Machine Learning

Introduction to Machine Learning

Marek Petrik

January 26, 2017

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

What is machine learning?

Arthur Samuel (1959, IBM):

> *Field of study that gives computers the ability to learn without being explicitly programmed*

The rise of machine learning

ICML: International Conference on Machine Learning

2009

500 attendees

2015

ICML Lille

International Conference on Machine Learning

06 - 11 JULY 2015
LILLE GRAND PALAIS

1600 attendees

2016

ICML@NYC

International Conference on Machine Learning

JUNE 19-24 2016 NEW YORK

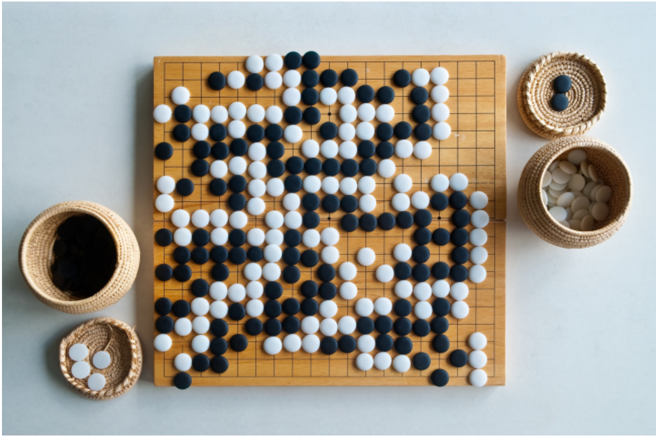
3300 attendees

Data is everywhere!

IBM Watson: Computers Beat Humans in Jeopardy

Fair use, https://en.wikipedia.org/w/index.php?curid=31142331

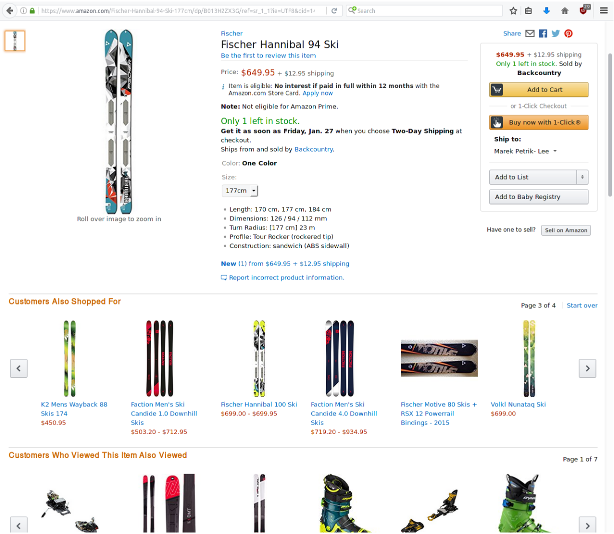
AlphaGo: Computers Beat Humans in Go



Photograph by Saran Poroong ff Getty Images/iStockphoto

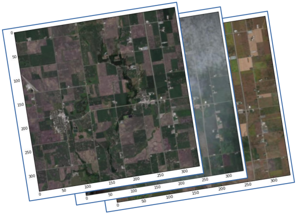
Personalized Product Recommendations

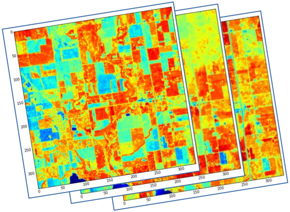
Online retailers mine purchase history to make recommendations

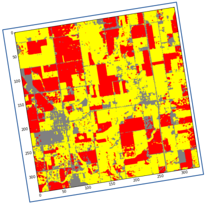


Identifying Crops from Space

Identify the type of crop / vegetation / urban space from satellite observations







Other Applications

1. Health-care: Identify risks of getting a disease
2. Health-care: Predict effectiveness of a treatment
3. Detect spam in emails
4. Recognize hand-written text
5. Speech recognition (speech to text)
6. Machine translation
7. Predict probability of an employee leaving

Activity: any other applications?

This Course: Introduction to Machine Learning

- Build a foundation for practice and research in ML
- Basic machine learning concepts: max likelihood, cross validation
- Fundamental machine learning techniques: regression, model-selection, deep learning
- Educational goals:
  1. How to apply basic methods
  2. Reveal what happens inside
  3. What are the pitfalls
  4. Expand understanding of linear algebra, statistics, and optimization

Course Overview

- Grading:

50%	6 Assignments
15%	Midterm exam
30%	Final exam
15%	Class project

- Assignments: posted on myCourses and the website
- Discuss questions: `https://piazza.com/unh/spring2017/cs780cs880`
- Programming language: R (or Python, but discouraged)

Course Texts

- Textbooks:
  - ISL James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning
  - ELS Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer Series in Statistics (2nd ed.)
- Other sources:
  - DL Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. [online pdf on github]
  - LA Strang, G. Introduction to Linear Algebra. (2016)
  - CO Boyd, S., & Vandenberghe, L. (2004). Convex Optimization. [online pdf]
  - RL Sutton, R. S., & Barto, A. (2012). Reinforcement learning. 2nd edition [online pdf draft]

Class Plan I

Date	Day	Topic
Jan 24	Tue	Statistical learning and R language
Jan 26	Thu	Linear regression with one variable
Jan 31	Tue	Linear regression with multiple variables
Feb 02	Thu	*No class*
Feb 07	Tue	Classification and logistic regression 1
Feb 09	Thu	Classification, naive Bayes and LDA
Feb 14	Tue	Linear algebra for machine learning: review
Mar 16	Thu	Linear algebra and optimization
Feb 21	Tue	Overfitting and resampling methods
Feb 23	Thu	Cross-validation and bootstrapping
Feb 28	Tue	Linear model selection, priors
Mar 02	Thu	Linear model selection and regularization

Class Plan II

Mar 07	Tue	Midterm review
Mar 09	Thu	**Midterm exam**; material until 2/23
Mar 14	Tue	*Spring break, no class*
Mar 16	Thu	*Spring break, no class*
Mar 21	Tue	Building nonlinear features
Mar 23	Thu	Nearest neighbor methods and GAMs
Mar 28	Tue	Tree-based methods and boosting
Mar 30	Thu	Support vector machines and other techniques
Apr 04	Tue	Unsupervised learning, PCA
Apr 06	Thu	Unsupervised learning, k-means
Apr 11	Tue	Reinforcement learning
Apr 13	Thu	Neural networks and deep learning
Apr 18	Tue	Neural networks and deep learning
Apr 20	Thu	Big data and machine learning

Class Plan III

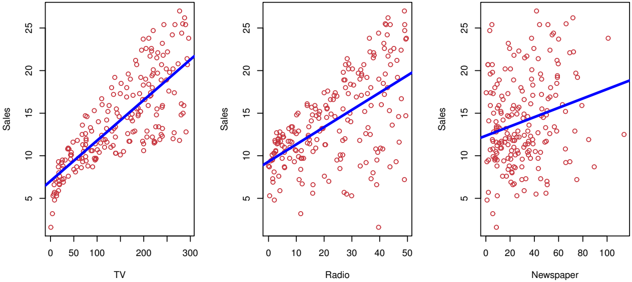
Apr 25	Tue	Machine learning in practice
Apr 27	Thu	Project presentations
May 02	Tue	Guest speaker
May 04	Thu	Final exam review
May 11-17	?	**Final exam**

What is Machine Learning

- Discover unknown function f:

$$Y = f(X)$$

- X = set of features, or inputs
- Y = target, or response



$$\text{Sales} = f(\text{TV}, \text{Radio}, \text{Newspaper})$$

Notation

$$Y = f(X) = f(X_1, X_2, X_3)$$
$$\text{Sales} = f(\text{TV}, \text{Radio}, \text{Newspaper})$$

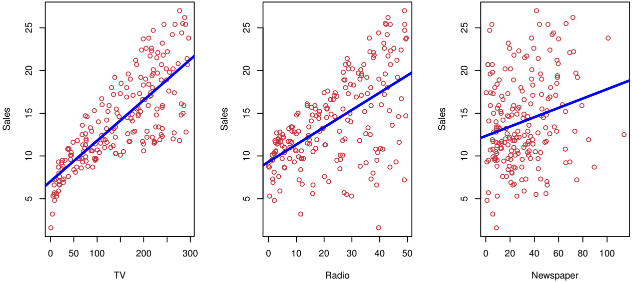
- $Y =$ Sales
- $X_1 =$ TV
- $X_2 =$ Radio
- $X_3 =$ Newspaper

Vector:

$$X = (X_1, X_2, X_3)$$

Dataset

$$\text{Sales} = f(\text{TV}, \text{Radio}, \text{Newspaper})$$



Dataset:

Y	X_1	X_2	X_3
10	101	20	35
20	66	41	85
11	101	43	78
25	25	10	61
5	310	51	11

rows are samples

Errors in Machine Learning: World is Noisy

- World is too complex to model precisely
- Many features are not captured in data sets
- Need to allow for errors ϵ in f:

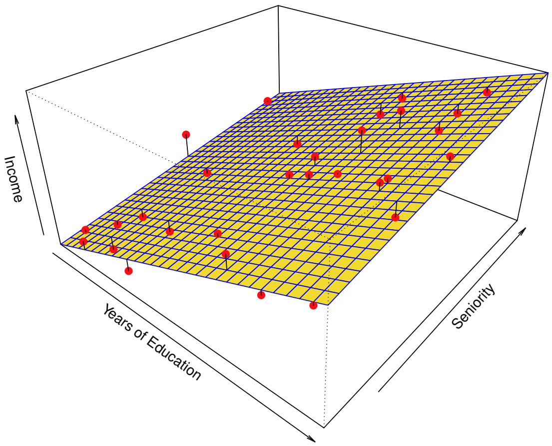
$$Y = f(X) + \epsilon$$

Machine Learning Algorithm

- **Input**:
  Training data-set with features and targets

- **Output**:
  Prediction function f

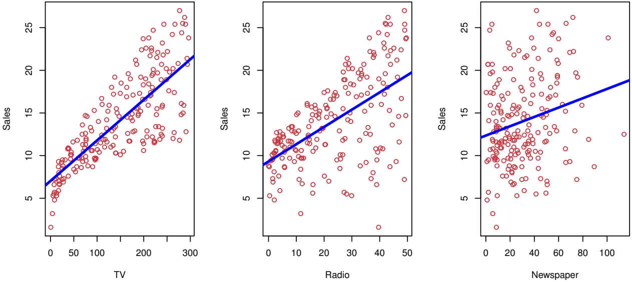
Parametric Prediction Methods



Linear models (linear regression)

$$\text{income} = f(\text{education}, \text{seniority}) = \beta_0 + \beta_1 \times \text{education} + \beta_2 \times \text{seniority}$$

Why Estimate f?



1. **Prediction**: Make predictions about future: Best medium mix to spend ad money?
2. **Inference**: Understand the relationship: What kind of ads work? Why?

Prediction or Inference?

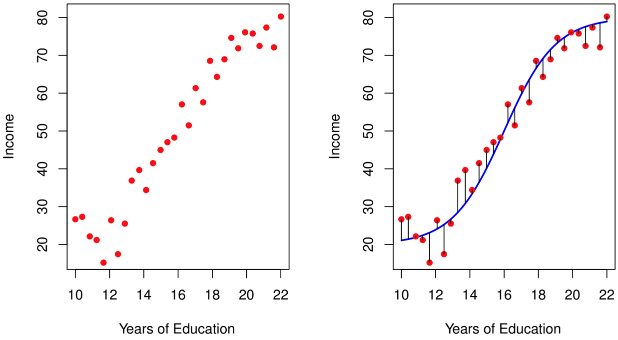
Application	Prediction	Inference
Identify risk of getting a disease		
Predict effectiveness of a treatment		
Recognize hand-written text		
Speech recognition		
Predict probability of an employee leaving		

Prediction or Inference?

Application	Prediction	Inference
Identify risk of getting a disease	✓	✓
Predict effectiveness of a treatment	✓	✓
Recognize hand-written text	✓	
Speech recognition	✓	
Predict probability of an employee leaving	✓	✓

Statistical View of Machine Learning

- Probability space Ω: Set of all adults
- Random variable: $X(\omega) = \mathbb{R}$: Years of education
- Random variable: $Y(\omega) = \mathbb{R}$: Salary



How Good are Predictions?

- Learned function $\hat{f}$
- Test data: $(x_1, y_1), (x_2, y_2), \ldots$
- **Mean Squared Error (MSE)**:

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

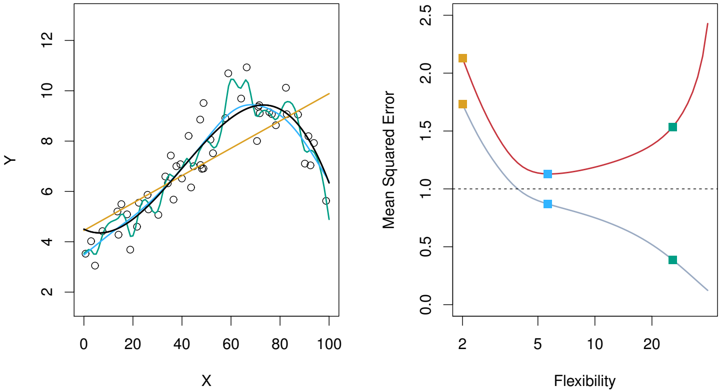
- This is the estimate of:

$$\text{MSE} = \mathbb{E}[(Y - \hat{f}(X))^2] = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} (Y(\omega) - \hat{f}(X(\omega)))^2$$

- Important: Samples x_i are i.i.d.

Do We Need Test Data?

- Why not just test on the training data?



- Flexibility is the degree of polynomial being fit
- <u>Gray line</u>: training error, <u>red line</u>: testing error

Bias-Variance Decomposition

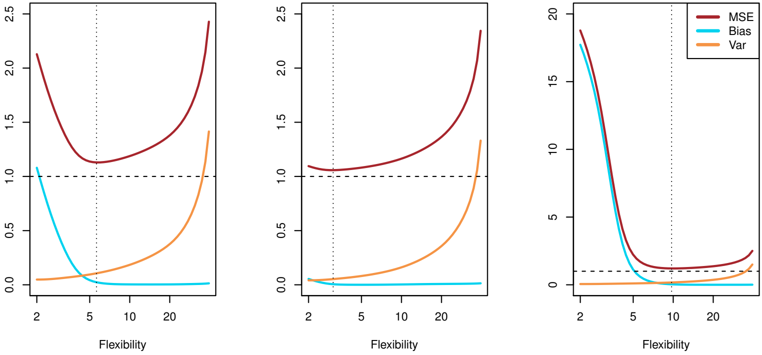
$$Y = f(X) + \epsilon$$

Mean Squared Error can be decomposed as:

$$\mathsf{MSE} = \mathbb{E}(Y - \hat{f}(X))^2 = \underbrace{\mathrm{Var}(\hat{f}(X))}_{\text{Variance}} + \underbrace{(\mathbb{E}(\hat{f}(X)))^2}_{\text{Bias}} + \mathrm{Var}(\epsilon)$$

- **Bias**: How well would method work with infinite data
- **Variance**: How much does output change with different data sets

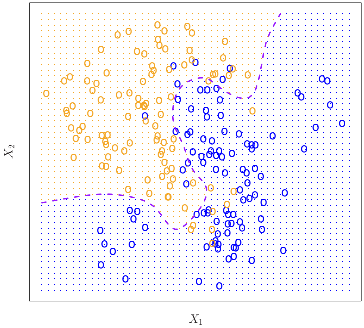
Bias-Variance Trade-off



Types of Function f

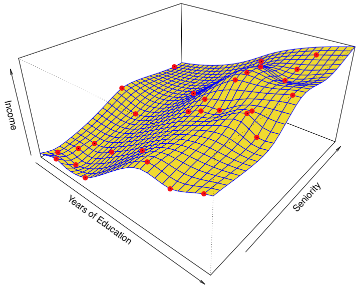
Classification: discrete target

$$f : \mathcal{X} \to \{1, 2, 3, \ldots, k\}$$



Regression: continuous target

$$f : \mathcal{X} \to \mathbb{R}$$



Regression or Classification?

Application	Regression	Classification
Identify risk of getting a disease		
Predict effectiveness of a treatment		
Recognize hand-written text		
Speech recognition		
Predict probability of an employee leaving		

Regression or Classification?

Application	Regression	Classification
Identify risk of getting a disease	✓	
Predict effectiveness of a treatment	✓	
Recognize hand-written text		✓
Speech recognition		✓
Predict probability of an employee leaving	✓	

Error Rate In Classification

- Learned function $\hat{f}$
- Test data: $(x_1, y_1), (x_2, y_2), \ldots$

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

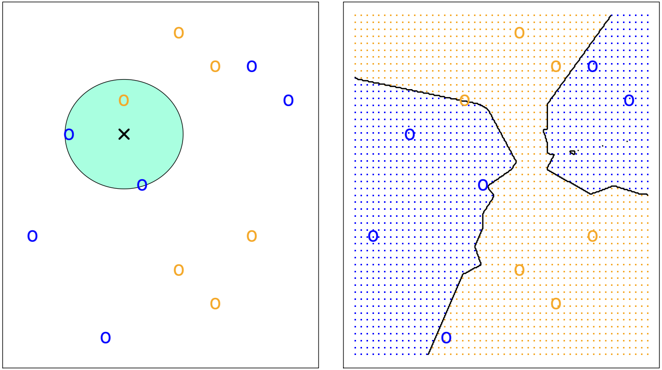
- Bayes classifier: assign each observation to the most likely class

$$f(x) = \Pr[Y = j \mid X = x]$$

- Bayes classifier would require known true function f
- Lower bound on the error

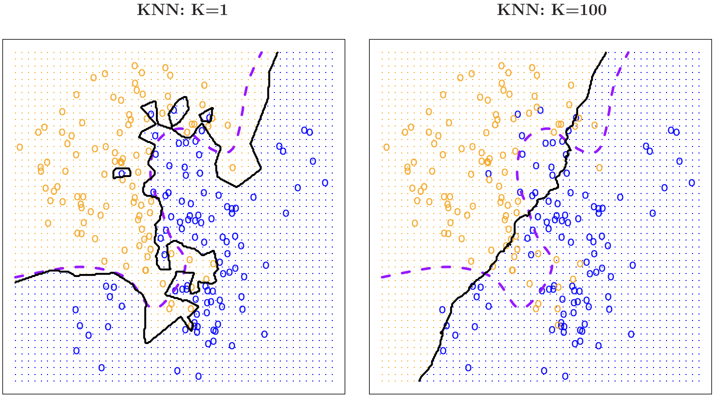
KNN: K-Nearest Neighbors

- Bayes classifier can only predict the values x in the training set
- <u>Idea</u>: Use similar training points when making predictions

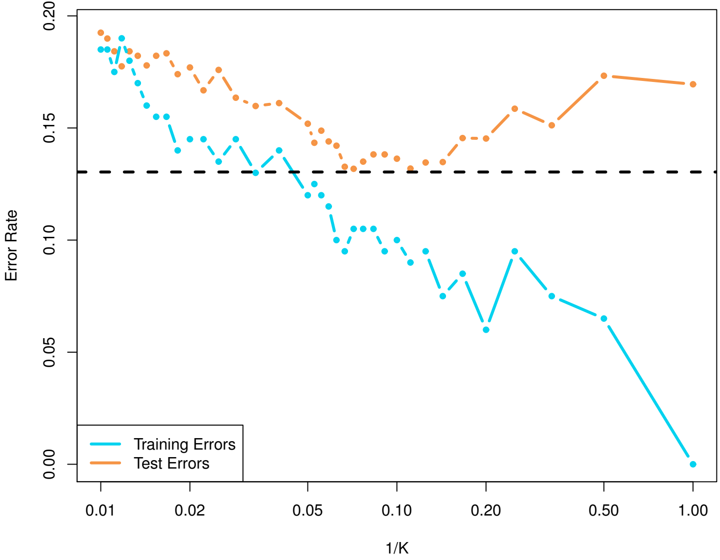


- Non-parametric method (unlike regression)

KNN: Choosing k



KNN: Training and Test Errors



R Language

- Download and install R: `http://cran.r-project.prg`
- Try using RStudio as an R IDE
- Read the R lab: ISL 2.3
- Use Piazza for questions

Alternatives to R

- **Python** (+ Numpy + Pandas + Matplotlib + Scikits Learn + Scipy)
  More popular, more flexible, but less mature
- **MATLAB**
  Linear algebra and control focus, but a lot of ML toolkits
- **Julia**
  New technical computing language, “ walks like python runs like c”, but least mature