DS 4400

Machine Learning and Data Mining I
Spring 2022

Alina Oprea
Associate Professor
Khoury College of Computer Science
Northeastern University

January 24 2022

Today's Outline

- Learning tasks
  - Supervised Learning: classification, regression
  - Unsupervised Learning
- ML terminology
- Learning challenges
  - Bias-Variance tradeoff
- Probability review

Course Information

- Website: http://www.ccs.neu.edu/home/alina/classes/Spring2022
- Canvas: https://canvas.northeastern.edu

- Gradescope: gradescope.com

- Communication: piazza.com

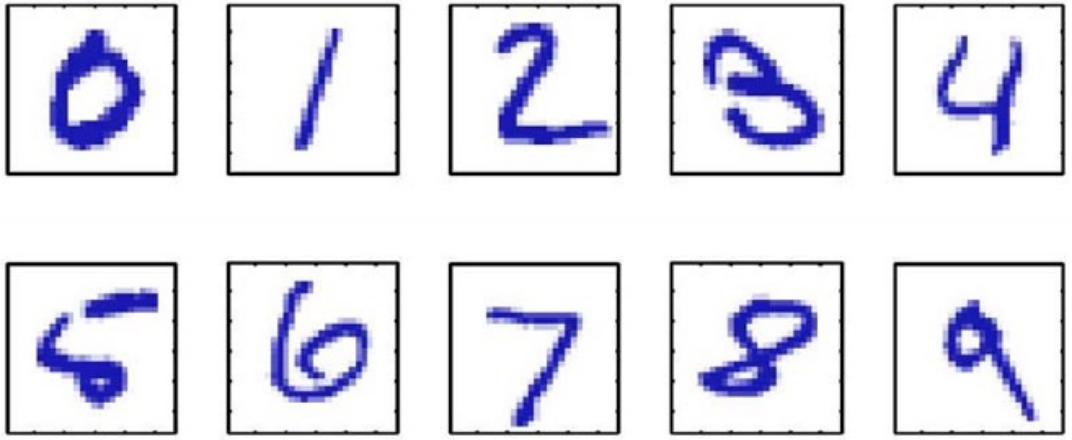
Learning Tasks

- Supervised learning
  - Classification
  - Regression
  - Examples
- Unsupervised learning
  - Clustering

Slides adapted from

- A. Zisserman, University of Oxford, UK
- S. Ullman, T. Poggio, D. Harari, D. Zysman, D Seibert, MIT
- D. Sontag, MIT
- Figures from “An Introduction to Statistical Learning”, James et al.

Example 1
Handwritten digit recognition



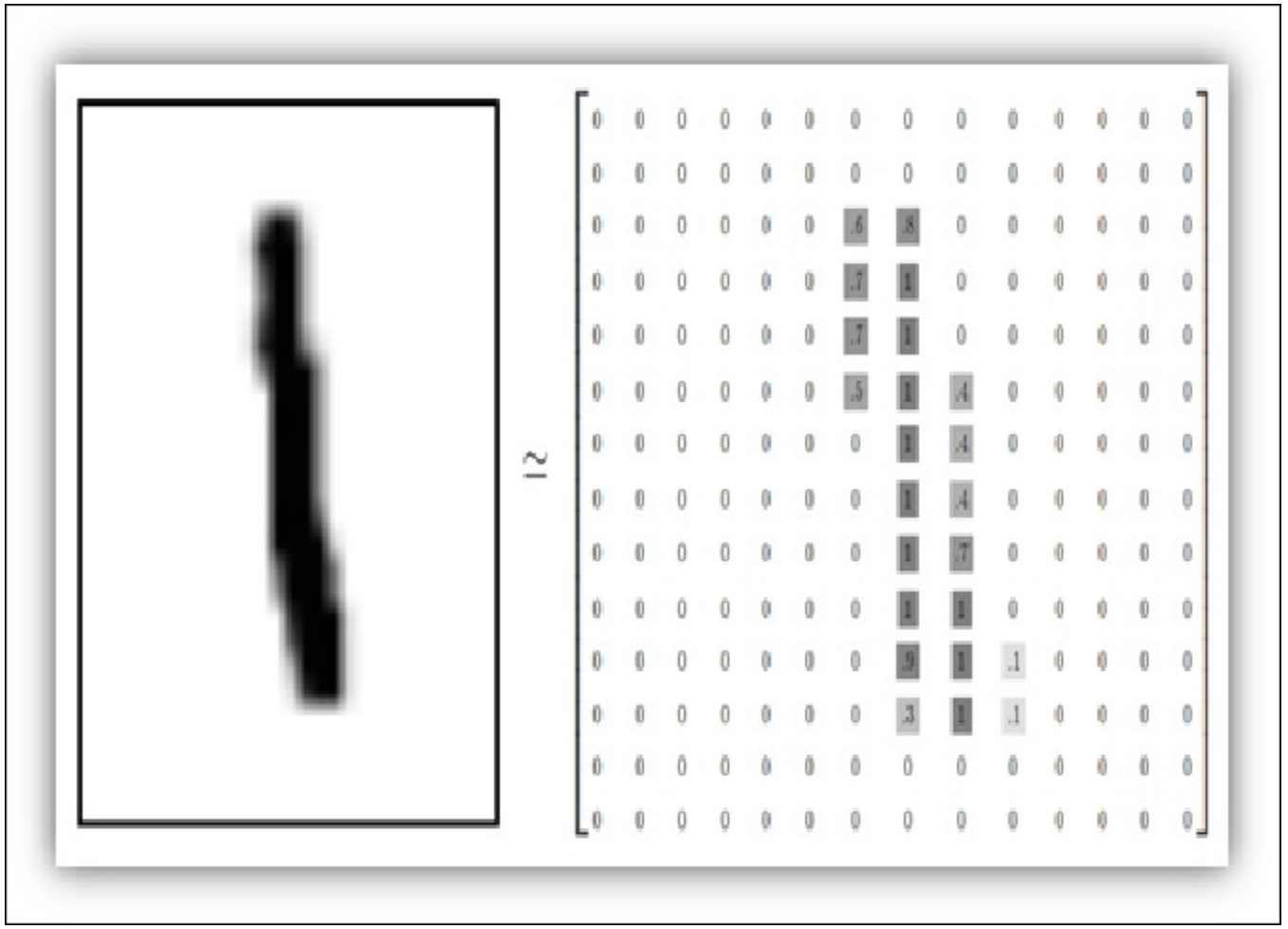
Images are 28 x 28 pixels

Represent input image as a vector $\mathbf{x} \in \mathbb{R}^{784}$
Learn a classifier $f(\mathbf{x})$ such that,

$$f : \mathbf{x} \rightarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

MNIST dataset: Predict the digit
Multi-class classifier

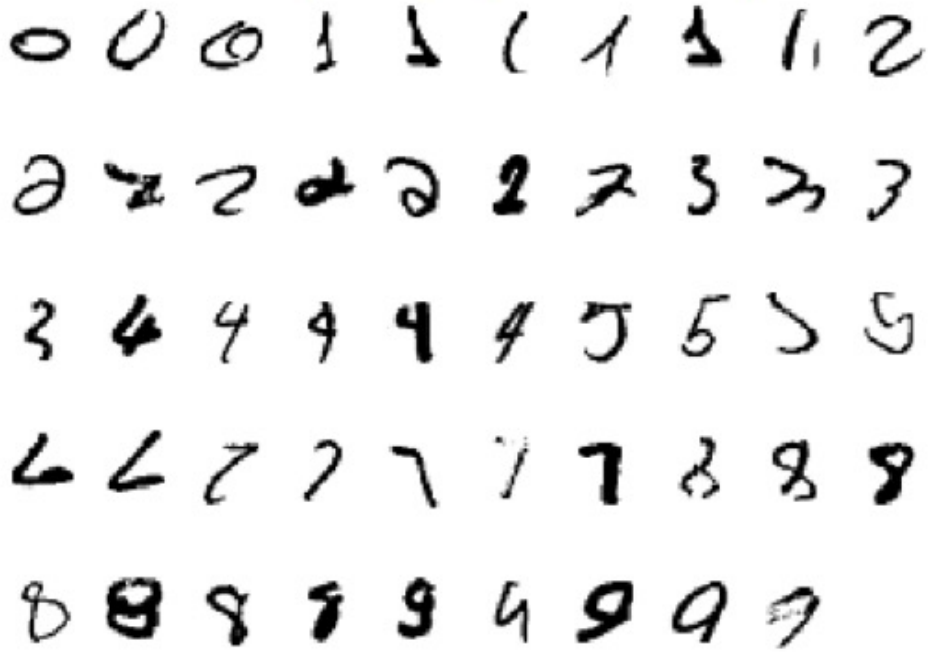
Data Representation



Model the problem

As a supervised classification problem

Start with training data, e.g. 6000 examples of each digit



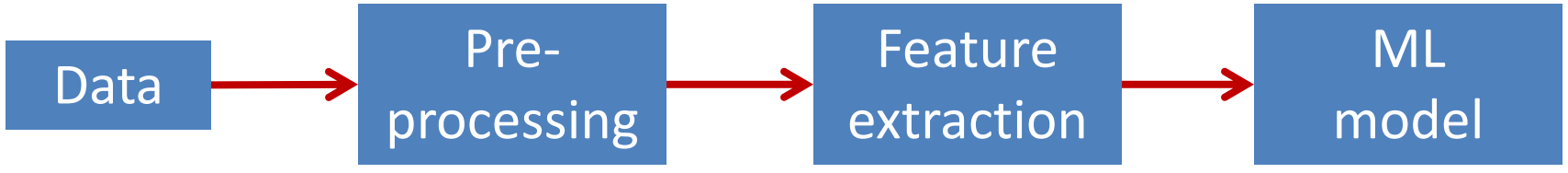
- Can achieve testing error of 0.4%
- One of first commercial and widely used ML systems (for zip codes & checks)

Other examples

- Spam classification
  - Is my email spam or not? Binary classification
  - Is the attachment safe?
- Weather prediction
  - Will it rain tomorrow or not?
- Healthcare classification
  - Is the patient sick or not?
- Image classification
  - What object does the image depict?
  - Where is the object in the image?

Supervised Learning: Classification

Training



Labeled

$x_i, y_i \in \{0,1\}$

Normalization

Feature Selection

Classification

$f(x)$

Testing

$y' = f(x') \in \{0,1\}$

Unlabeled

x'

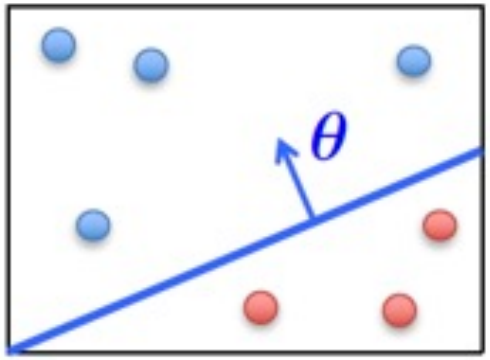
f

Positive
Negative

Classification

Classification

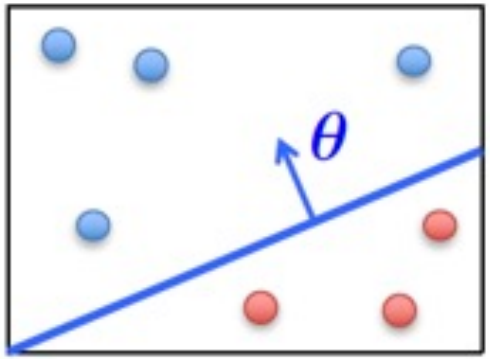
- Training data
  - $x_i = [x_{i,1}, \ldots x_{i,d}]$: vector of image pixels (features)
  - Size d = 28x28 = 784
  - y_i: image label
- Models (hypothesis)
  - Example: Linear model (parametric mod
    - $f(x) = wx + b$
  - Classify 1 if $f(x) >$ T ; 0 otherwise
- Classification algorithm
  - Training: Learn model parameters w, b to minimize objective
  - Output: "optimal" model
- Testing
  - Apply learned model to new data and generate prediction $f(x)$



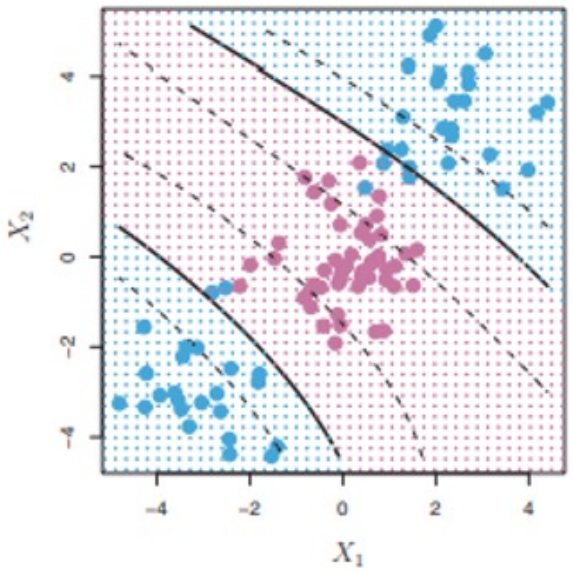
Objectives

- What are we trying to optimize?
  - Minimize error
  - Maximize accuracy

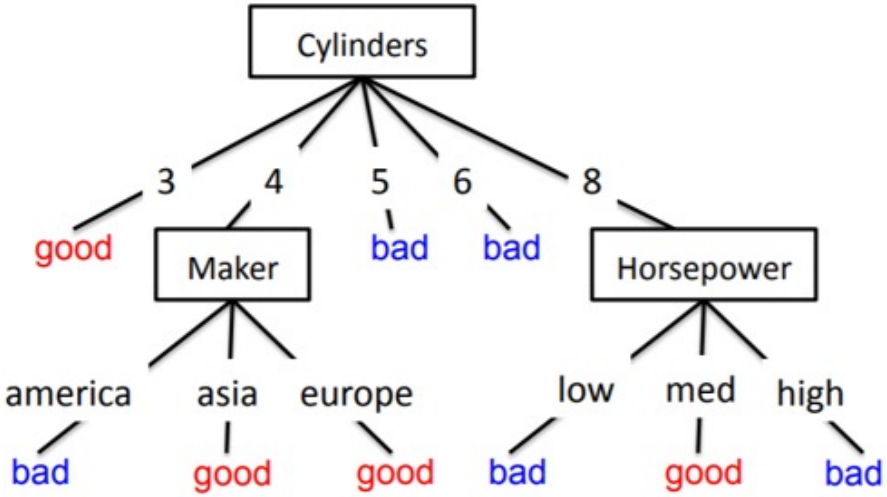
Example Classifiers



Linear classifiers: logistic regression, SVM, LDA



SVM polynomial kernel



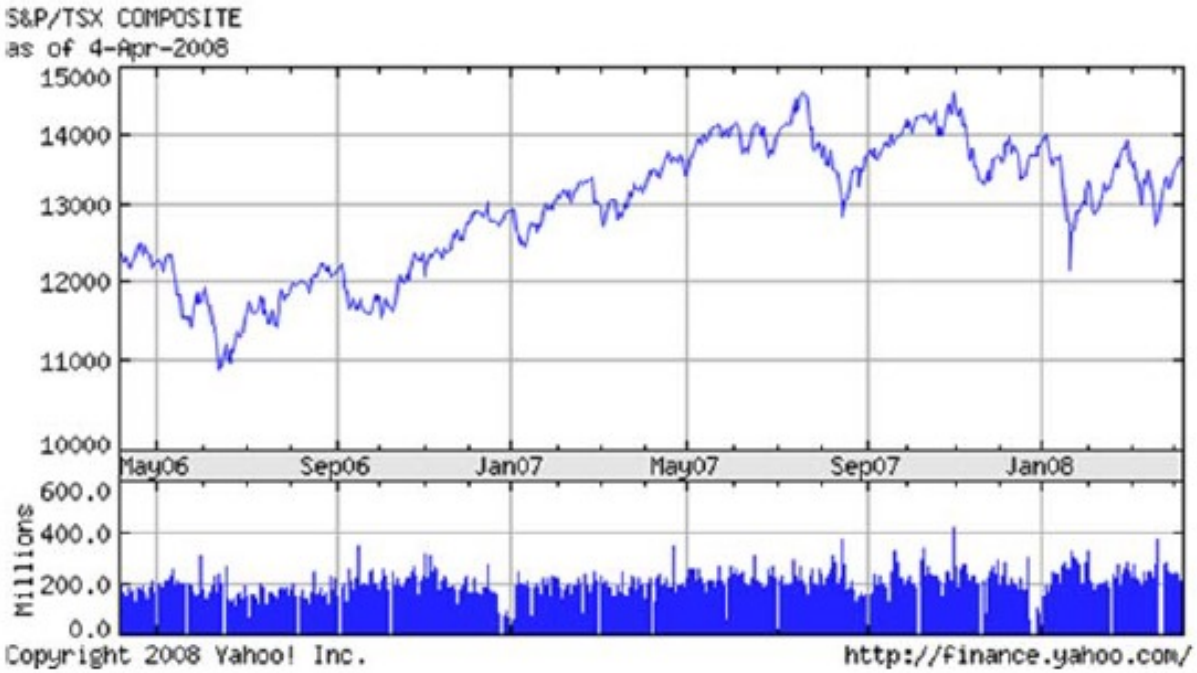
Decision trees

Why Multiple Models?

- There is no free lunch in statistics / ML!

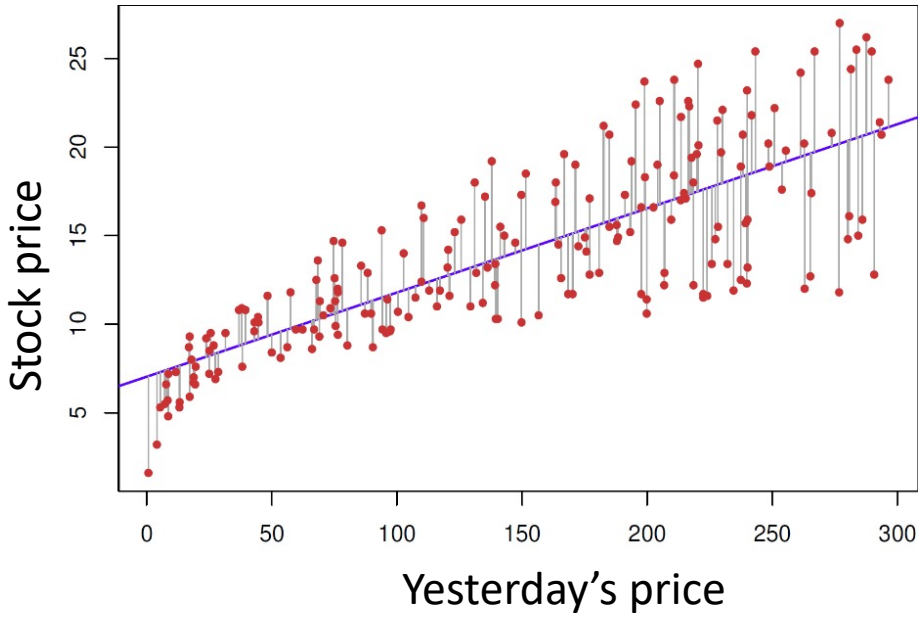
- There is no single model that dominates all
- Performance depends on many things, such as:
  - Data distribution
  - Data dimensionality
  - Quality of data and labeling

Example 2
Stock market prediction



- Task is to predict stock price at future date
- This is a regression task, as the output is continuous

Regression



Linear regression
1 dimension

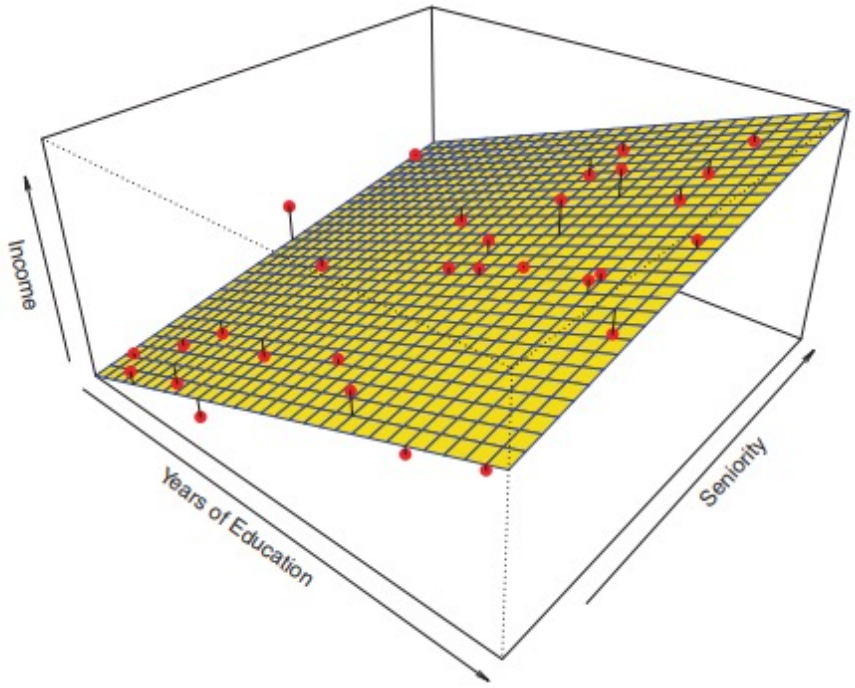
- Suppose we are given a training set of N observations

$$(x_1, \ldots, x_N) \text{ and } (y_1, \ldots, y_N)$$

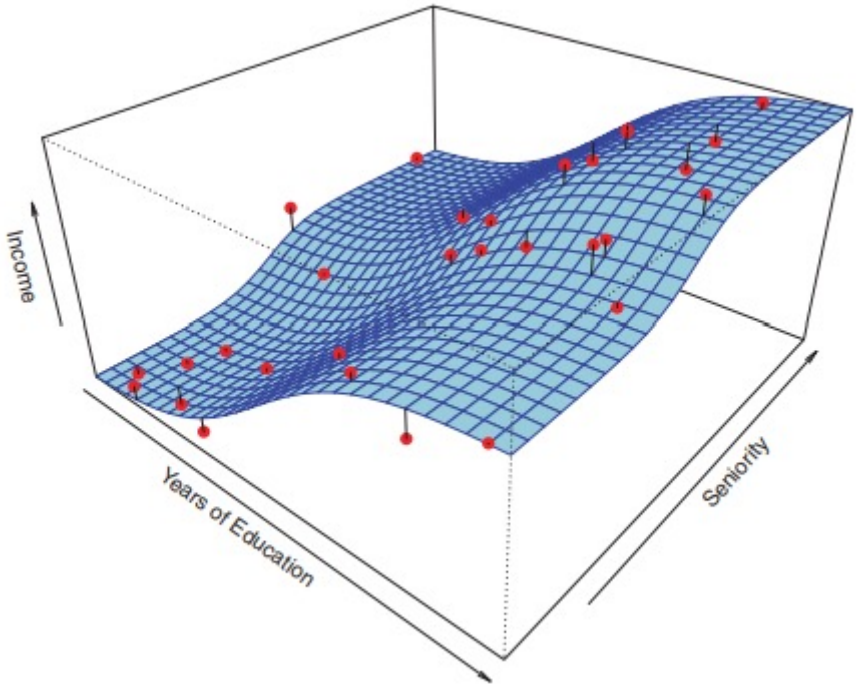
- Regression problem is to estimate y(x) from this data

$x_i = (x_{i1}, \ldots, x_{id})$ - d predictors (features)
y_i - response variable, numerical

Income Prediction



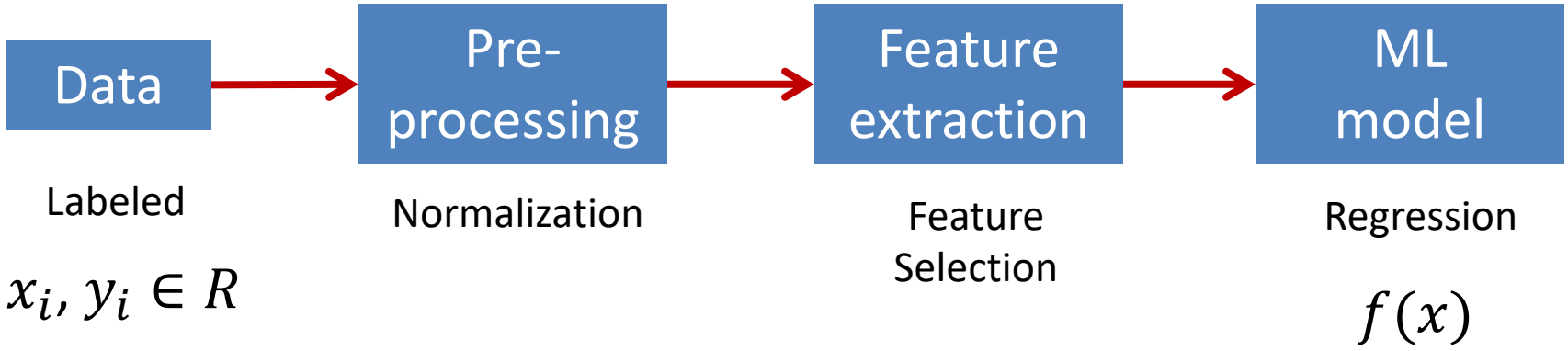
Linear Regression



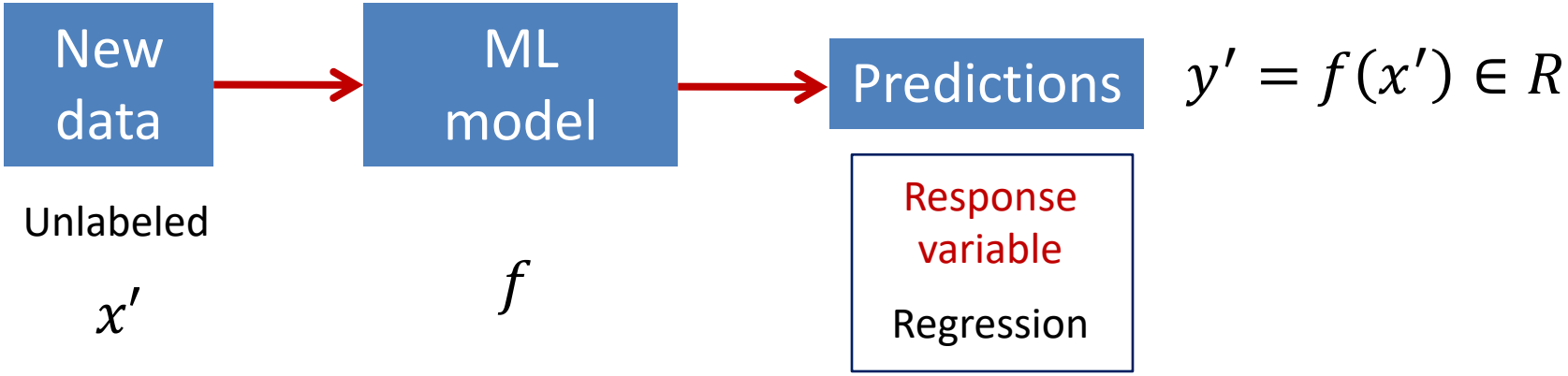
Non-Linear Regression
Polynomial/Spline Regression

Supervised Learning: Regression

Training

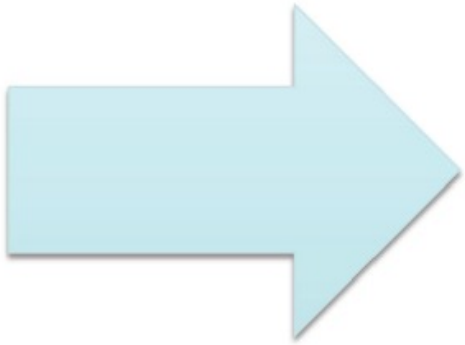


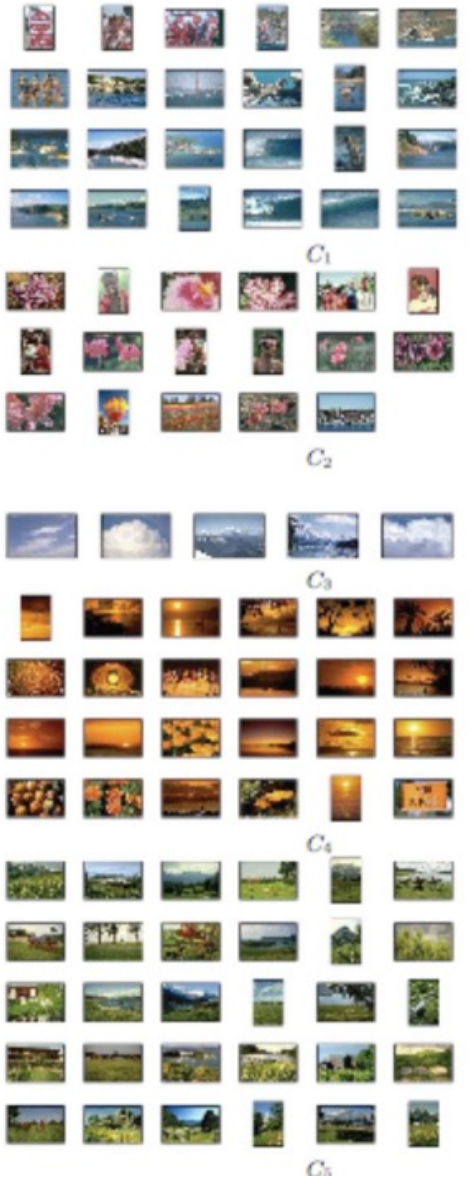
Testing



Example 3: image search

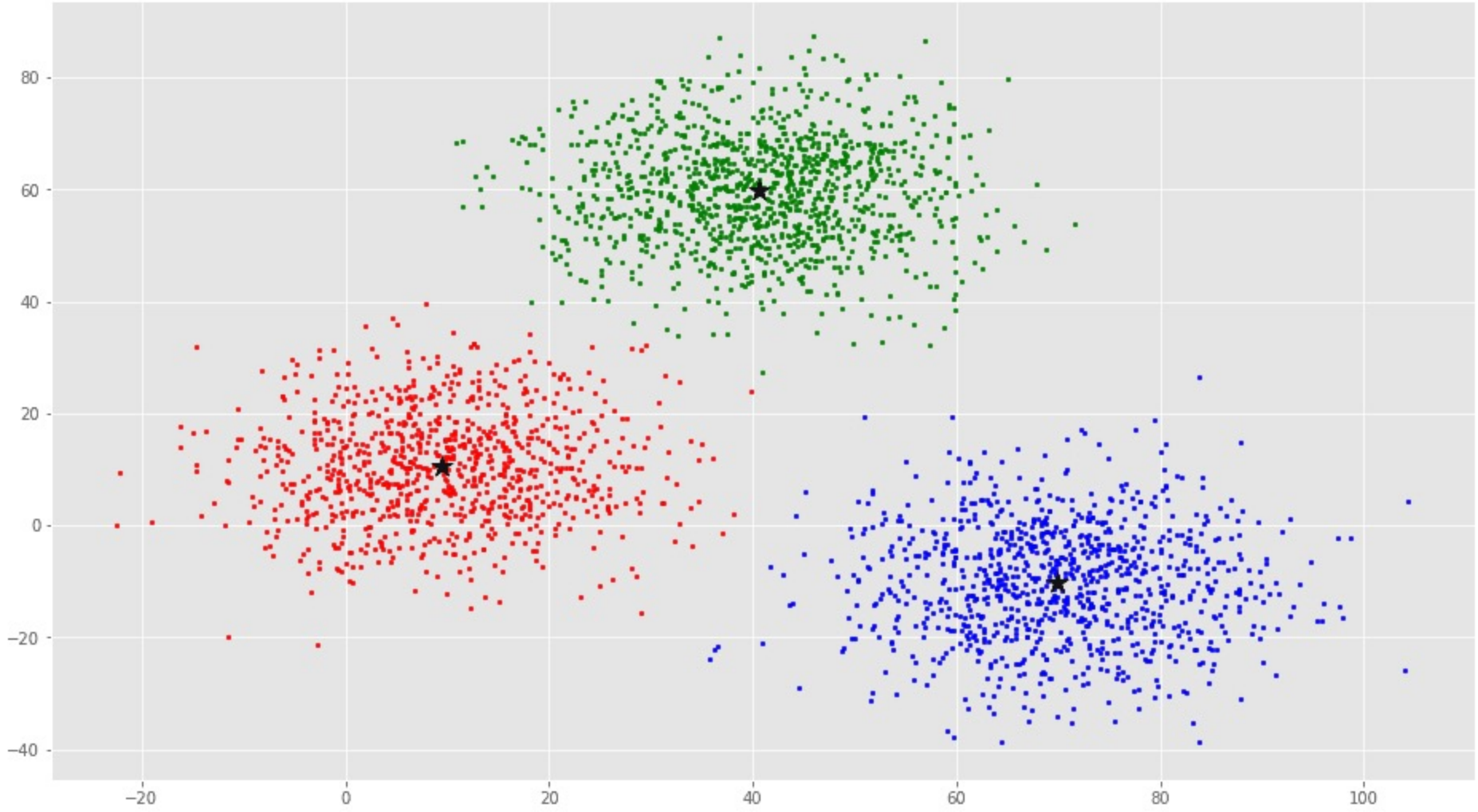
Clustering images





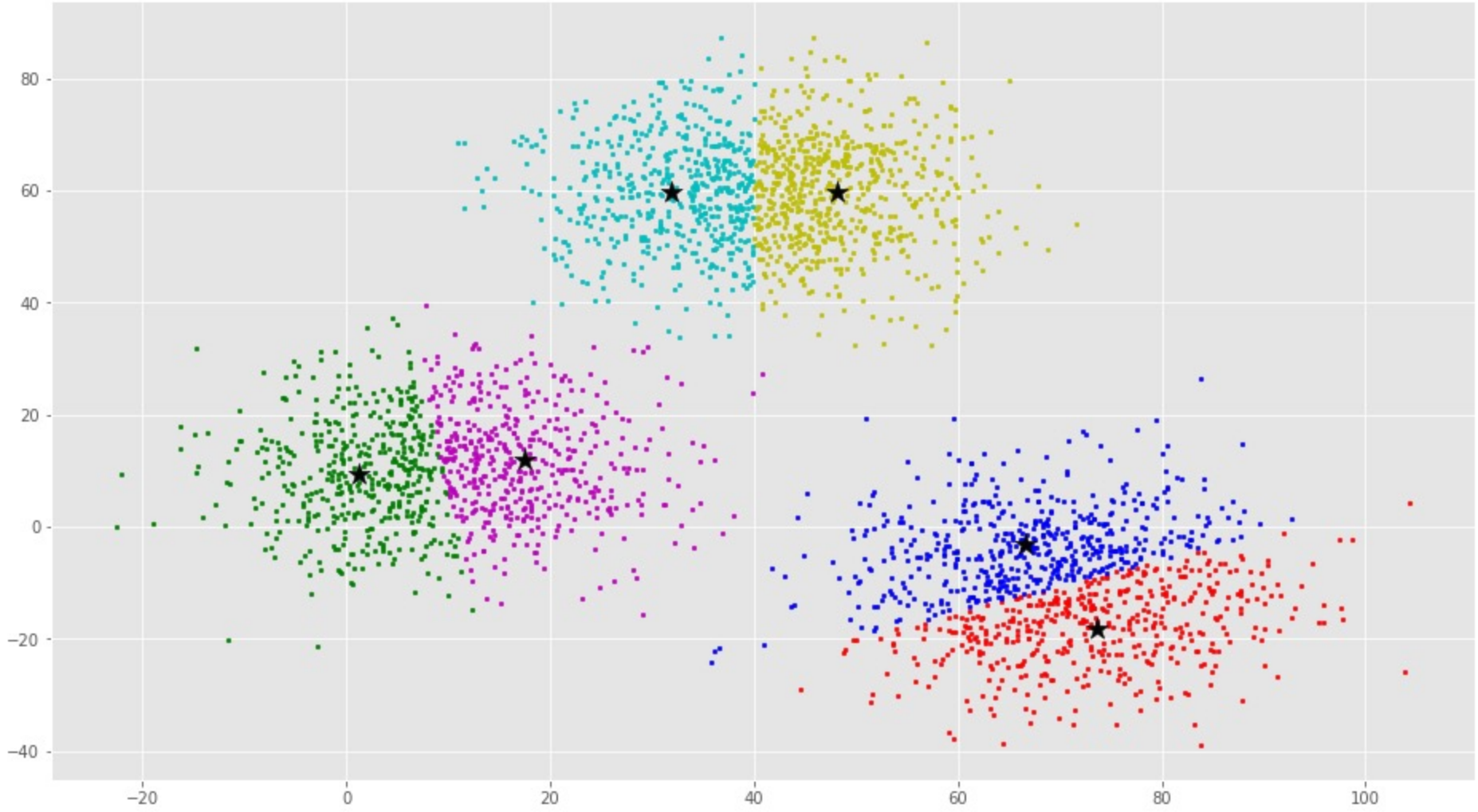
Find similar images to a target one

K-means Clustering



K=3

K-means Clustering



K=6

Unsupervised Learning

- Clustering
  - Group similar data points into clusters
  - Example: k-means, hierarchical clustering, density-based clustering
- Dimensionality reduction
  - Project the data to lower dimensional space
  - Example: PCA (Principal Component Analysis), UMAP
- Feature learning
  - Find feature representations
  - Example: Autoencoders

Supervised Learning Tasks

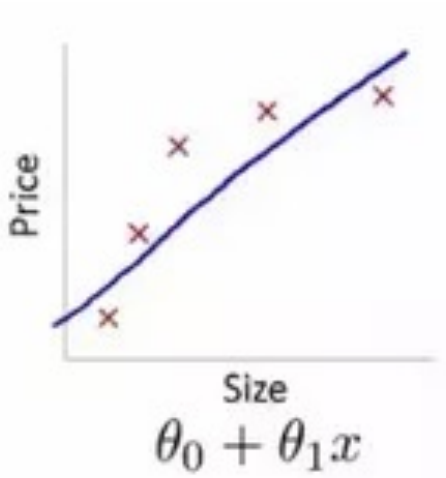
- Classification
  - Learn to predict class (discrete)
  - Minimize classification error
- Regression
  - Learn to predict response variable (numerical)
  - Minimize MSE (Mean Square Error)
- Both classification and regression
  - Training and testing phase
  - "Optimal" model is learned in training and applied in testing

Learning Challenges

- Chapters 2.2.1 and 2.2.2 from ISL book
- Goal
  - Classify well new testing data
  - Model generalizes well to new testing data
  - Minimize error (MSE or classification error) in testing
- Variance
  - Amount by which model would change if we estimated it using a different training data set
- Bias
  - Error introduced by approximating a real-life problem by a much simpler model
  - E.g., for linear models (linear regression) bias is high

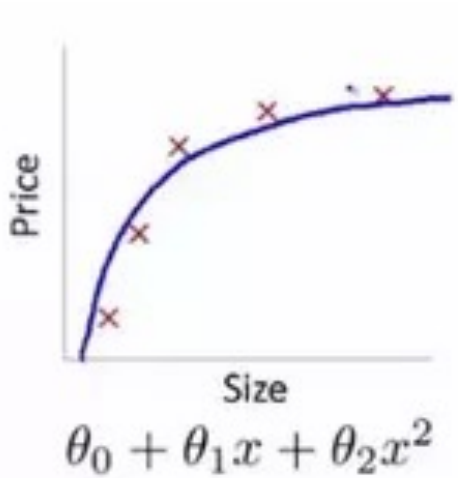
Bias-Variance tradeoff

Example: Regression



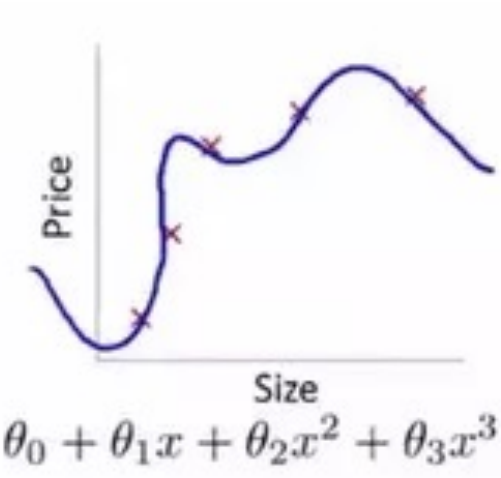
$\theta_0 + \theta_1 x$

High bias
(underfit)



$\theta_0 + \theta_1 x + \theta_2 x^2$

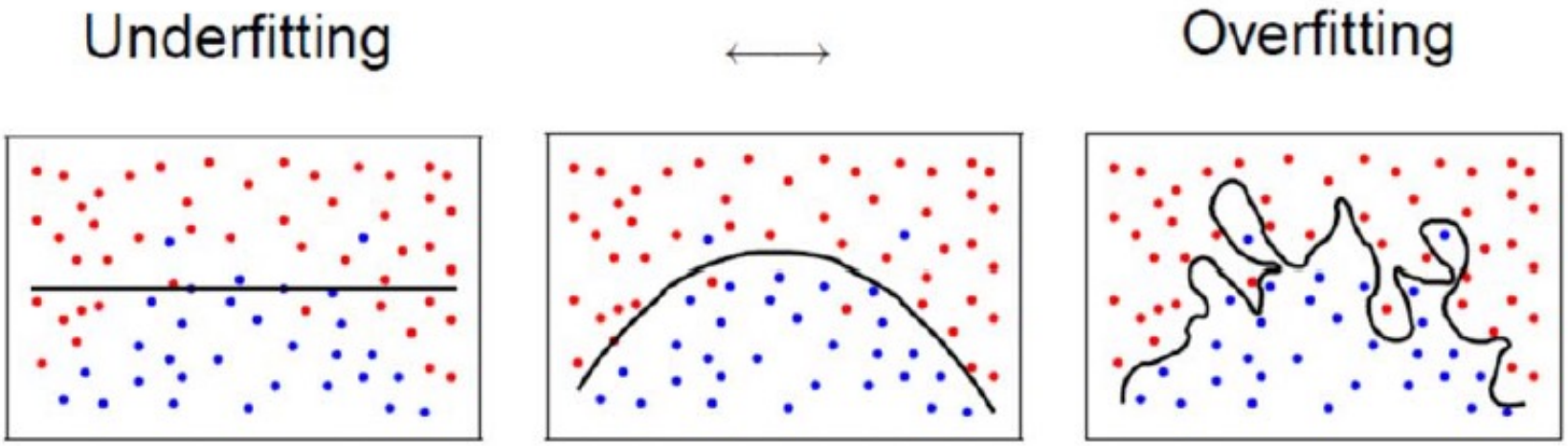
"Just right"



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

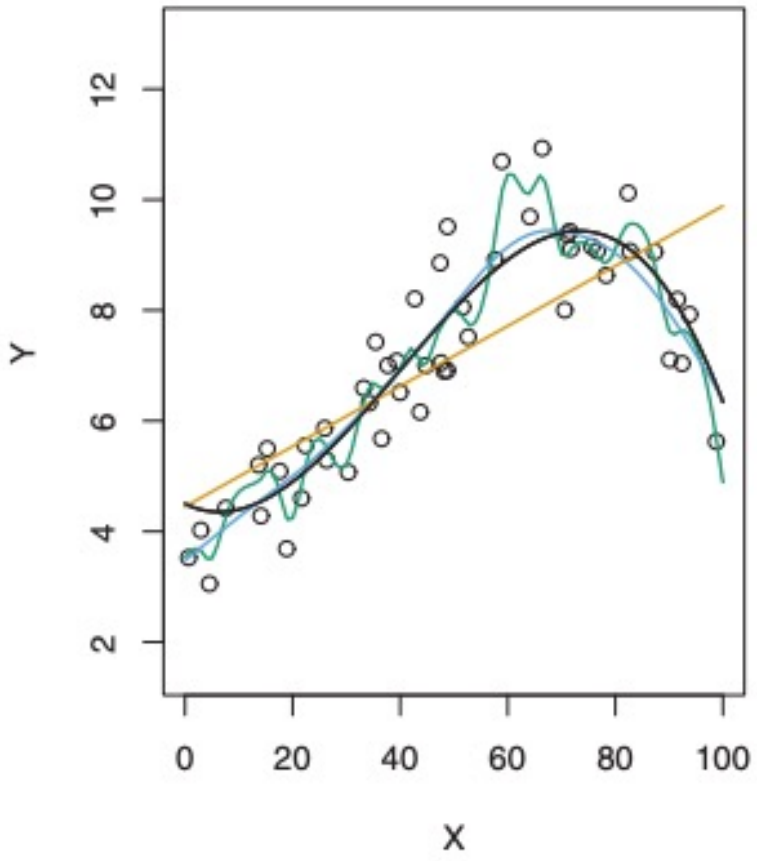
High variance
(overfit)

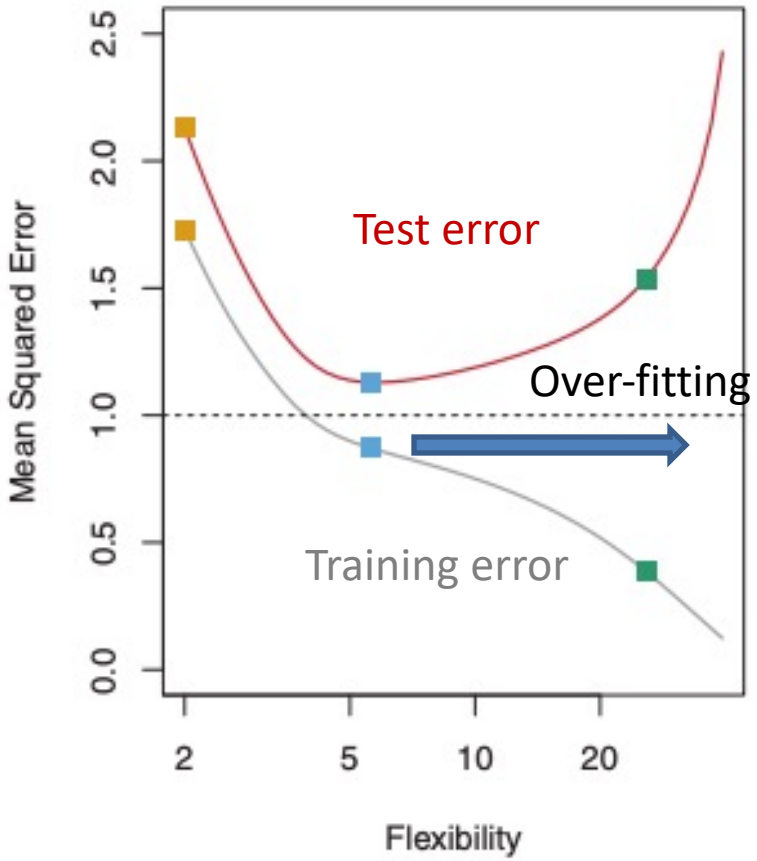
Generalization Problem in Classification



- Again, need to control the complexity of the (discriminant) function

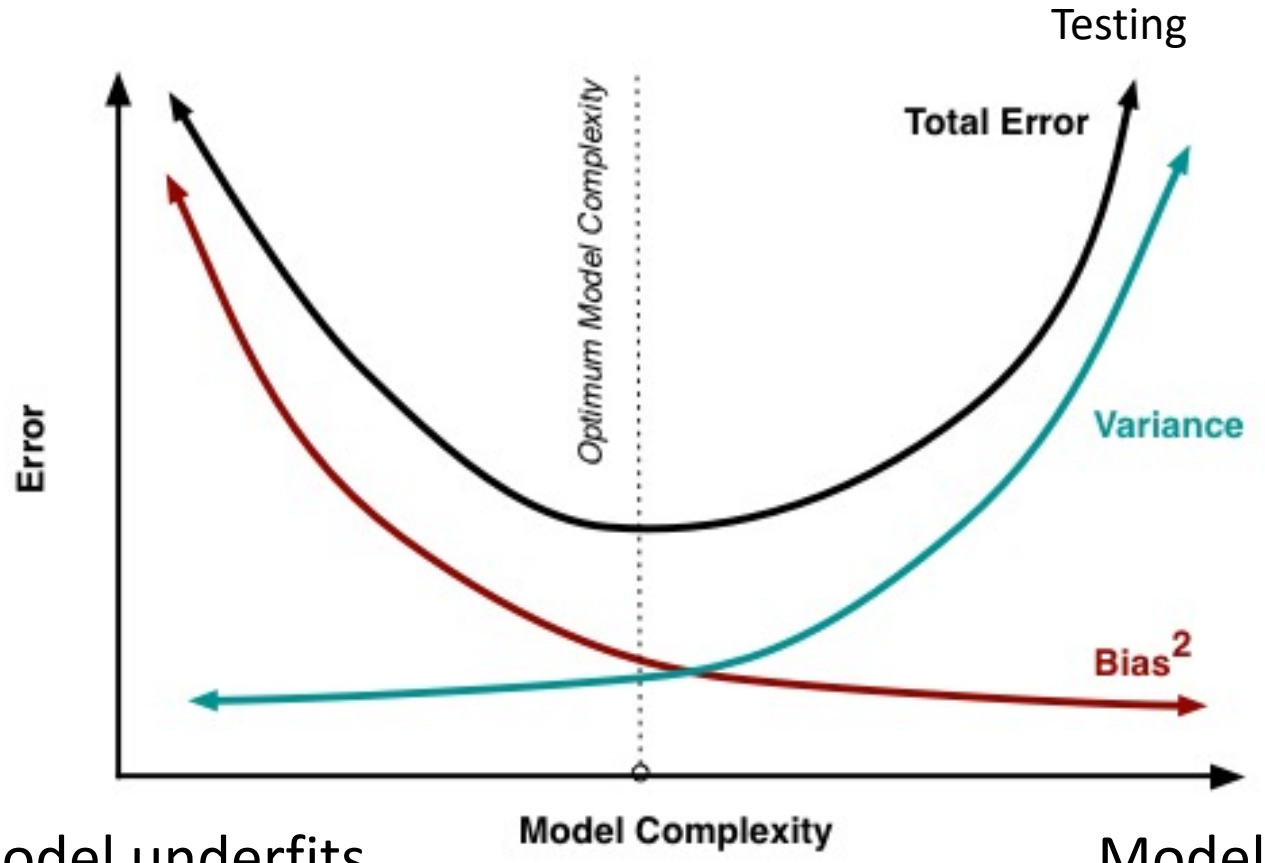
Training and testing error





ISL, Chapter 2.2.2

Bias-Variance Tradeoff



Model underfits

Model overfits

Test error is sum of bias, variance and noise

Occam's Razor

- William of Occam: Monk living in the 14th century
- Principle of parsimony:

"One should not increase, beyond what is necessary, the number of entities required to explain anything"

- When many solutions are available for a given problem, we should select the simplest one

Select the simplest machine learning model that gets reasonable accuracy for the task at hand

Recap

- ML is a subset of AI designing learning algorithms
- Learning tasks are *supervised* (e.g., classification and regression) or *unsupervised* (e.g., clustering)
  - Supervised learning uses labeled training data
- Learning the "best" model is challenging
  - Design algorithm to minimize the error in testing
  - Minimize training error is not the best strategy
  - Bias-Variance tradeoff
  - Need to generalize on new, unseen test data
  - Occam's razor (prefer simplest model with good performance)

Acknowledgements

- Slides made using resources from:
  - Andrew Ng
  - Eric Eaton
  - David Sontag
- Thanks!