DS 4400

Machine Learning and Data Mining I
Spring 2024

David M. Liu
Khoury College of Computer Science
Northeastern University

Friday January 12 2024

Today's Outline

- Learning tasks [Recap]
  - Supervised Learning: classification, regression
  - Unsupervised Learning

- ML terminology
- Learning challenges
  - Bias-Variance tradeoff

- Probability review

Office Hours Update

- Tuesday
  – Hosted by Jai from 2-3:30pm
- Wednesday
  – Hosted by Dhanush from 12-1:30pm
- Thursday
  – Hosted by David from 2-3:30pm
- Friday
  – Hosted by Caleb from 2-3:30pm

Office hours will be virtual and hosted on the Khoury Office Hours Portal*

Course Preparation

Thank you for completing the student survey!

Many folks expressed experience with calculus, probability and Python but less with linear algebra. Next Tuesday's class will provide a recap on linear algebra.

Administrative Questions?

News

The New York Times

State Legislators, Wary of Deceptive Election Ads, Tighten A.I. Rules

Sophisticated political deepfakes have warped elections overseas. Can U.S. legislators act fast enough to make A.I. campaign ads more transparent?

Share full article

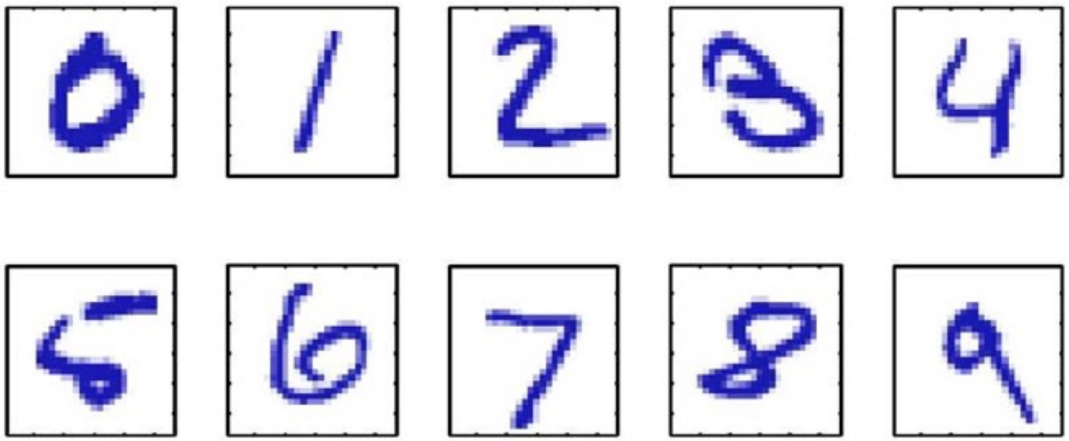
Recap from last class

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.

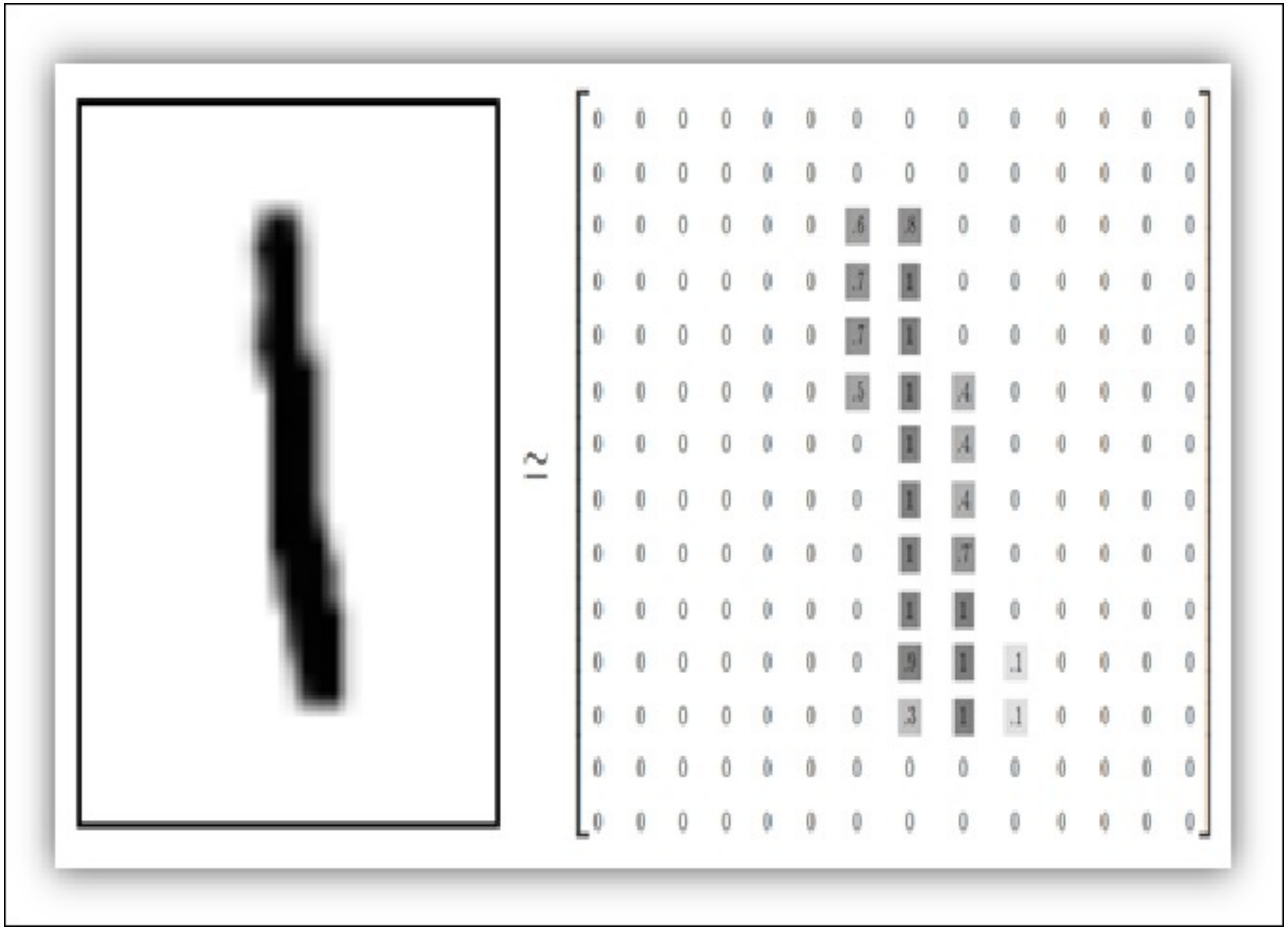
Example 1
Handwritten digit recognition



Images are 28 x 28 pixels

MNIST dataset: Predict the digit
Multi-class classifier

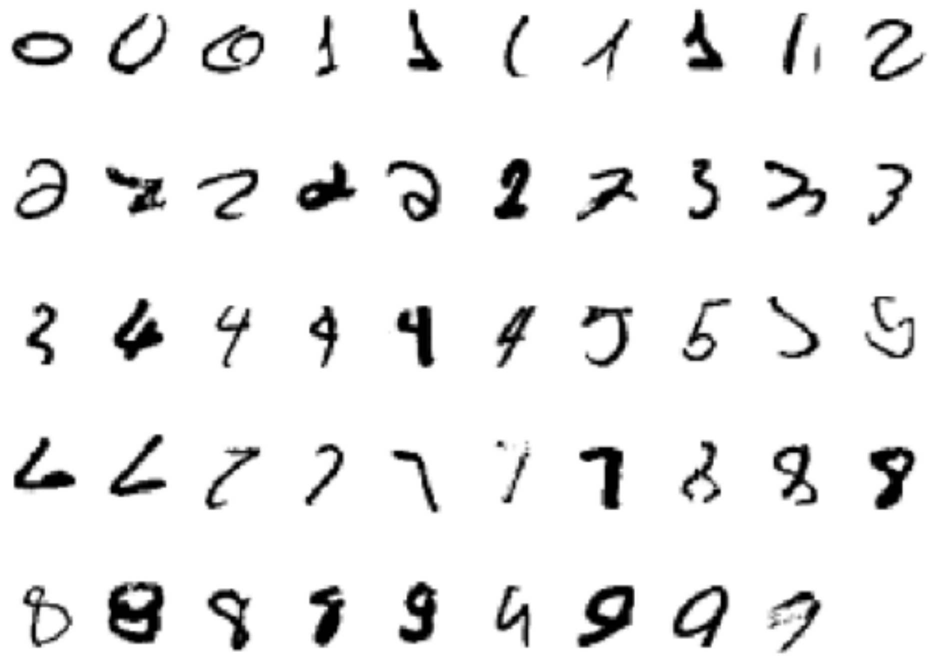
Data Representation



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

Data → Pre-processing → Feature extraction → ML model

Labeled

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

Normalization

Feature Selection

Classification

$f(x)$

Testing

New data → ML model → Predictions $y' = f(x') \in \{0,1\}$

Unlabeled

x'

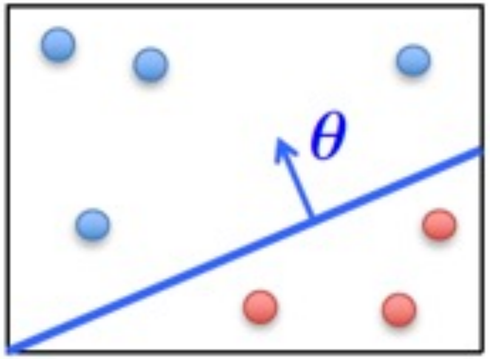
f

Positive
Negative

Classification

Classification

- Training data
  - $x_i = [x_{i,1}, \dots x_{i,d}]$: vector of image pixels (features)
  - Size $d = 28\text{x}28 = 784$
  - y_i: image label
- Models (hypothesis)
  - Example: Linear model (parametric mod
    - $f(x) = wx + b$
  - Classify 1 if $f(x) > \mathrm{T}$; 0 otherwise



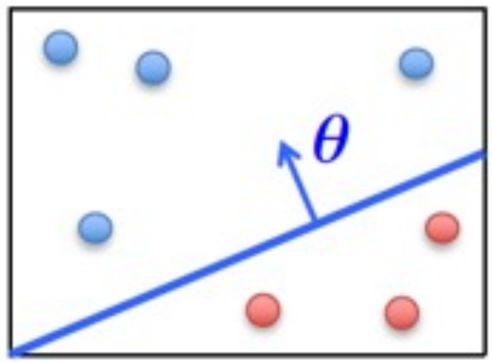
- Classification algorithm
  - Training: Learn model parameters w, b to minimize objective
  - Output: “optimal” model
- Testing
  - Apply learned model to new data and generate prediction $f(x)$

Objectives

- What are we trying to optimize?

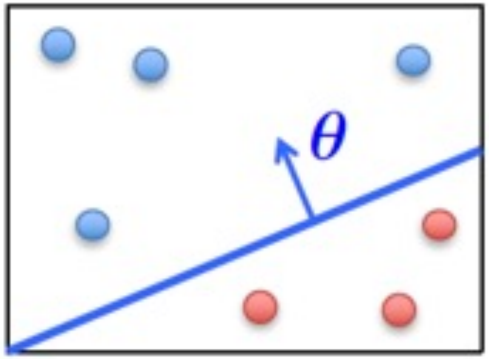
Example Classifiers

Example Classifiers

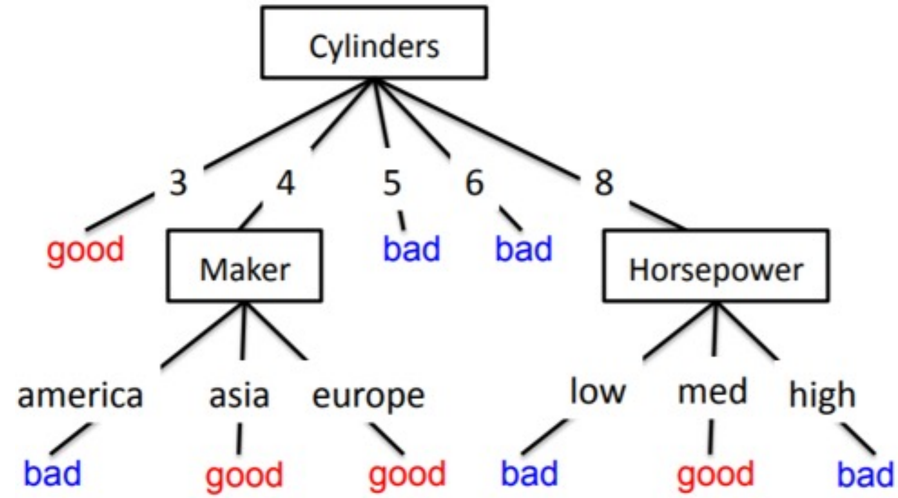


Linear classifiers: logistic regression, SVM, LDA

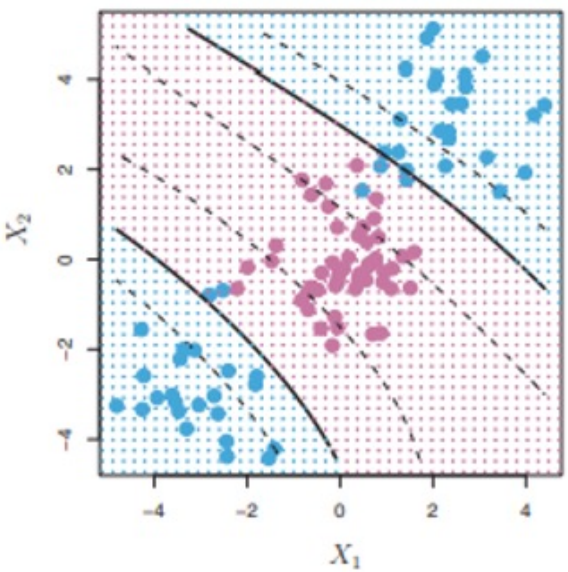
Example Classifiers



Linear classifiers: logistic regression, SVM, LDA



Decision trees



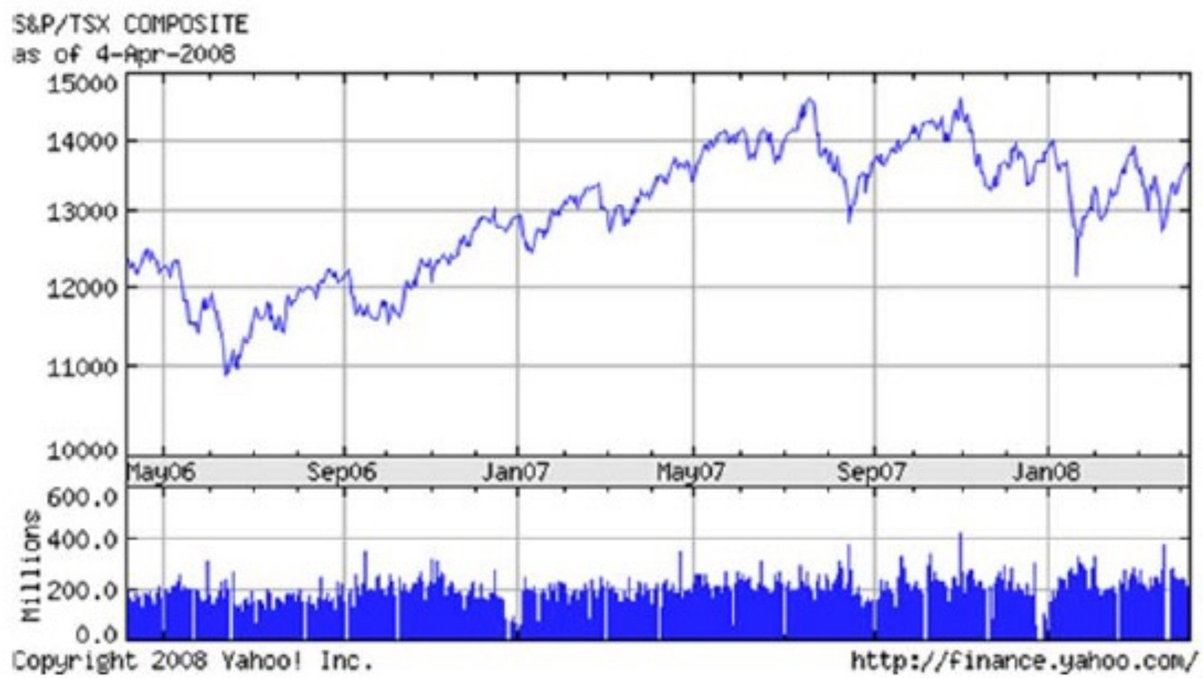
SVM polynomial kernel

Why Multiple Models?

- There is no free lunch in statistics / ML!

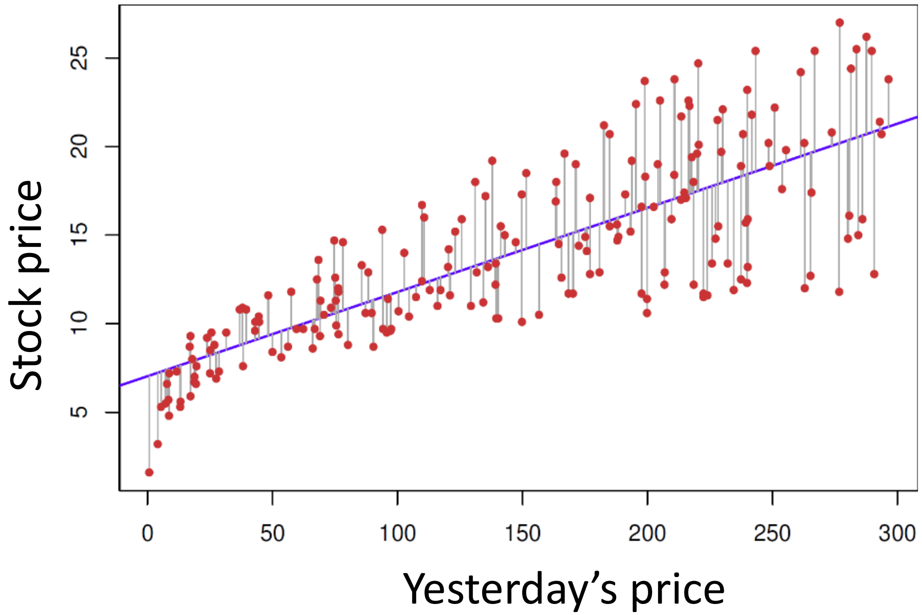
- There is no single model that dominates all
- Performance depends on many things, such as:
  - Data distribution
  - Data dimensionality
  - Quality of data and labeling

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

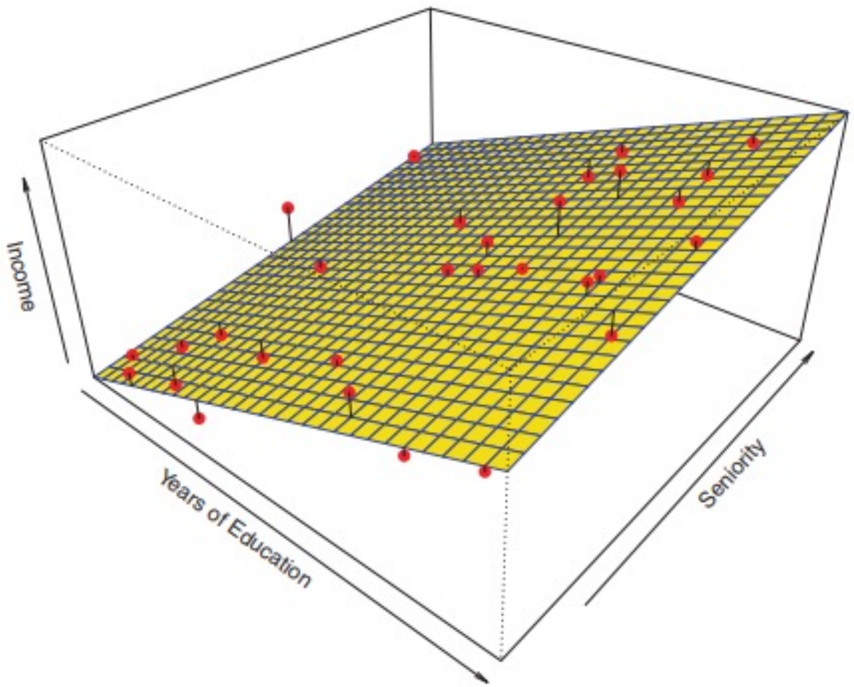
- Suppose we are given a training set of N observations

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

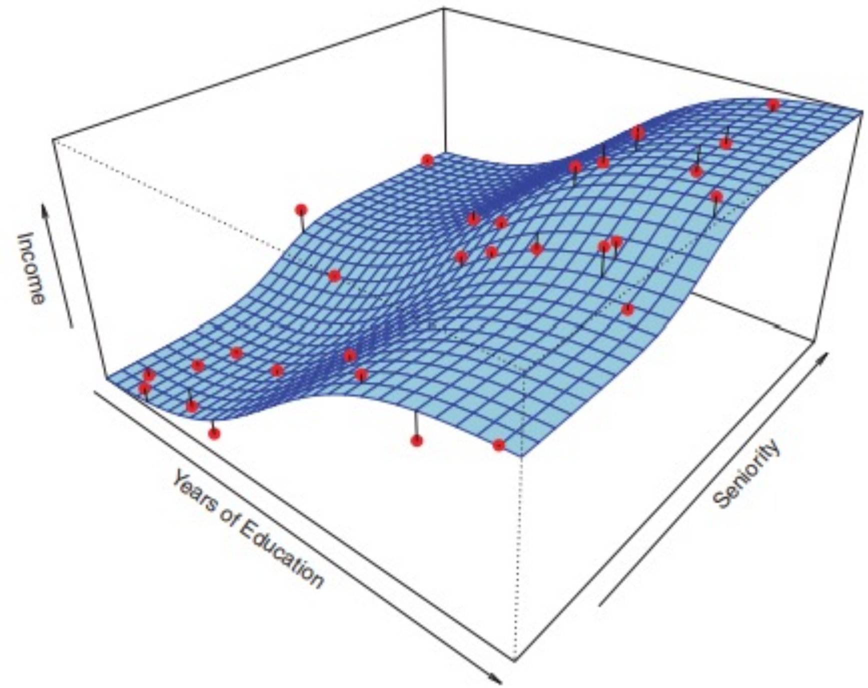
- Regression problem is to estimate y(x) from this data

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

Income Prediction



Linear Regression



Non-Linear Regression
Polynomial/Spline Regression

Supervised Learning: Regression

Training

Data → Pre-processing → Feature extraction → ML model

Labeled

$x_i, y_i \in R$

Normalization

Feature Selection

Regression

$f(x)$

Testing

New data → ML model → Predictions

Unlabeled

x'

f

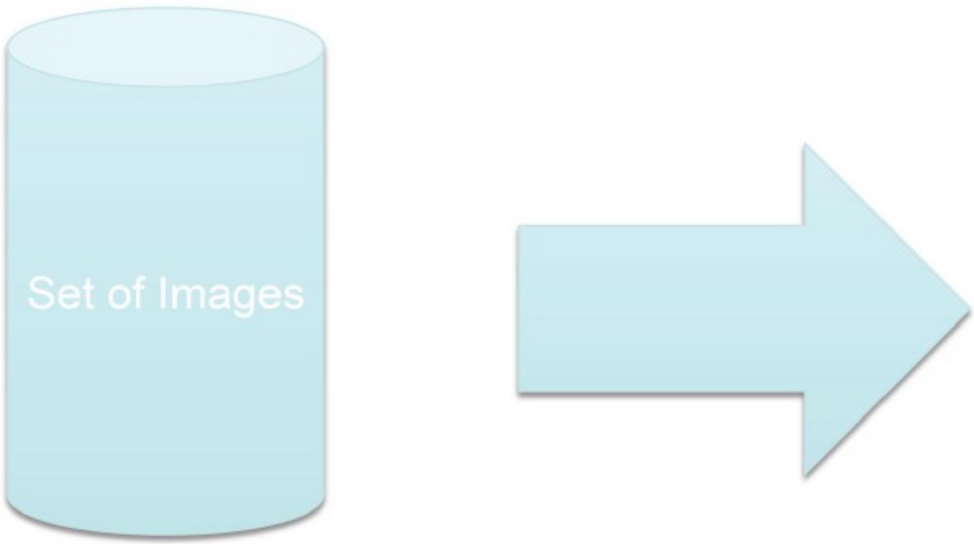
$y' = f(x') \in R$

Response variable

Regression

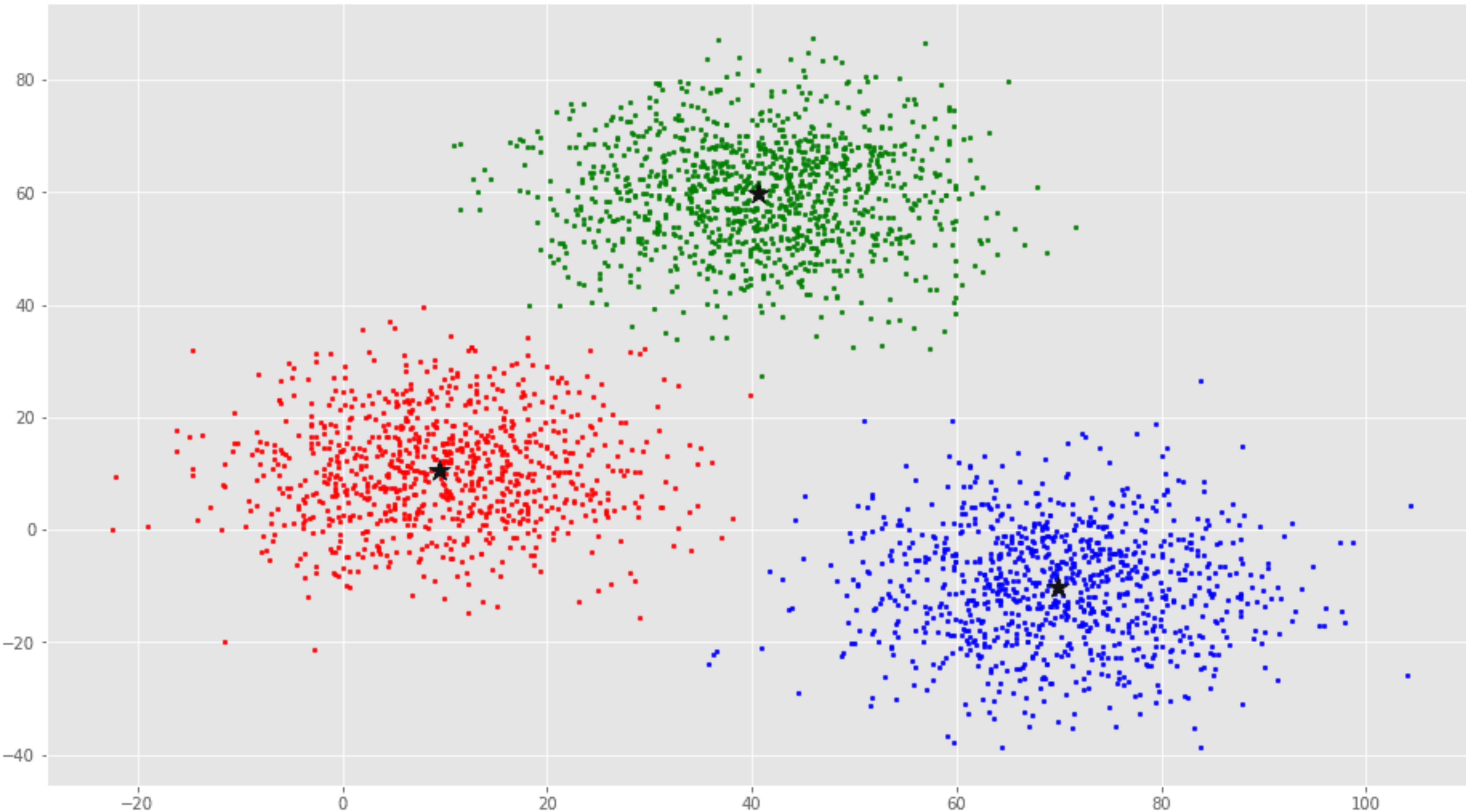
Example 3: image search

Clustering images



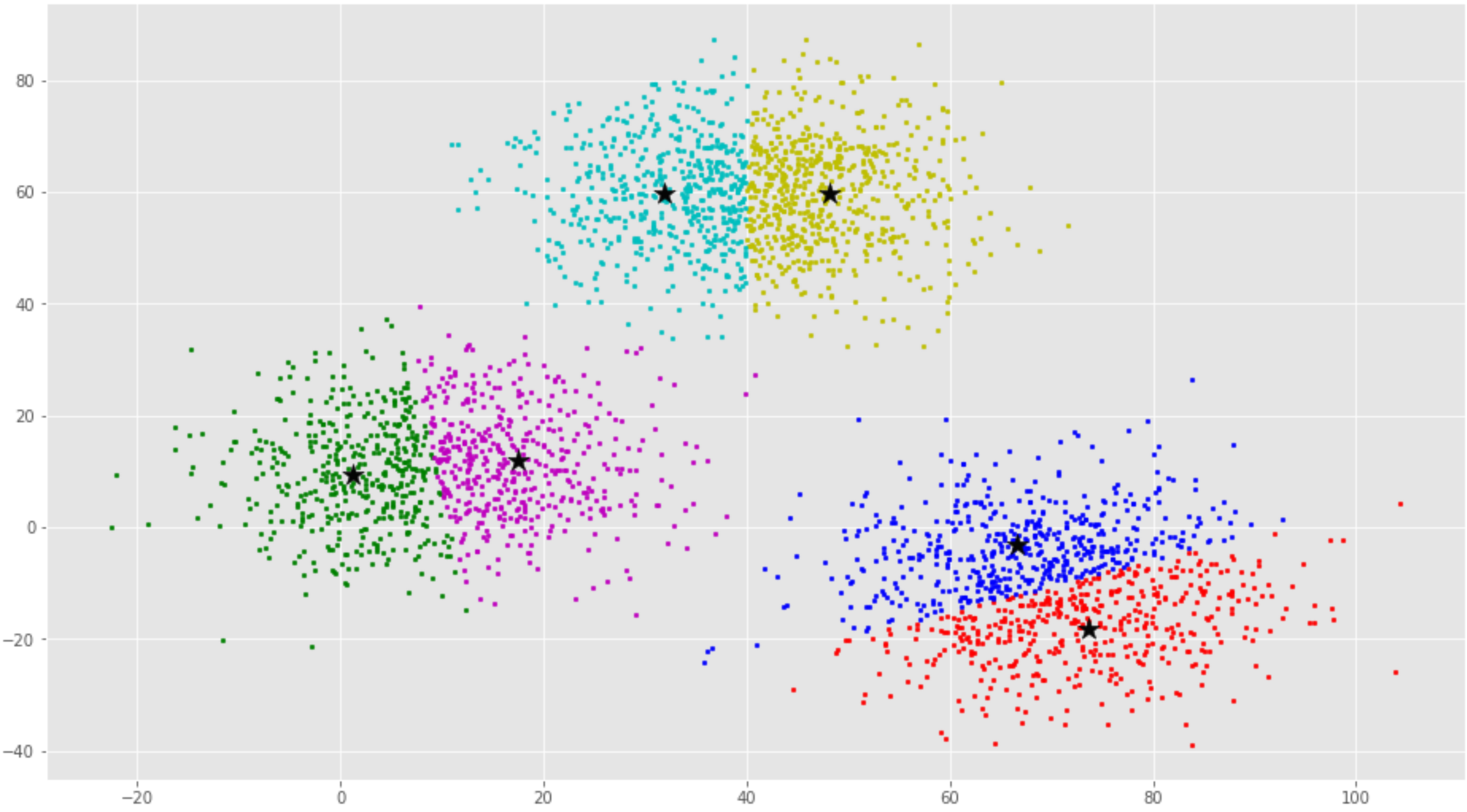
Find similar images to a target one

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

New content

Supervised Learning Tasks

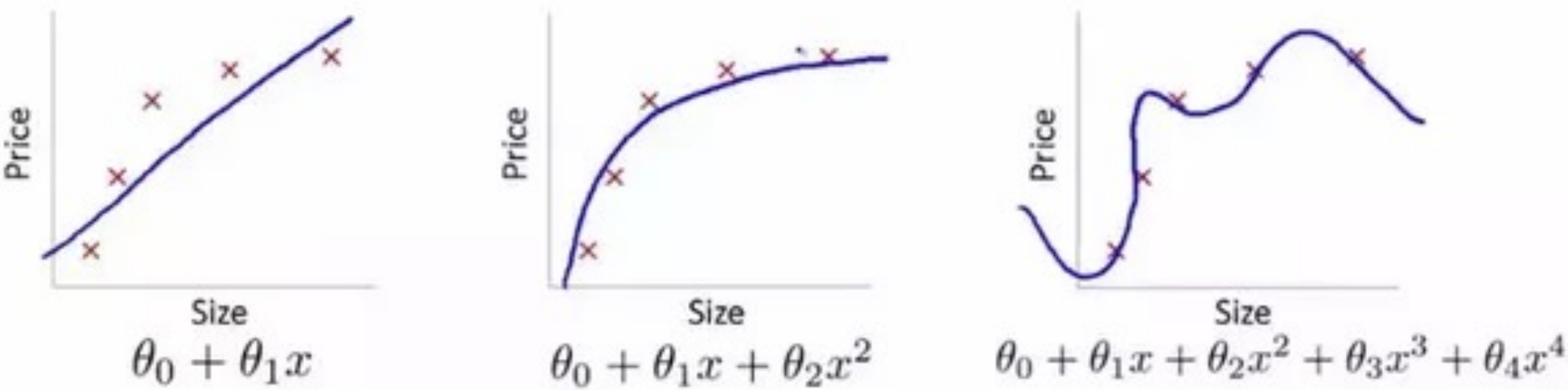
- Classification
  - Learn to predict class (discrete)
  - Minimize classification error
- Regression
  - Learn to predict response variable (numerical)
  - Minimize MSE (Mean Square Error)
- Both classification and regression
  - Training and testing phase
  - "Optimal" model is learned in training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

Learning Challenges

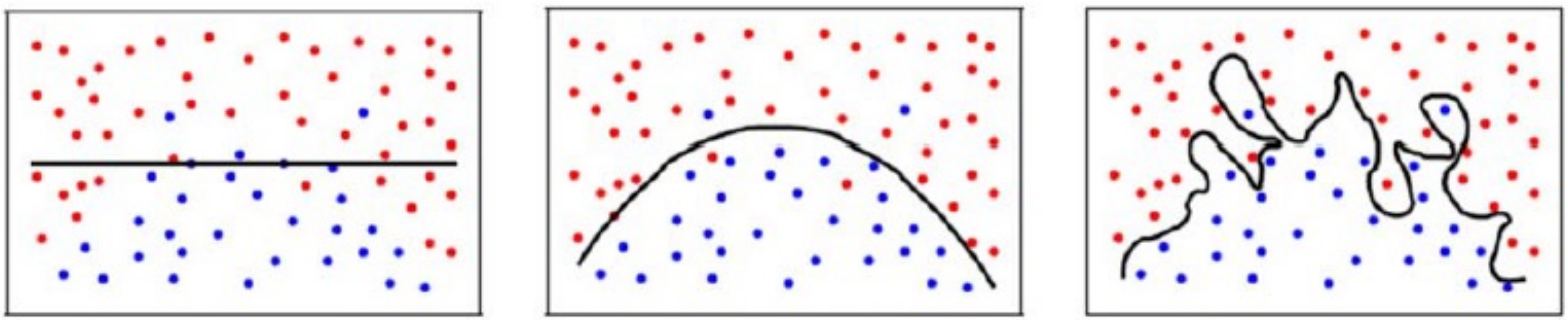
- Chapters 2.2.1 and 2.2.2 from ISL book
- Goal
  - Classify well new testing data
  - Model generalizes well to new testing data
  - Minimize error (MSE or classification error) in testing
- Variance
  - Amount by which model would change if we estimated it using a different training data set
- Bias
  - 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

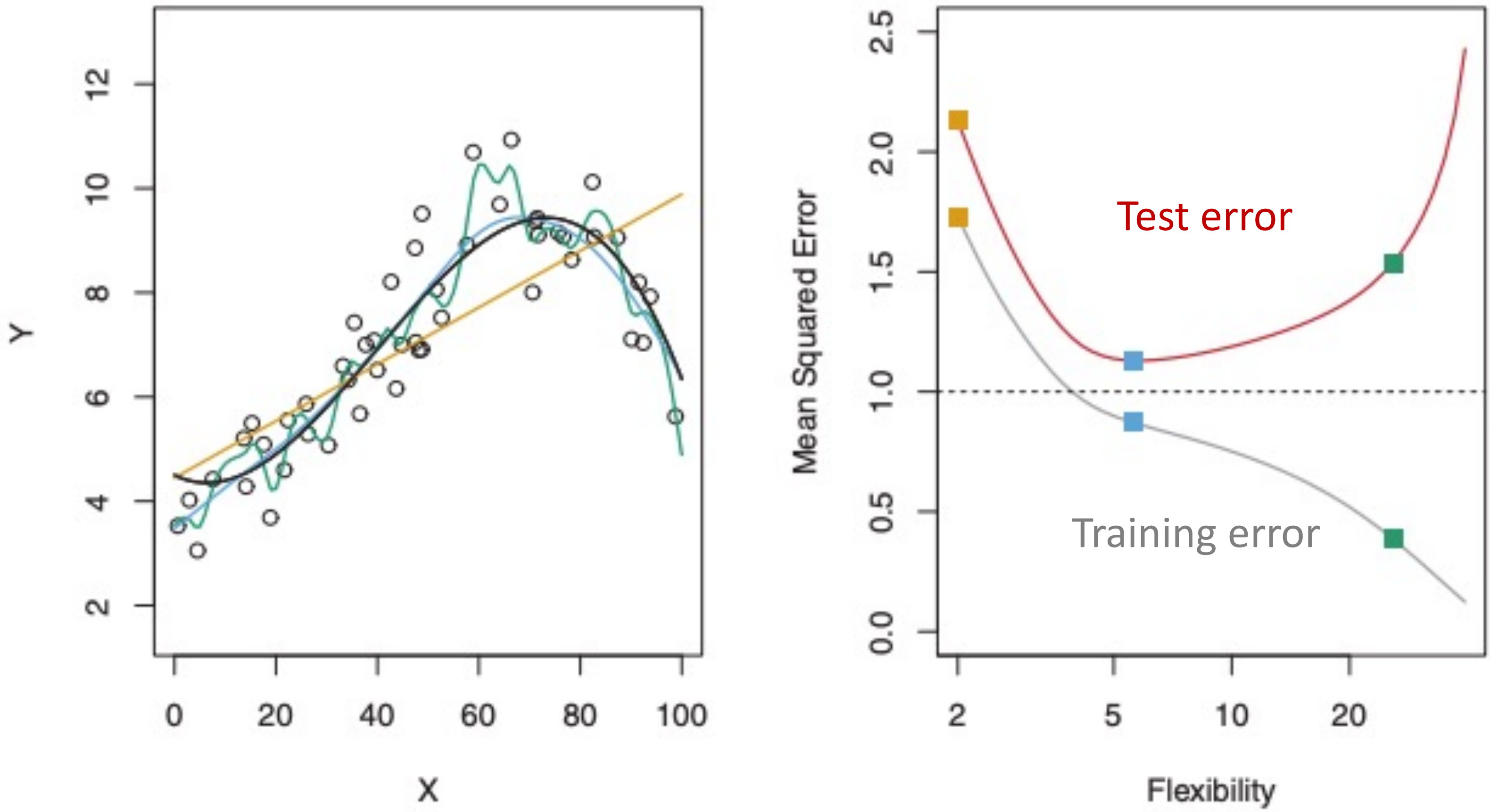


Generalization Problem in Classification



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

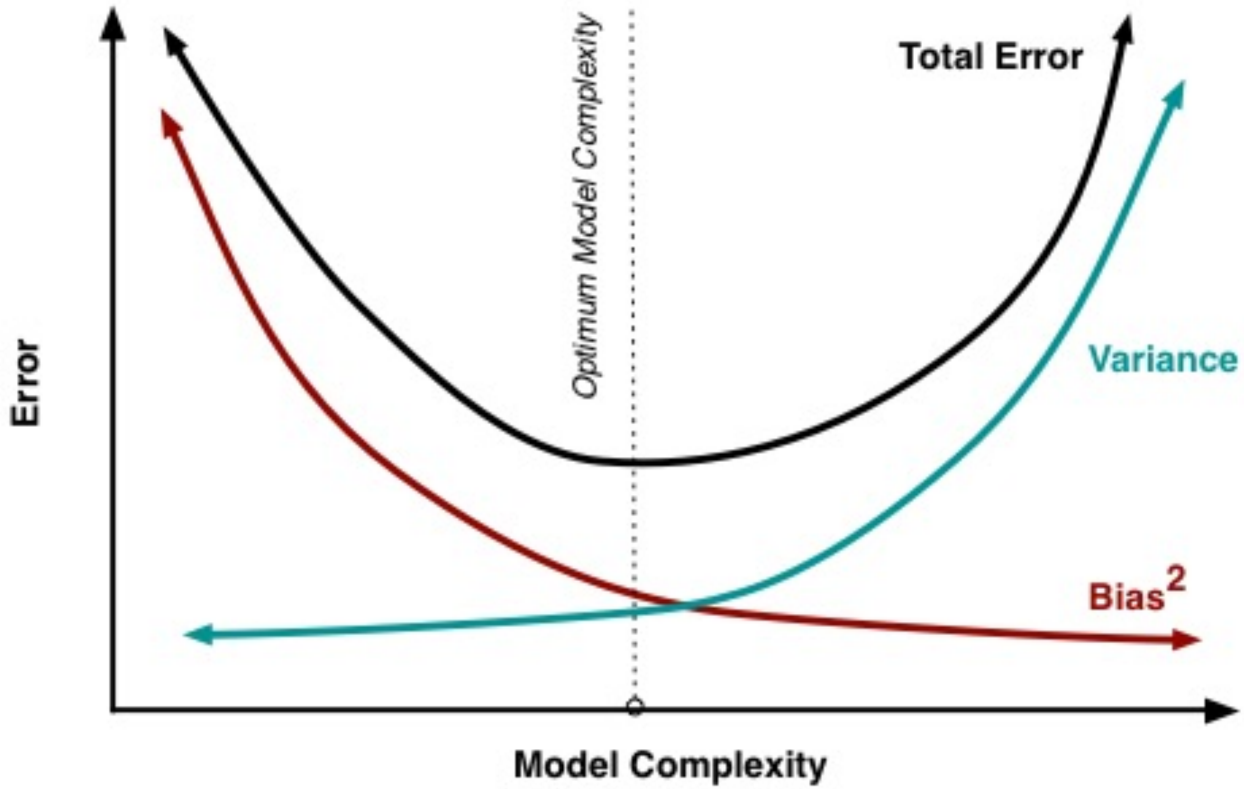
Training and testing error



ISL, Chapter 2.2.2

Bias-Variance Tradeoff

Testing



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)

Probability review

Probability Resources

- Review notes from Stanford's machine learning class
- David Blei's probability review
- Books:
  - Sheldon Ross, A First course in probability

Discrete Random Variables

- Let A denote a random variable
  - A represents an event that can take on certain values
  - Each value has an associated probability
- Examples of binary random variables:

  A = It will snow tomorrow

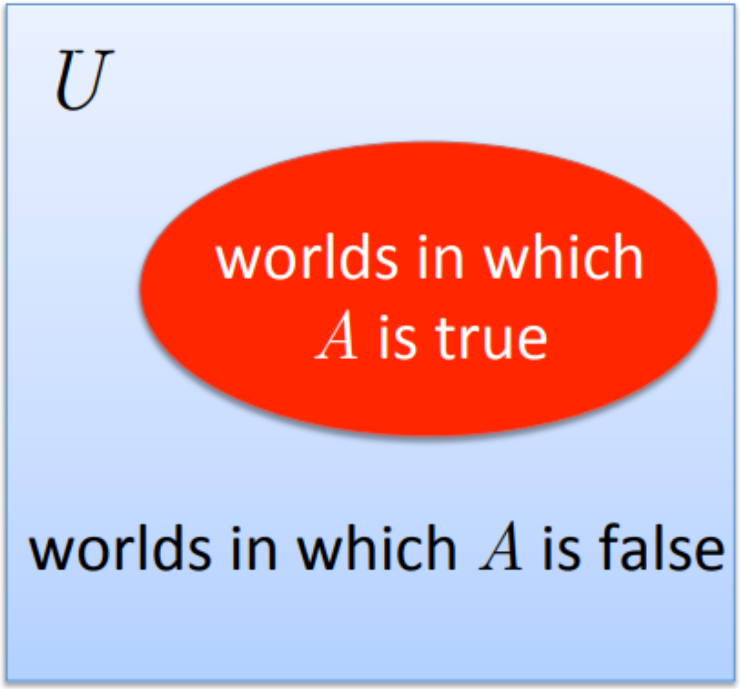
  B = The patient will recover
- $\mathrm{P}(A)$ is "the fraction of possible worlds in which A is true"

Visualizing A

- Universe U is the event space of all possible worlds
  - Its area is 1
  - $\mathrm{P}(U) = 1$
- $\mathrm{P}(A)$ = area of red oval
- Therefore:

$$P(A) + P(\neg A) = 1$$

$$P(\neg A) = 1 - P(A)$$



Working with Probabilities

- $0 \leq P(A) \leq 1$
- $P(U) = 1; P(\Phi) = 0$
- $P(\neg A) = 1 - P(A)$

Examples discrete RV

- Bernoulli RV
  - X is modelling a coin toss
  - Output: 1 (head) or 0 (tail)
  - P[X=1] = p; P[X=0] = 1-p
- Y is the number of points in a fair dice
  - $k \in \{1, \ldots, 6\}$, P[Y = k]=
  - P[Y = even] =

Example discrete RV

- Z is the sum of two fair dice
  - What is $\mathrm{P}[Z = k]$ for $k \in \{2, \ldots, 12\}$?
  - What is *k* for which this probability is maximum?

Expectation and variance

Expectation for discrete random variable X

$$\boxed{E[X] = \sum_{v} v Pr[X = v]}$$

Bernoulli: P[X=1] =p; P[X=0] = 1-p

Expectation and variance

Expectation for discrete random variable X

$$\boxed{E[X] = \sum_{v} v Pr[X = v]}$$

Properties

- $E[aX] = a\,E[X]$
- $E[X + Y] = E[X] + E[Y]$
- $E[f(X)] = \sum_v f(v) Pr[X = v]$

Variance: $\mathrm{Var}[X] = E\left[\left(X - E(X)\right)^2\right]$

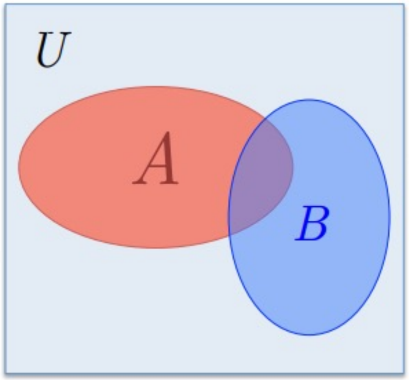
Variance of Bernoulli

- Variance: $\mathrm{Var}[X] = E(X^2) - E^2(X)$

Bernoulli: P[X=1] =p; P[X=0] = 1-p

Conditional Probability

- $\mathrm{P}(A \mid B)$ = Fraction of worlds in which B is true that also have A true



What if we already know that B is true?

That knowledge changes the probability of A

- Because we know we're in a world where B is true

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

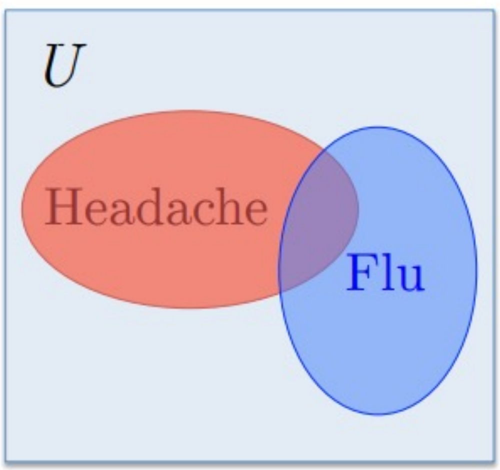
$$P(A \wedge B) = P(A \mid B) \times P(B)$$

Events A and B are **independent** if $\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$

Inference from Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$



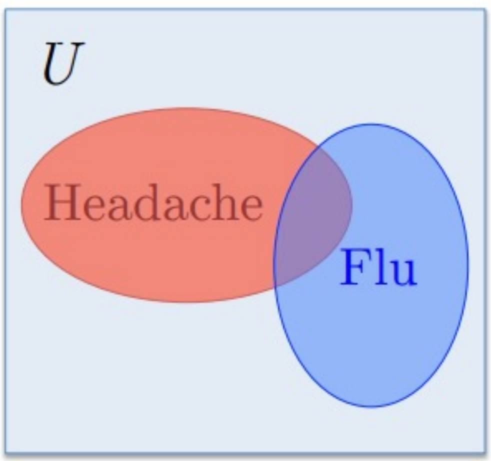
P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2

"Headaches are rare and flu is rarer, but if you're coming down with the flu there's a 50-50 chance you'll have a headache."

Inference from Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$



P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2

One day you wake up with a headache. You think: "Drat! 50% of flus are associated with headaches so I must have a 50-50 chance of coming down with flu."

Is this reasoning good?

Inference from Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$

$\mathrm{P(headache)}$ = 1/10
$\mathrm{P(flu)}$ = 1/40
$\mathrm{P(headache \mid flu)}$ = 1/2

Want to solve for:
$\mathrm{P(headache} \wedge \mathrm{flu)}$ = ?
$\mathrm{P(flu \mid headache)}$ = ?

Inference from Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$

$P(\text{headache}) = 1/10$
$P(\text{flu}) = 1/40$
$P(\text{headache} \mid \text{flu}) = 1/2$

Want to solve for:
$P(\text{headache} \wedge \text{flu}) = ?$
$P(\text{flu} \mid \text{headache}) = ?$

$$\begin{aligned} P(\text{headache} \wedge \text{flu}) &= P(\text{headache} \mid \text{flu}) \times P(\text{flu}) \\ &= 1/2 \times 1/40 = 0.0125 \end{aligned}$$

$$\begin{aligned} P(\text{flu} \mid \text{headache}) &= P(\text{headache} \wedge \text{flu}) \,/\, P(\text{headache}) \\ &= 0.0125 \,/\, 0.1 = 0.125 \end{aligned}$$

Bayes Theorem

Bayes' Rule

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$$

- Exactly the process we just used
- The most important formula in probabilistic machine learning

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

Multi-Value Random Variable

- Suppose A can take on more than 2 values
- A is a *random variable with arity k* if it can take on exactly one value out of $\{v_1, v_2, \ldots, v_k\}$
- Thus…

$$P(A = v_i \wedge A = v_j) = 0 \quad \text{if } i \neq j$$

$$P(A = v_1 \vee A = v_2 \vee \ldots \vee A = v_k) = 1$$

$$\boxed{1 = \sum_{i=1}^{k} P(A = v_i)}$$

EXAMPLE

Marginalization

- We can also show that:

$$P(B) = P(B \wedge [A = v_1 \vee A = v_2 \vee \ldots \vee A = v_k])$$

$$\boxed{P(B) = \sum_{i=1}^{k} P(B \wedge A = v_i)} = \sum_{i=1}^{k} P(B \mid A = v_i) P(A = v_i)$$

- This is called **marginalization** over A

EXAMPLE