OSM Lecture and Exercises

Akito Sakurai

Handouts

- At

http://www.sakurai.comp.ae.keio.ac.jp/classes/2015OSMLandE.html

or

Machine Learning

Akito Sakurai

Contents

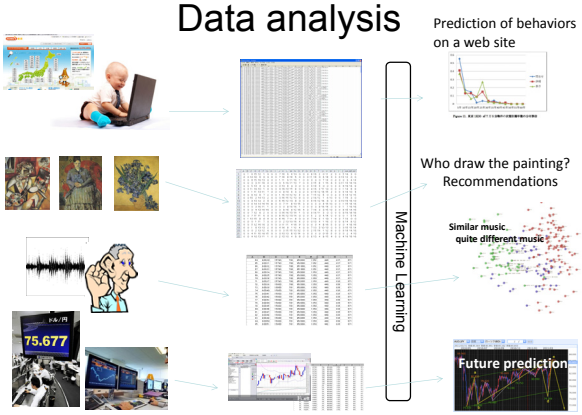
- Very brief introduction (this)
- Bayesian inference and naïve Bayes
  - With an exercise
- Decision tree
  - With an exercise

What is the Machine Learning

- ML is a science/engineering field of
  - For a given set of samples (tiny or huge), predict some attributes of an unseen samples
    - By building a model of the given data, or
    - Directly (without building a model), or

    infer some actions to get benefits.
  - ML is a branch of AI (artificial intelligence) and also is a backbone of many AI research fields

Data analysis



Types of Machine Learning

- Supervised Learning
  - Classification
  - Regression
  - Time series analysis
- Unsupervised Learning
  - Clustering
  - Association rule mining
- Reinforcement Learning

Supervised Learning

- P(x,y): true distribution (not known)
- D: training samples drawn from P(x,y)

	x	y
D	57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,1,1,0,0,0,0,0,0,0	0
	78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0	1
	69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0	0
	18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
	54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,1,0,0,1,0,0,0	1
	84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0	0
	89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,1,0,0	1
	49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0	0
	40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
	74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0	1
	77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,1	0

We want to infer

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0 ?

Numbers are from Ramakrishnan and Gehrke. Database Management Systems, 3rd Edition.

Supervised Learning

Well-defined goal:

Learn Q(x,y) that is a good approximation to P(x,y) from training sample D

Or

Infer y for a given and unseen x, where (x,y) is drawn from P(x,y)

Well-defined error metrics:

Accuracy, RMSE, ROC, …

Un-Supervised Learning

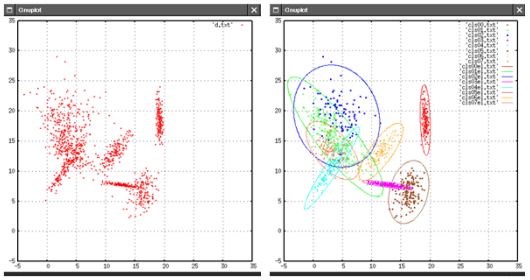
Training dataset:

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,1,1,0,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,1,0,0,1,0,0,0	1
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0	0
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,1,0,0	1
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0	0
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0	1
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,1	0

Unseen data:

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0 ?

Clustering by EM algorithm



http://jormungand.net/projects/em/