Introduction to Machine Learning

Mingjun Zhong

Department of Computing Science

University of Aberdeen

Adapted from A. Zisserman, Geert Jan Bex, Jan Ooghe, and Ehsan Moravveji’s slides

Today

- Course practicalities
- Learning outcomes:
  - What is ML?
  - The four canonical ML problems
  - ML frameworks

Practicalities

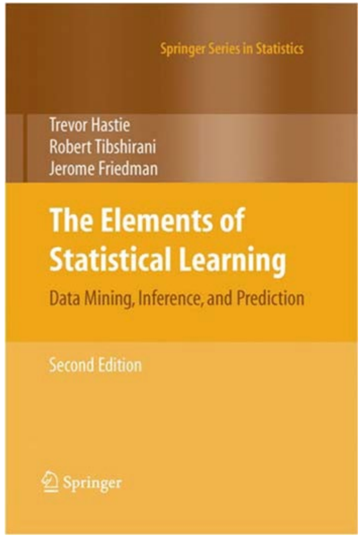
- Lecturers: Yaji Sripada & Mingjun Zhong
- Lectures: On campus (Teaching weeks 9-19)
- Practical:

	Day	From	To	Room	Week
Lecture	Friday	12:00	13:00	Fraser Noble 3	9-19
Practical	Tuesday	14:00	16:00	MR107 MacRobert Building	10-19
Practical	Friday	10:00	12:00	MR107 MacRobert Building	10-19

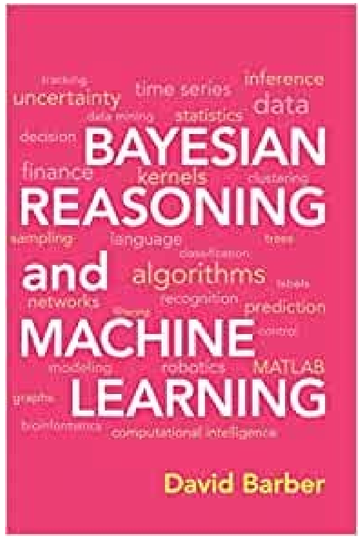
- Assessments: MCQ class test (25%) + Individual Project Report (75%). Late hand-in penalties as per the programme handbook.
- You should
  - Be able to program using Python
  - Use a bit of linear algebra + calculus

Recommended books

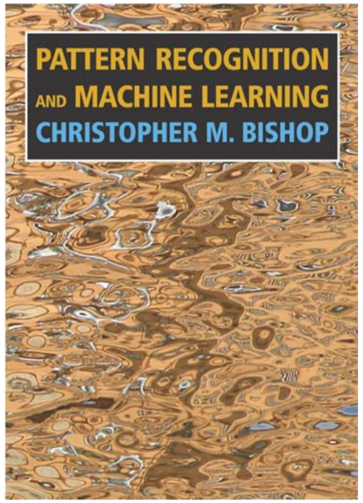
- Elements of Statistical Learning
- Hastie, Tibshirani, Friedman
- Springer 2009, second edition
- Good explanations


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

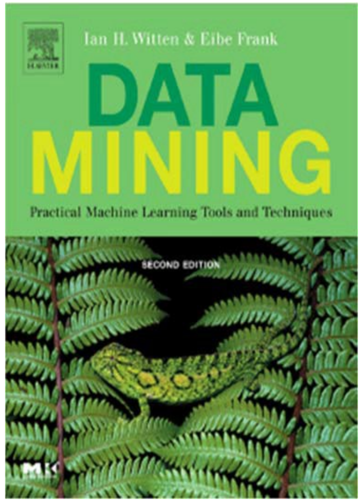
- Bayesian Reasoning and Machine Learning
- David Barber
- Cambridge University Press
- Good for computer scientists



- Pattern Recognition and Machine Learning
- Christopher Bishop
- Springer 2006
- Good explanations on classification and regression



- Data Mining: Practical Machine Learning Tools and Techniques
- Ian Witten & Eibe Frank
- Morgan Kaufmann, 2005
- Readable and practical guide


Ian H. Witten & Eibe Frank
DATA MINING
Practical Machine Learning Tools and Techniques
SECOND EDITION

What is Machine Learning?

- Approaches/methodologies to artificial intelligence (AI)
- Making computers be intelligent (thinking and/or acting)
- Making the data do the work
- Study/development of computer algorithms that can learn knowledge for AI through experience and by the use of data
- Typically, these algorithms have a large number of parameters whose values are learnt from data
- Applications (for which it is impossible to define rules by hand):
    - Face detection
    - Image classification
    - Weather prediction
    - Stock prediction
    - Speech recognition
    - Etc.

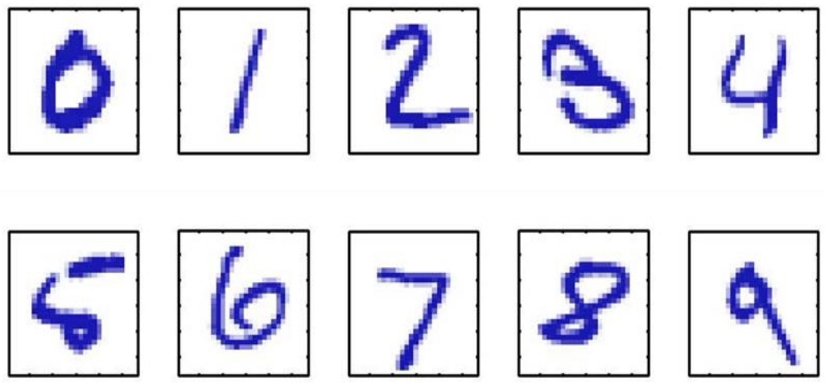
Why Machine Learning?

- *We are drowning in information and starving for knowledge*. –John Naisbitt.
- **Era of big data**:
  - In 2025, there are over 1.1 billion websites
  - 518,000 hours of video are uploaded to YouTube every day
  - Amazon ships about 12.9 million packages a day
- **No human being can deal with the data avalanche**!
- **Machine Learning is a key tool for every data scientist!**

Example: hand-written digit recognition

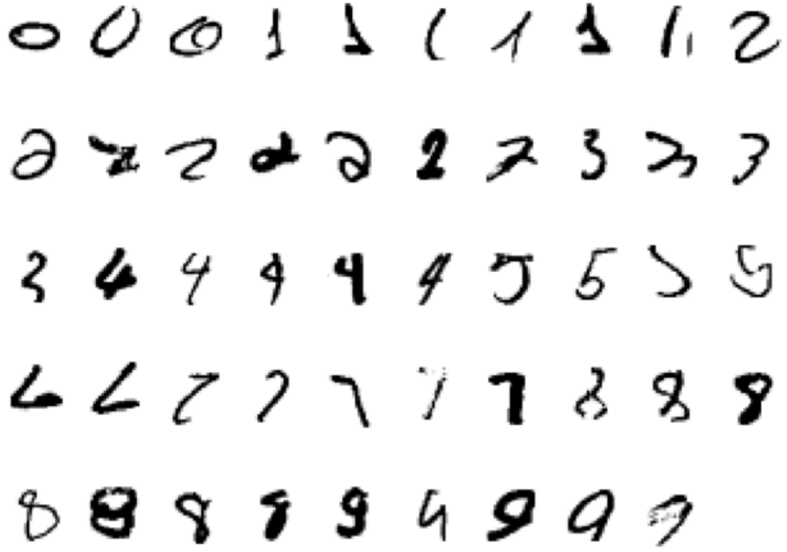
- Images are 28*28 pixels
- Represent input image as a vector $x \in R^{784}$
- Learn a classifier $f\,(\mathrm{x})$ such that,

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

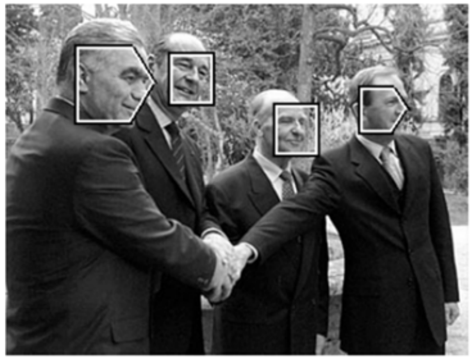


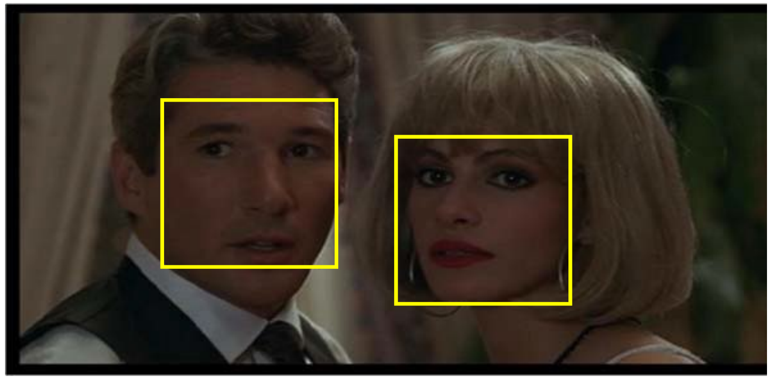
Example: hand-written digit recognition

- How to learn a classifier?
- 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 the first commercial and widely used ML systems (for zip codes & checks)



Example: face detection





- Again, a supervised classification problem
- Need to classify an image window into three classes:
  - non-face
  - frontal-face
  - profile-face

Example: face detection

Classifier is learnt from labelled data

Training data for frontal faces

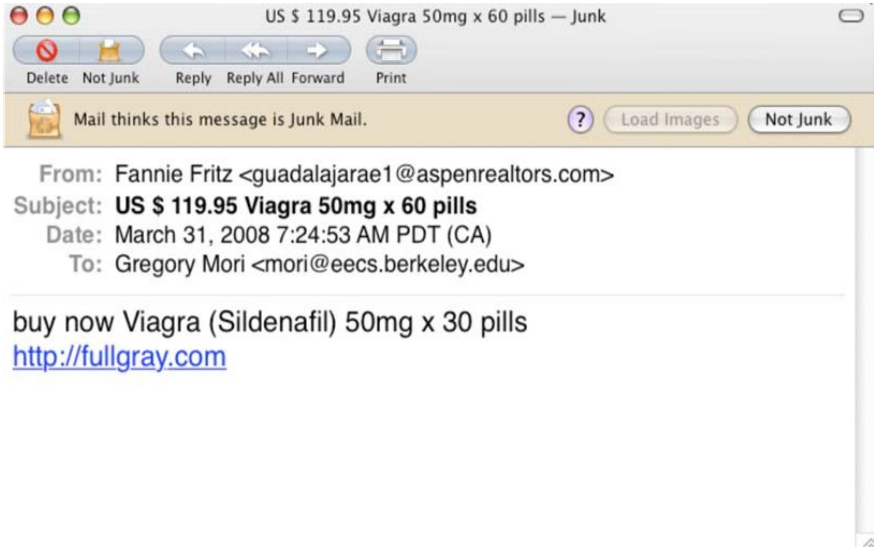
- 5000 faces
  - All near frontal
  - Age, race, gender, lighting
- 10^8 non faces
- faces are normalized
  - scale, translation

Example: Spam detection

- This is a classification problem
- Task is to classify email into spam/non-spam

•Data x_i is word count, e.g. of 'Viagra', 'outperform', “you may be surprized to be contacted” …

- Requires a learning system as “enemy” keeps innovating

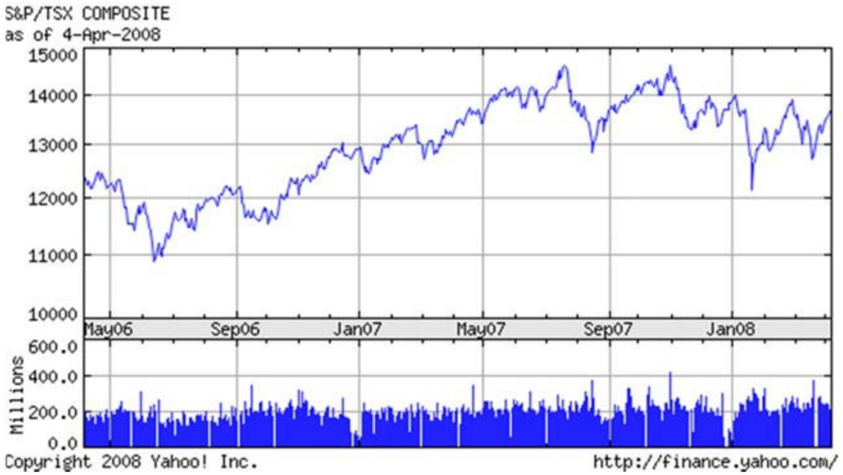

US $ 119.95 Viagra 50mg x 60 pills — Junk
Delete
Not Junk
Reply
Reply All
Forward
Print
Mail thinks this message is Junk Mail.
?
Load Images
Not Junk

From: Fannie Fritz <guadalajarae1@aspenrealtors.com>
Subject: **US $ 119.95 Viagra 50mg x 60 pills**
Date: March 31, 2008 7:24:53 AM PDT (CA)
To: Gregory Mori <mori@eecs.berkeley.edu>

buy now Viagra (Sildenafil) 50mg x 30 pills
http://fullgray.com

Example: Stock price prediction

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

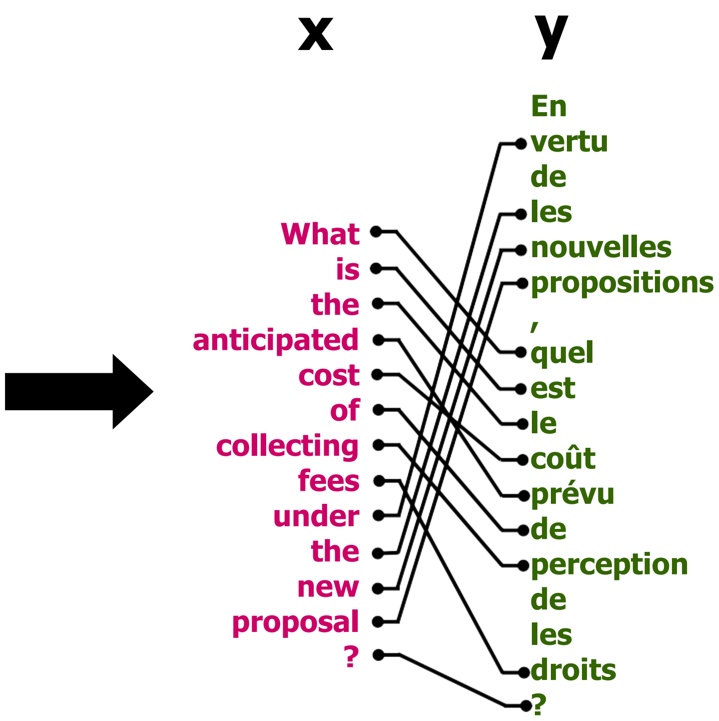


Example: machine translation

Use of aligned text

What is the anticipated cost of collecting fees under the new proposal?

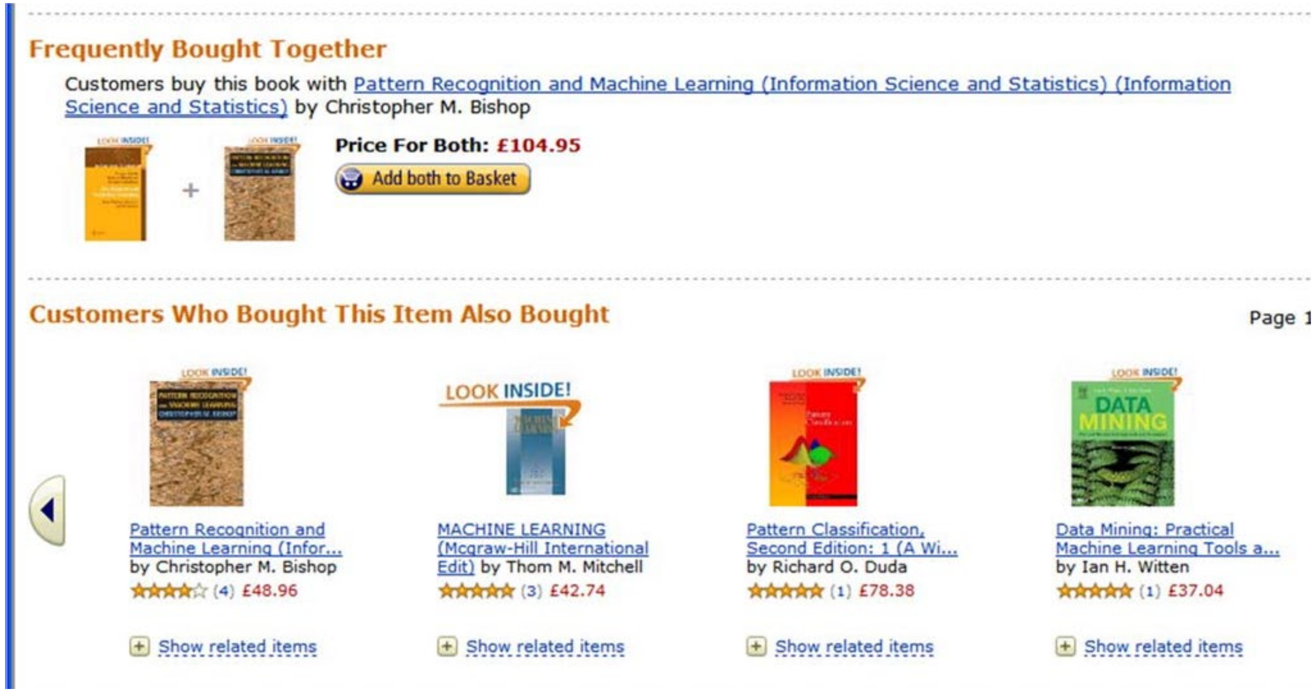
En vertu des nouvelles propositions, quel est le coût prévu de perception des droits?



e.g. Google translate

Example: recommender systems

People who bought Hastie’s book …



Frequently Bought Together

Customers buy this book with Pattern Recognition and Machine Learning (Information Science and Statistics) (Information Science and Statistics) by Christopher M. Bishop

Price For Both: £104.95

Customers Who Bought This Item Also Bought

Page 1

Pattern Recognition and Machine Learning (Infor... by Christopher M. Bishop ★★★★☆ (4) £48.96

⊞ Show related items

MACHINE LEARNING (Mcgraw-Hill International Edit) by Thom M. Mitchell ★★★★★ (3) £42.74

⊞ Show related items

Pattern Classification, Second Edition: 1 (A Wi... by Richard O. Duda ★★★★★ (1) £78.38

⊞ Show related items

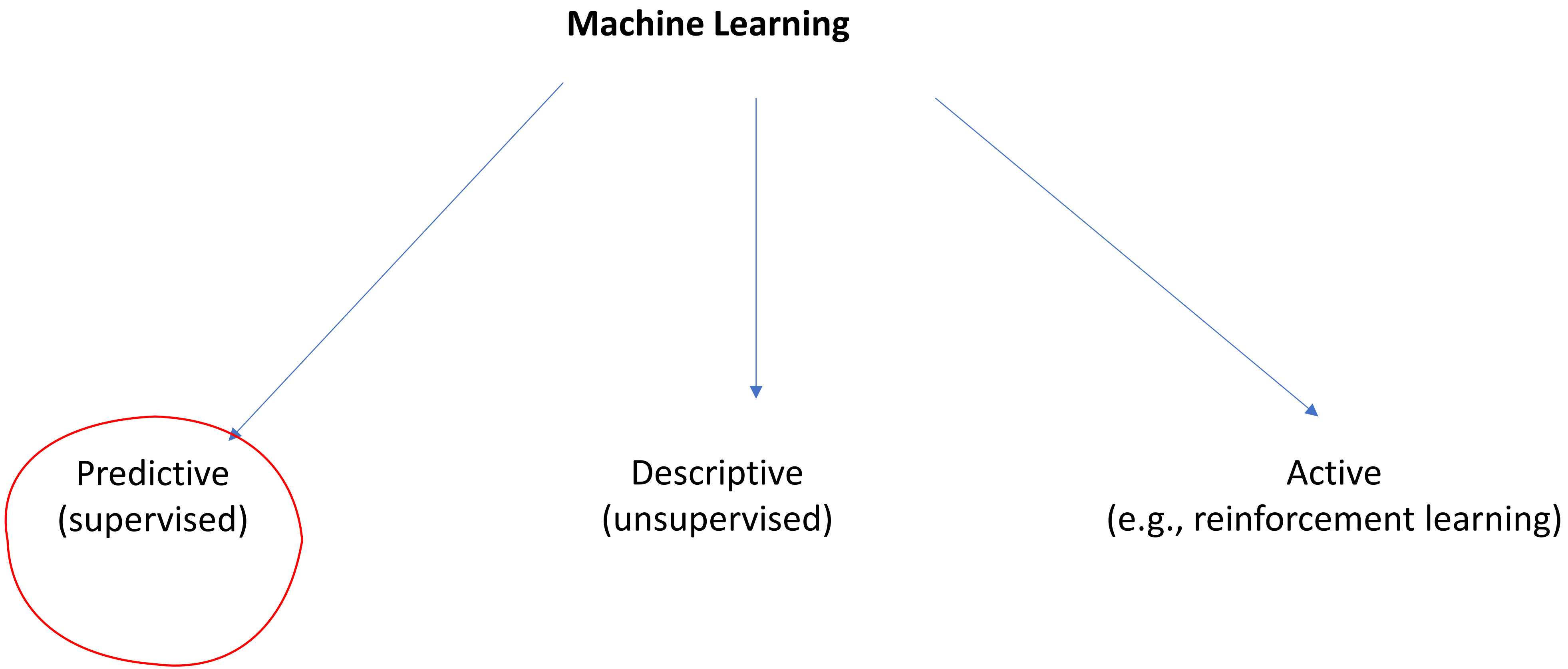
Data Mining: Practical Machine Learning Tools a... by Ian H. Witten ★★★★★ (1) £37.04

⊞ Show related items

Example: AlphaGo

A computer program that plays the board game Go

Types of Machine Learning



Machine Learning

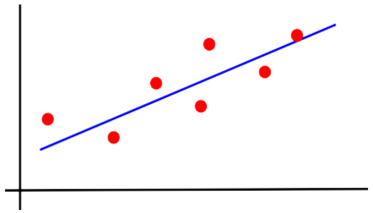
Predictive
(supervised)

Descriptive
(unsupervised)

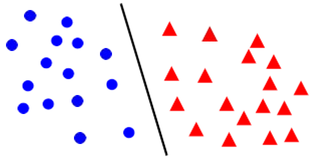
Active
(e.g., reinforcement learning)

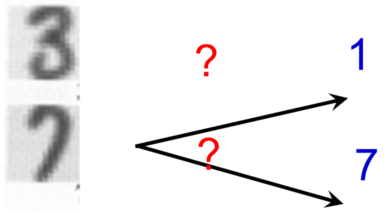
Four canonical ML problems

1. Regression - supervised
   - estimate parameters, e.g. of weight vs height

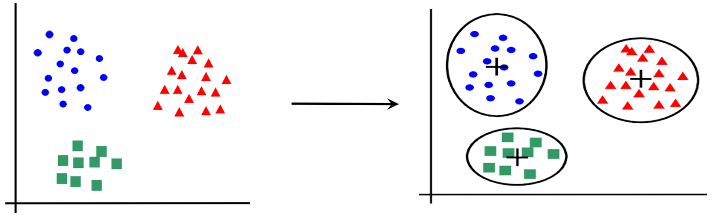


2. Classification - supervised
   - estimate class, e.g. handwritten digit classification

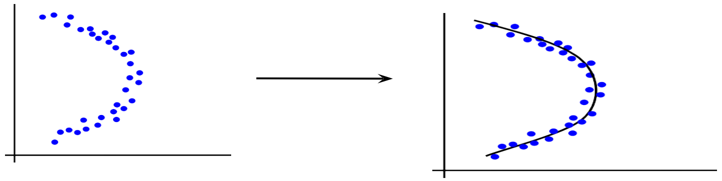




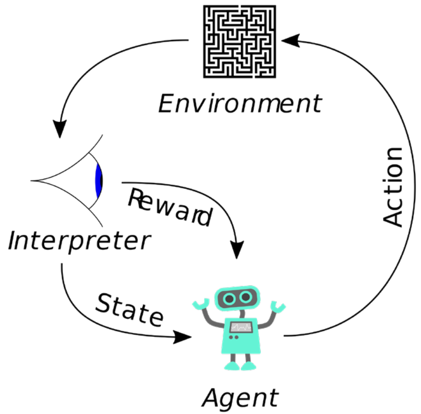
3. Unsupervised learning – model the data
   - clustering



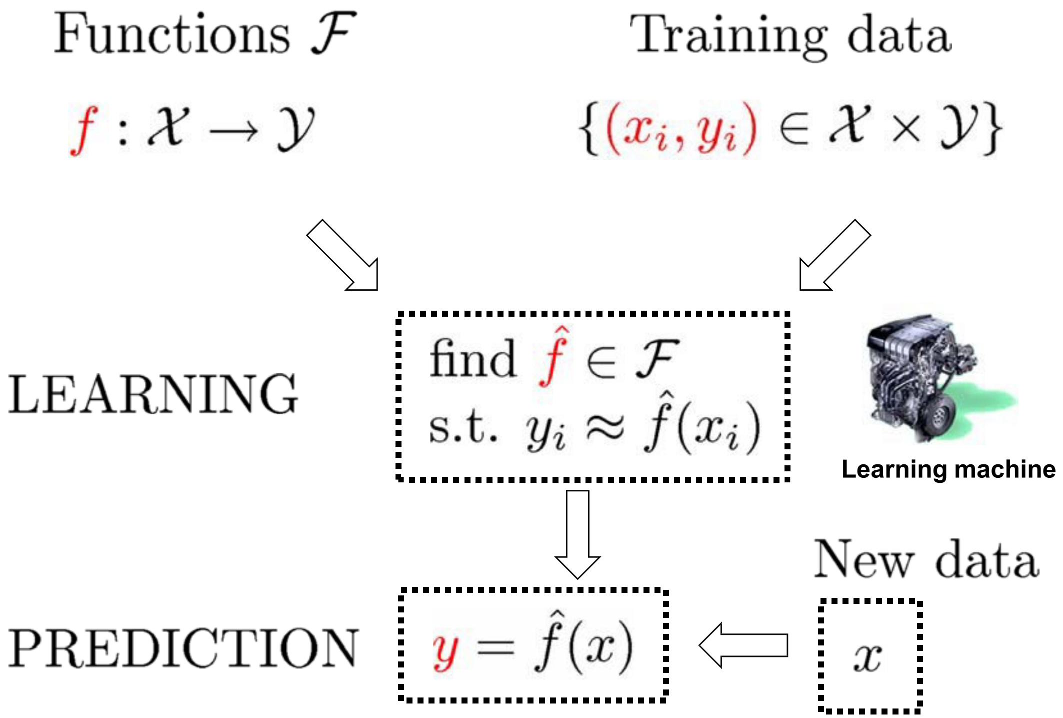
   - dimensionality reduction



4. Reinforcement learning
   - decision making



Supervised learning: an overview



Functions $\mathcal{F}$

$f : \mathcal{X} \to \mathcal{Y}$

Training data

$\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}$

LEARNING

find $\hat{f} \in \mathcal{F}$

s.t. $y_i \approx \hat{f}(x_i)$

Learning machine

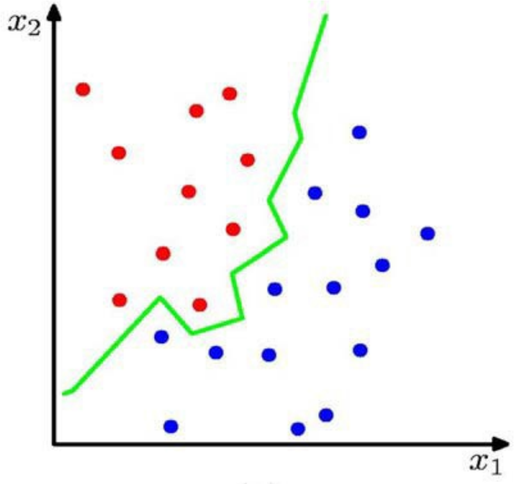
New data

x

PREDICTION

$y = \hat{f}(x)$

Classification

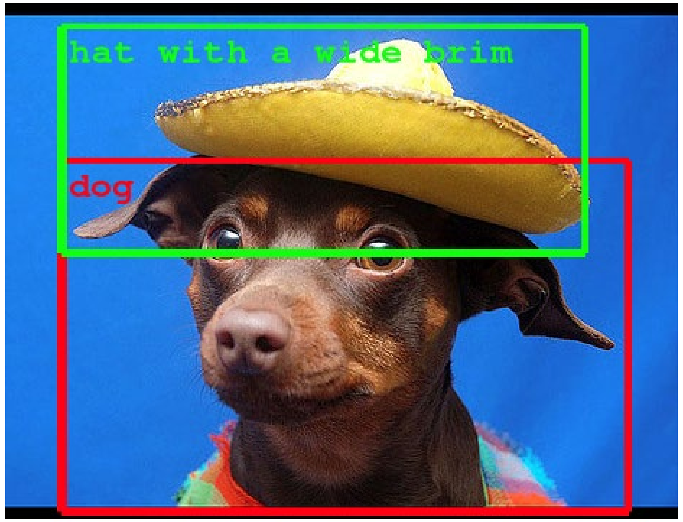


- Suppose we are given a training set of N observations

$$(x_1, \ldots, x_N) \text{ and } (y_1, \ldots, y_N), x_i \in \mathrm{R}^d, y_i \in \{-1, 1\}$$

- Classification problem is to estimate f(x) from this data such that

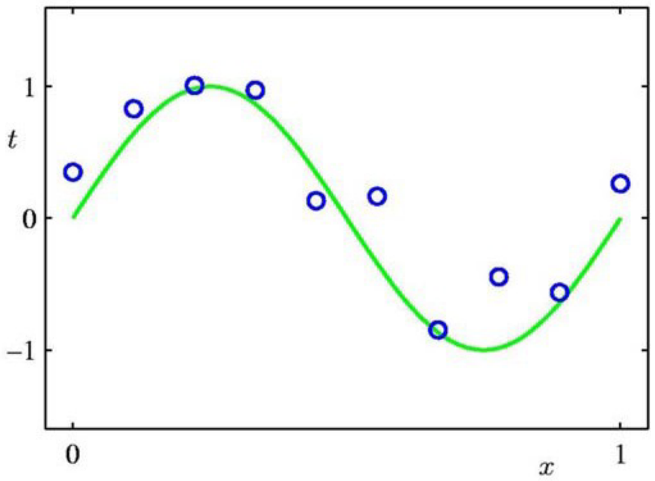
$$f(x_i) = y_i$$



Object recognition

https://ai.googleblog.com/2014/09/building-deeper-understanding-of-images.html

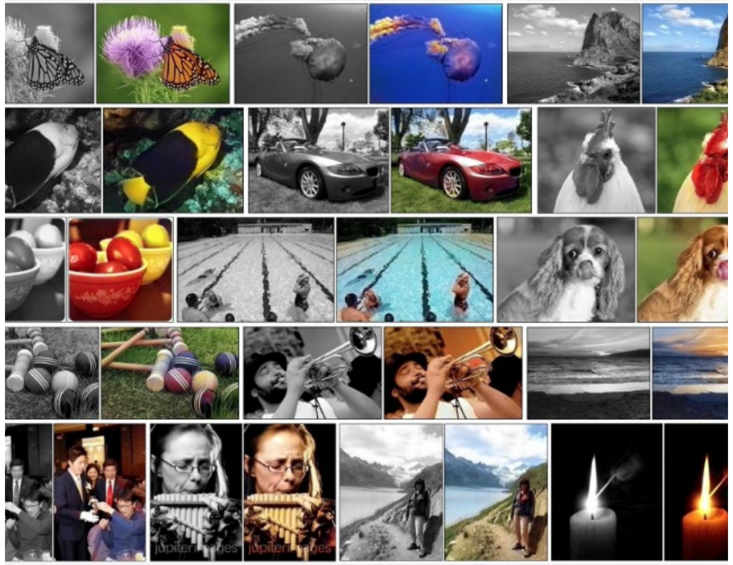
Regression



- Suppose we are given a training set of N observations

$$(x_1, \ldots, x_N) \text{ and } (y_1, \ldots, y_N), x_i, y_i \in \mathbb{R}$$

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



Colorize B&W images automatically

https://tinyclouds.org/colorize/

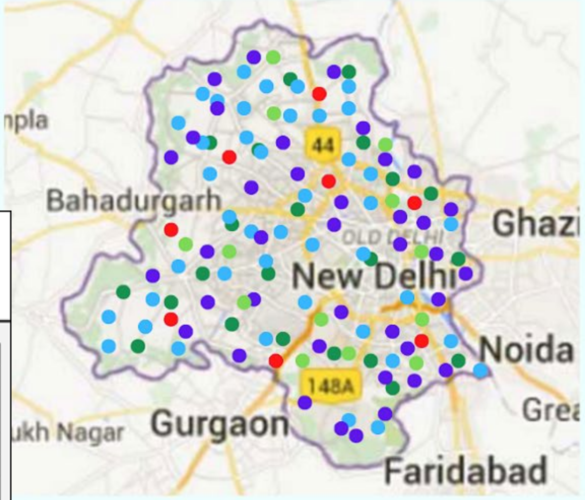
Clustering

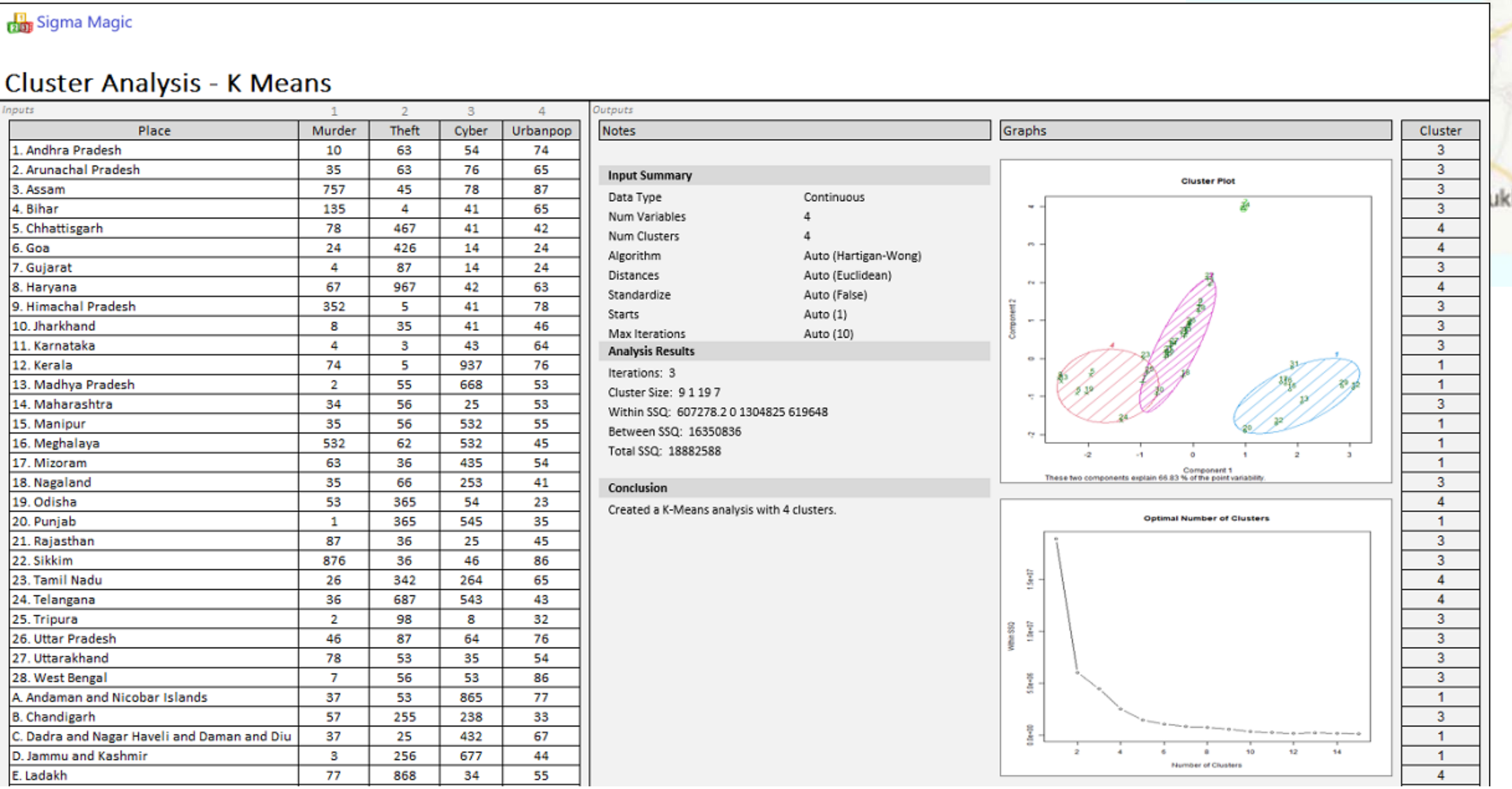
Crime prediction using k-means clustering

https://sigmamagic.com/blogs/crime-analysis-using-k-means-clustering/

CRIME PATTERN ANALYSIS

Get the data set of Criminal activities and determine geospatial points of the crime in an area → Use proper clustring techniques to identify patterns → Analyse patterns and draw conclusions





Cluster Analysis - K Means

Place	Murder	Theft	Cyber	Urbanpop	Cluster
	1	2	3	4	
1. Andhra Pradesh	10	63	54	74	3
2. Arunachal Pradesh	35	63	76	65	3
3. Assam	757	45	78	87	3
4. Bihar	135	4	41	65	3
5. Chhattisgarh	78	467	41	42	4
6. Goa	24	426	14	24	4
7. Gujarat	4	87	14	24	3
8. Haryana	67	967	42	63	4
9. Himachal Pradesh	352	5	41	78	3
10. Jharkhand	8	35	41	46	3
11. Karnataka	4	3	43	64	3
12. Kerala	74	5	937	76	1
13. Madhya Pradesh	2	55	668	53	1
14. Maharashtra	34	56	25	53	3
15. Manipur	35	56	532	55	1
16. Meghalaya	532	62	532	45	1
17. Mizoram	63	36	435	54	1
18. Nagaland	35	66	253	41	3
19. Odisha	53	365	54	23	4
20. Punjab	1	365	545	35	1
21. Rajasthan	87	36	25	45	3
22. Sikkim	876	36	46	86	3
23. Tamil Nadu	26	342	264	65	4
24. Telangana	36	687	543	43	4
25. Tripura	2	98	8	32	3
26. Uttar Pradesh	46	87	64	76	3
27. Uttarakhand	78	53	35	54	3
28. West Bengal	7	56	53	86	3
A. Andaman and Nicobar Islands	37	53	865	77	1
B. Chandigarh	57	255	238	33	3
C. Dadra and Nagar Haveli and Daman and Diu	37	25	432	67	1
D. Jammu and Kashmir	3	256	677	44	1
E. Ladakh	77	868	34	55	4

Notes

Input Summary

Data Type	Continuous
Num Variables	4
Num Clusters	4
Algorithm	Auto (Hartigan-Wong)
Distances	Auto (Euclidean)
Standardize	Auto (False)
Starts	Auto (1)
Max Iterations	Auto (10)

Analysis Results

Iterations: 3

Cluster Size: 9 1 19 7

Within SSQ: 607278.2 0 1304825 619648

Between SSQ: 16350836

Total SSQ: 18882588

Conclusion

Created a K-Means analysis with 4 clusters.

Reinforcement learning

Reinforcement learning for self-driving cars

https://www.thinkautonomous.ai/blog/deep-reinforcement-learning-for-self-driving-cars-an-intro/

ML algorithms

- Regression:
  Ridge regression, LASSO, Support Vector Machines, Decision Trees, Random Forest, Multilayer Neural Networks, Deep Neural Networks, ...

- Classification:
  Naive Bayes, Support Vector Machines, Random Forest, Multilayer Neural Networks, Deep Neural Networks, ...

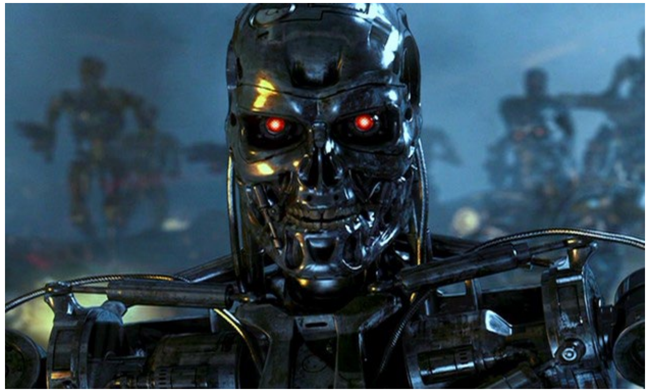
- Clustering:
  k-Means, Hierarchical Clustering, ...

- Reinforcement learning:
  - Q-learning, policy gradient methods, Monte Carlo methods, deep neural networks

Issues

- Many machine learning/AI projects fail (Gartner claims 85 %)

- Ethics, e.g., Amazon boss tells staff AI means their jobs are at risk in coming years(https://www.theguardian.com/technology/2025/jun/18/amazon-boss-tells-staff-ai-means-their-jobs-are-at-risk-in-coming-years); ML makes mistakes, e.g., wrong medicine



Reasons for failure

- Asking the wrong question
- Trying to solve the wrong problem
- Not having enough data
- Not having the right data
- Having too much data
- Hiring the wrong people
- Using the wrong tools
- Not having the right model

Frameworks

- Programming languages
  - Python
  - R
  - C++
  - …
- Many libraries
  - scikit-learn
  - XGBoost
  - NLTK
  - PyTorch
  - TensorFlow
  - Keras
  - …

Fast-evolving ecosystem!

classic machine learning

Natural Language Toolkit

deep learning frameworks

Scikit-learn

- Nice end-to-end framework
  - data exploration (+ Pandas + HoloViews (data analysis & visualization))
  - data preprocessing (+ Pandas)
    - cleaning/missing values
    - normalization
  - training
  - testing
  - application
- "Classic" machine learning only
- https://scikit-learn.org/stable/

PyTorch

- An end-to-end platform for machine learning
  - ML APIs
  - Traditional & deep learning
  - Easy deployment and development using GPU
  - Pre-trained models
  - Open source
- https://pytorch.org/

TensorFlow

- An end-to-end platform for machine learning
  - ML APIs
  - Traditional & deep learning
  - Easy deployment and development using GPU
  - Pre-trained models
  - Mobile, Embedded devices dev.
  - Open source
- https://www.tensorflow.org/

Keras

- High-level framework for deep learning
- TensorFlow backend
- Layer types
  - dense
  - convolutional
  - pooling
  - embedding
  - recurrent
  - activation
  - …
- https://keras.io/

TensorFlow vs PyTorch

- **PyTorch vs TensorFlow**: Both are powerful frameworks with unique strengths; PyTorch is favoured for research and dynamic projects, while TensorFlow excels in large-scale and production environments.
- **Ease of Use**: PyTorch offers a more intuitive, Pythonic approach, ideal for beginners and rapid prototyping. TensorFlow, with its recent updates, is becoming more user-friendly.
- **Performance and Scalability**: TensorFlow is optimized for performance, particularly in large-scale applications. PyTorch provides flexibility and is beneficial for dynamic model adjustments.
- **Community and Resources**: TensorFlow has a broad, established community with extensive resources, whereas PyTorch has a rapidly growing community, especially popular in academic research.
- **Real-World Applications**: PyTorch is prominent in academia and research-focused industries, while TensorFlow is widely used in industry for large-scale applications.
- **Future Prospects**: Both frameworks are evolving, with PyTorch focusing on usability and TensorFlow on scalability and optimization.
- **Making the Right Choice**: Your decision should be based on the project’s needs – PyTorch for flexibility and research, TensorFlow for scalability and production

Summary

- Machine learning: algorithms for AI
- Applying in industry
- Four canonical ML problems
- ML frameworks