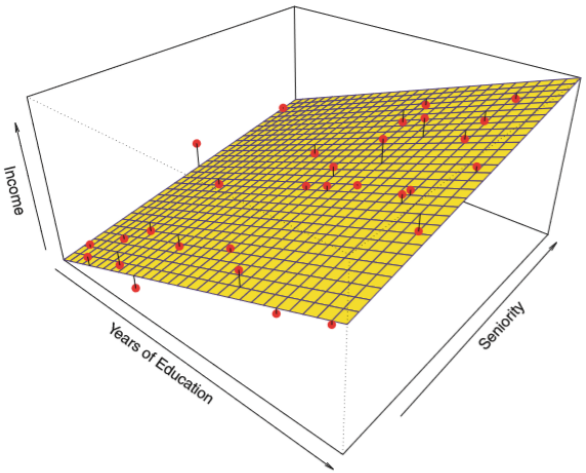
- 2-step model based approach to estimate f.
- Step 1 - make an assumption about the form of f. For example, that f is linear
- $f(X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
- Advantage - only have to estimate $p + 1$ coefficients to model f
- Step 2 - fit or train a model/procedure to estimate β
- The most common method for this is Ordinary Least Squares.

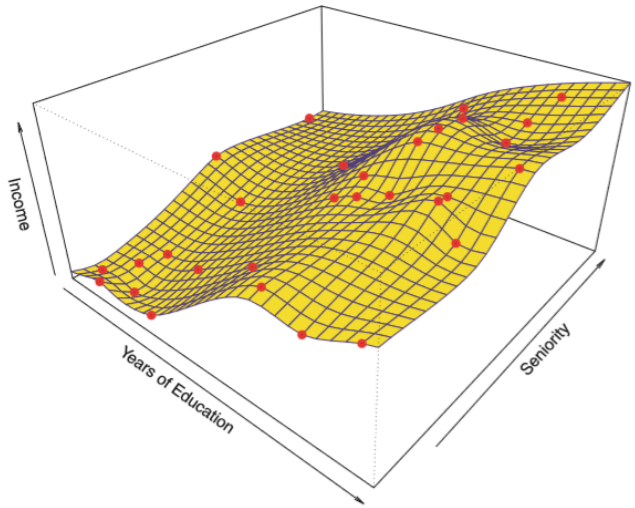
Solution to minimize the error

最小二乗法
https://www.matongxue.com/madocs/818/

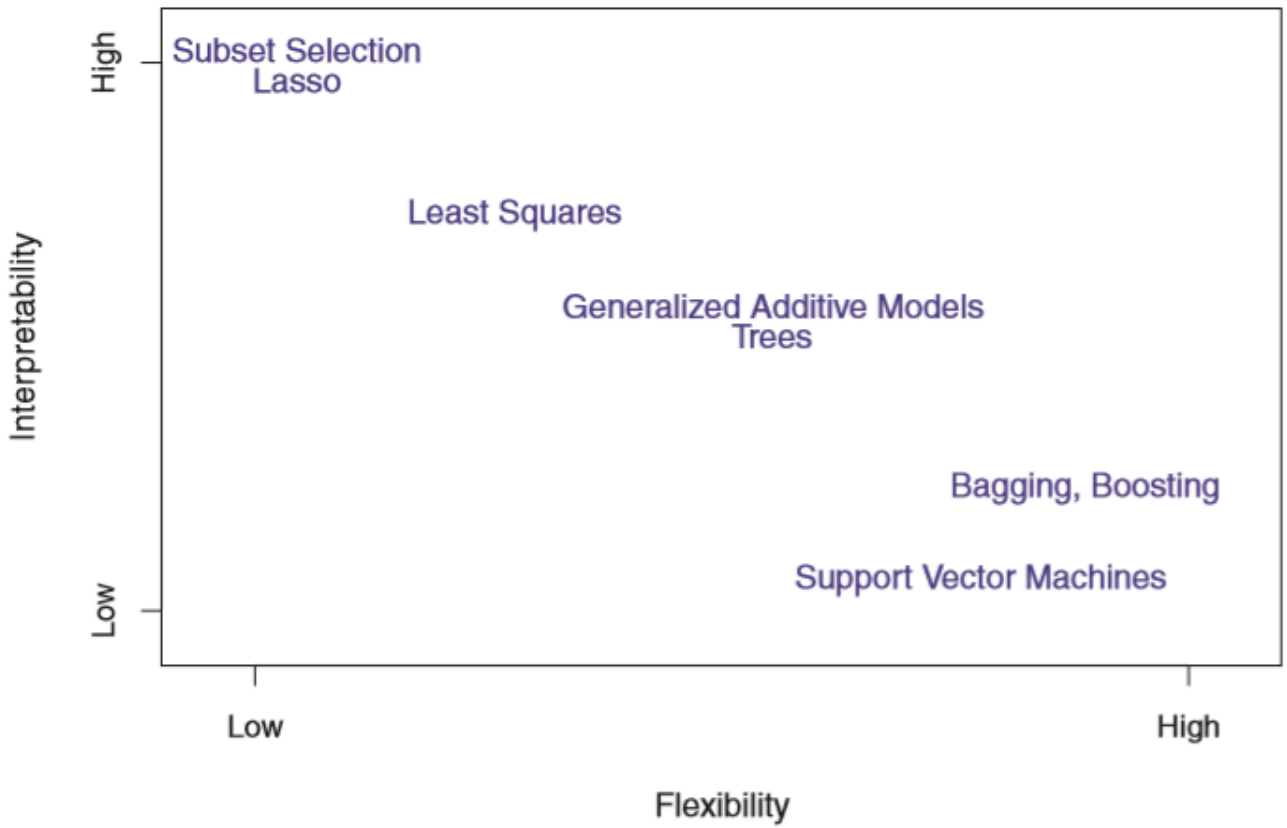
MSE ?

Parametric vs. Non-parametric





Flexibiilty vs. Interpretability



Model Accuracy

- Find some measure of error
- $MSE = \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \hat{f}(x_i) \right)^2$
- Note that this error is empirical error
- Methods trained to reduce MSE on training data S - is that enough?

Model Accuracy

- Find some measure of error
- $MSE = \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \hat{f}(x_i) \right)^2$
- Note that this error is empirical error
- Methods trained to reduce MSE on training data S - is that enough?
- why do we care about previously seen data? Instead we care about unseen test data to understand out the model predicts.
- so how do we select methods that reduce MSE on test data?

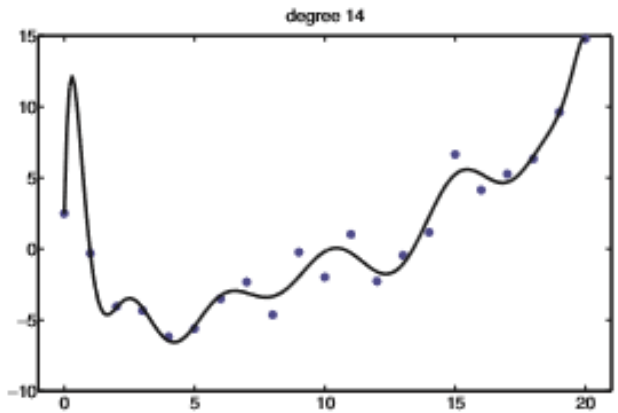
Bias Variance Tradeoff

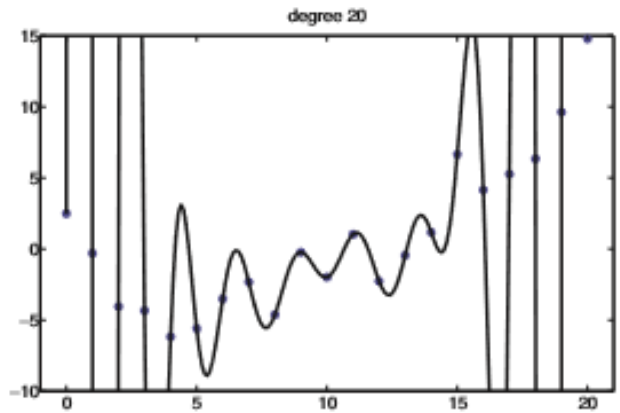
- Variance - the amount $\hat{f}$ would change with different training data
- Bias - the error introduced by approximating $\hat{f}$ with a simpler model
- $\mathbb{E}(y - \hat{f}(x))^2 = Var(\hat{f}(x)) + [Bias(\hat{f}(x))]^2 + Var(\epsilon)$
- Error Rate $= \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}(y_i \neq \hat{y}_i)$

Bias Variance Tradeoff

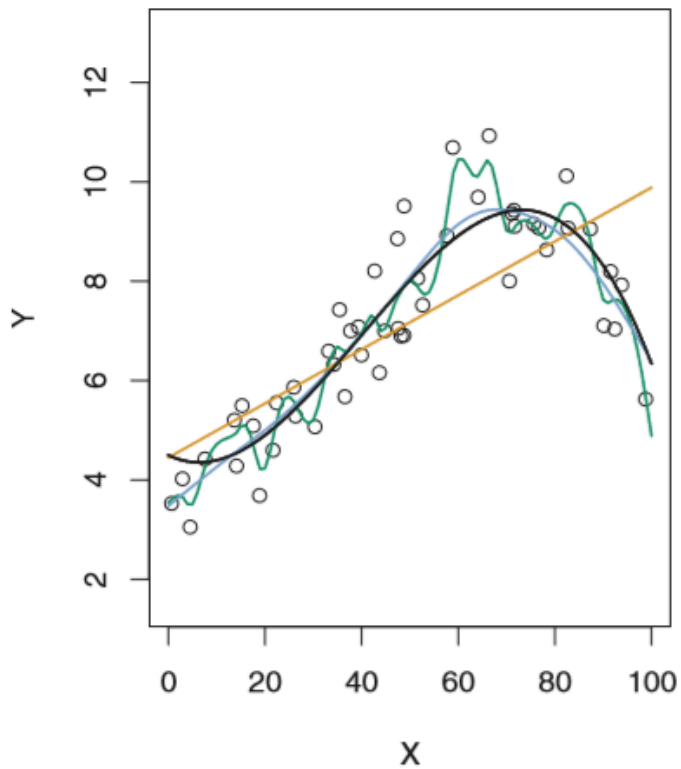
- Error Rate $= \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}(y_i \neq \hat{y}_i)$
- Test error is minimized by understanding the conditional probability $p(Y = j|X = x_0)$
- Bayes Error Rate $1 - \mathbb{E}(max_j p(Y = j|X))$
- K-Nearest Neighbor to solve this: $p(Y = j|X = x_0) = \frac{1}{K}\sum_{i \in N_0} \mathbb{I}(y_i = j)$
- More on this later!

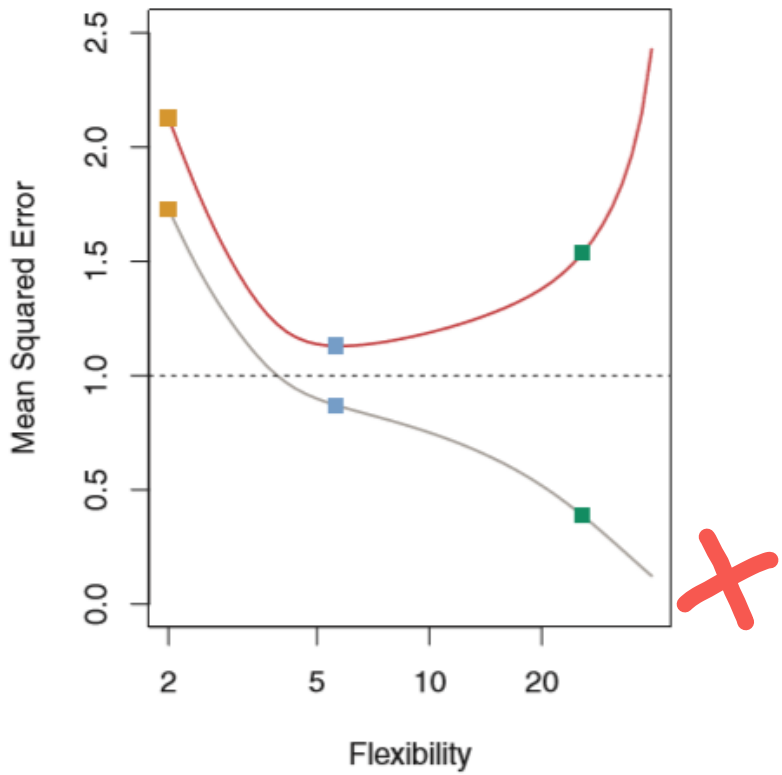
Key Challenge: Avoid Overfitting





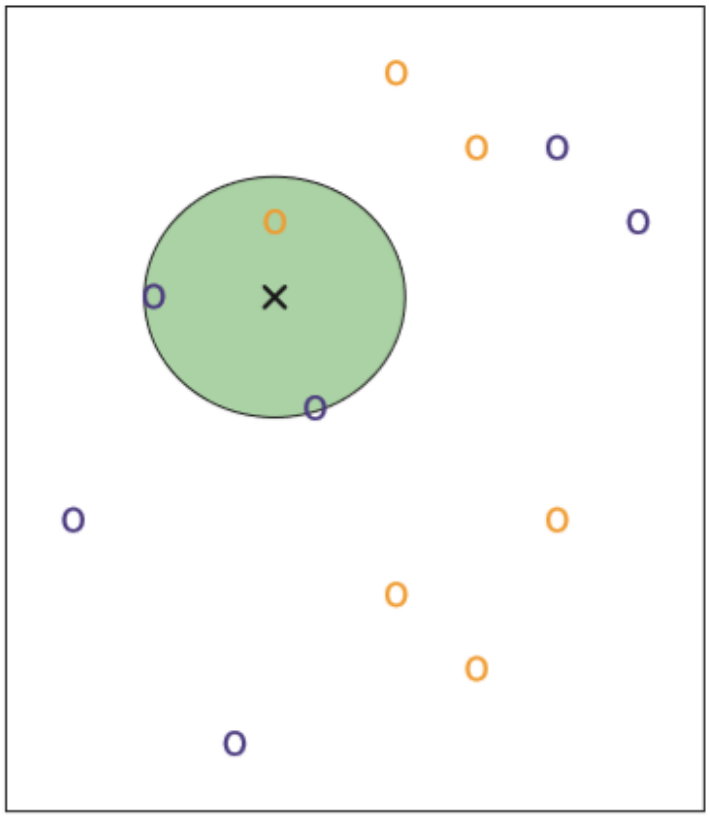
U-Shape Tradeoff

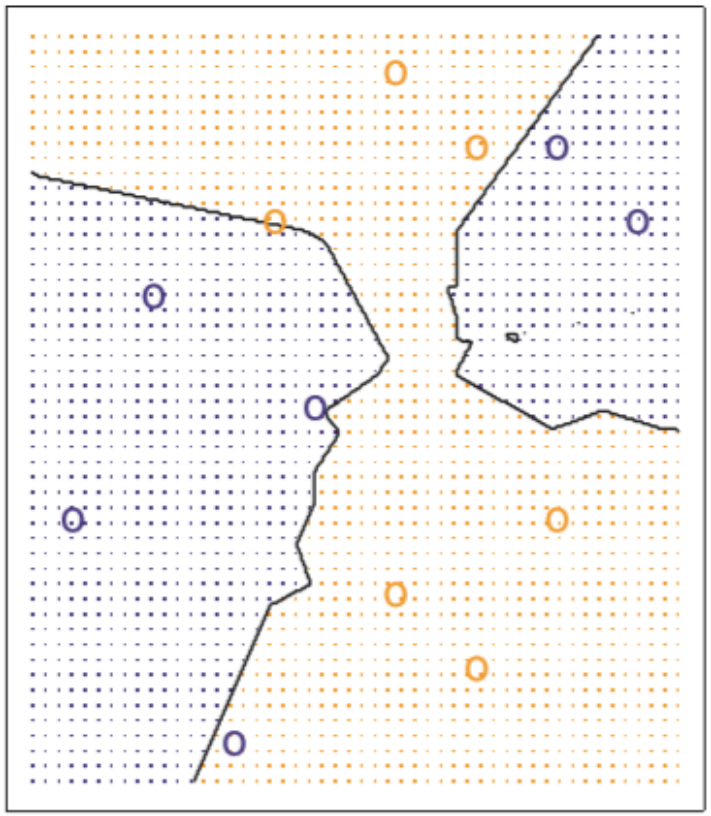




K Nearest Neighbor

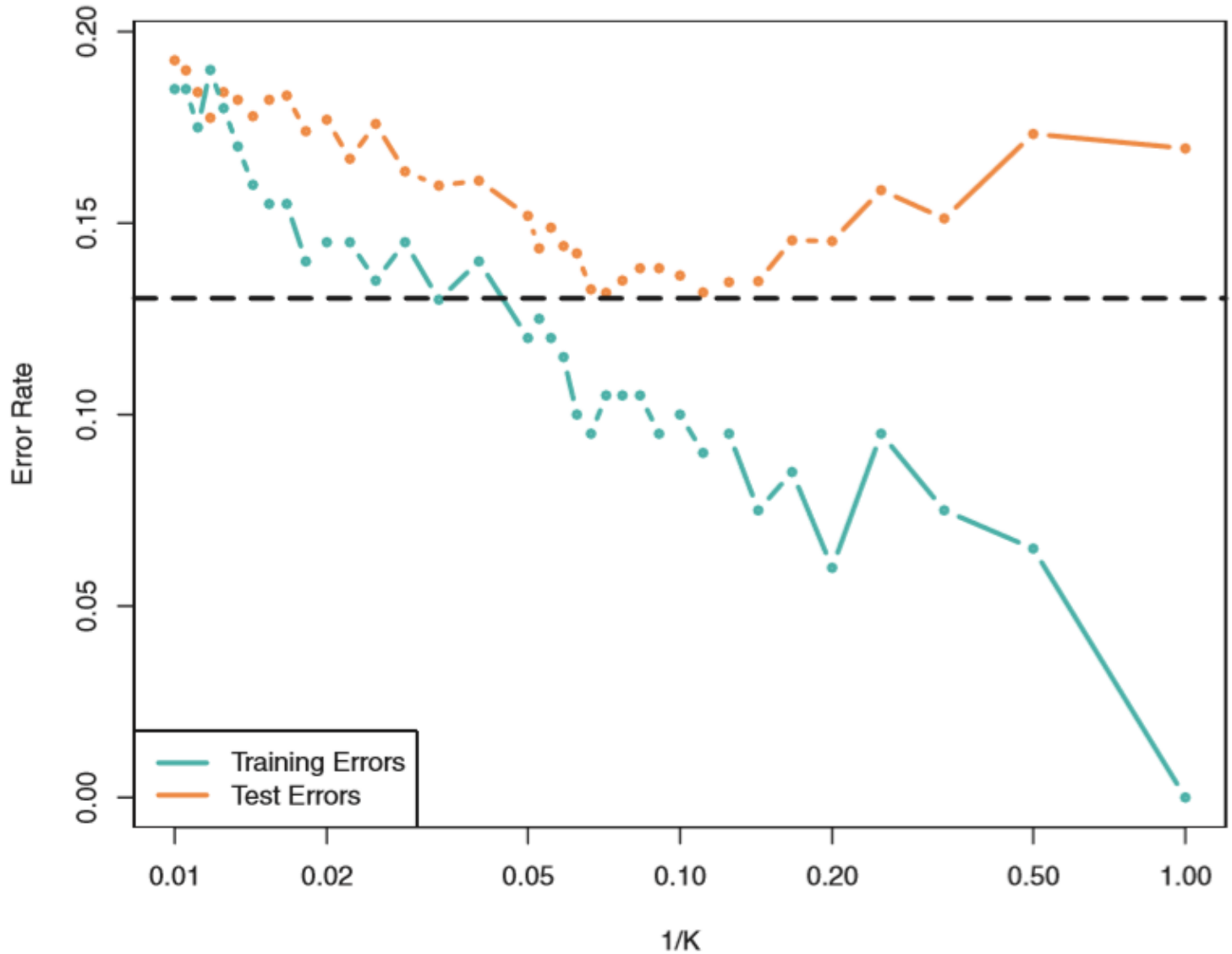
- $p(y = c|x, D, k) = \frac{1}{k} \sum_{i \in N_k(x,D)} \mathbb{I}(y_i = c)$





-

Curse of Dimensionality



No Free Lunch Theorem

- "All models are wrong but some models are useful", G. Box, 1987
- No single ML system works for everything
- Principal Component Analysis a way to solve curse of dimensionality
- Machine learning is not magic: it can't get something out of nothing, but it can get more from less!

Takeaways and Next Time

- Course Structure
- What are machine learning tasks?
- What are some basics of learning?
- Can a learner always achieve zero prediction error?
- Next session: Model Selection, Testing methodologies
- Sources: Foundations of machine learning - Mohri et al, and An Introduction to Statistical Learning - James et al