(1) Introduction

Note Title 3/26/2008

Chp 1.

Alpaydin.

Data Mining.

ability to store vast amounts of data.

Need to understand regularities in the data. Not perfect understanding

Good and Useful Approximations.

Examples :

Financial - Credit / Fraud / Stock Market.
Manufacturing - optimization / Troubleshoot / Control.
Medicine - Medical Diagnosis.
Telecomunications - Network optimization / service.
Science - Data Physics, Biology.
Web - Search, Analysis.
A.I - Vision, language, robotics.

(2) Machine Learning. in practice

programming computers to optimize a performance criterion using example data or past experience.

A model defined up to some parameters. learning is the execution of a computer program to optimize these parameters using training data / past experience.

The model may be

predictive. to make predictions in future.

or descriptive to gain knowledge from data.

Machine Learning involves

Statistics → build mathematical models with uncertainty. Make inferences from sample

Computer Science → efficient algorithms for optimization problems of learning, storing and process data.

After learning → algorithms for inference, storage.

Mathematics → optimization / geometric formulations.

(3) Examples of Machine Learning Application

Learning Associations

Retail: Basket Analysis

find associations between products bought by customers.

If people who buy X typically also buy Y – then client who buys X is a potential customer for Y.

Want an association rule

Conditional probability $P(Y|X)$.

E.G. $P(\text{chips} \mid \text{beer}) = 0.7$,

then 70% customers who buy beer will also buy chips

More advanced – make distinctions between customers

$P(Y|X, D)$ D – customer attributes

e.g. gender, age, marital status.

Bookseller – products are books or authors

Web Portal – links to webpages, what links will user click, download pages in advance.

(4) Classification

Credit Scoring.

Bank loans money at interest.
What risk is associated with loan?
What probability that customer will fail/default to pay back all/part of the money.

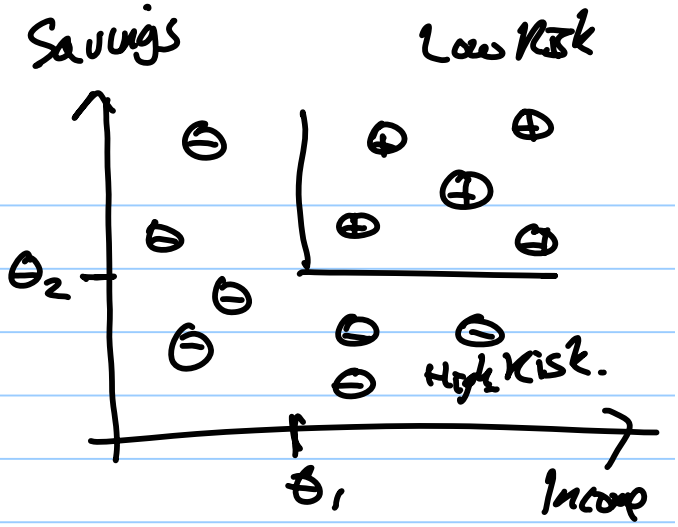
Credit Scoring — bank calculates the risk given the amount of credit and information about customer.. Attributes
Information — income, savings, collateral, profession, age, financial history.
Bank has records of past loans including defaults

Bank wants to infer a general rule coding the association between customers attributes and his risk.

Classification problem — two classes:
(a) low-risk, (b) high-risk.
Information about customer's attributes are input to a classifier whose task is to assign the customer to one of these two classes.

(5) Classification (cont)

⊕ & ⊖ are data instances
"+" classified as low risk
"−" classified as high risk.



Example: two attributes (for simplicity)
savings, income.

Example classification rule:

IF income $> \theta_1$ AND savings $> \theta_2$
THEN low-risk
ELSE high-risk.

This is an example of a discriminant function that separates the data into two classes.

Main application is prediction. If future is similar to the past, then we can make predictions for novel instances

In some cases, instead of 0/1 decision, we may want to calculate prob. $P(Y|X)$
(i.e. learn an association).

(6) Pattern Recognition Examples:

Optical character Recognition (OCR)

recognize character codes from images.

variability in writing styles

exploit redundancy of language - successive characters are not independent but are constrained by the coords of the language.

Medical Diagnosis

inputs are relevant information about the patient and the classes are the illnesses.

Inputs - Age, gender, medical history, symptoms

Some tests may not have been applied and these inputs are missing.

Tests are expensive, take time and we only want to apply them to gain valuable info

Wrong decision is very bad ~ classifier must take this into account.

Speech recognition.

(7) Knowledge Extraction.

learning a rule from data also gives knowledge extraction.

The rule is a simple model that explains the data — looking at this model gives an explanation for the process underlying the data.

This knowledge can be used - e.g. to advertise to low-risk customers for bank loans

Learning also performs compression. Since we get an explanation that is simpler than the data. It requires less memory to store and less computation to process.

(If you know the law of addition, you don't need to remember the sum of all possible pairs of numbers)

Outlier detection - find instances which do not obey the rule and are exceptions.

-> eg. to detect anomalies requiring attention (e.g. fraud).

(8) Regression

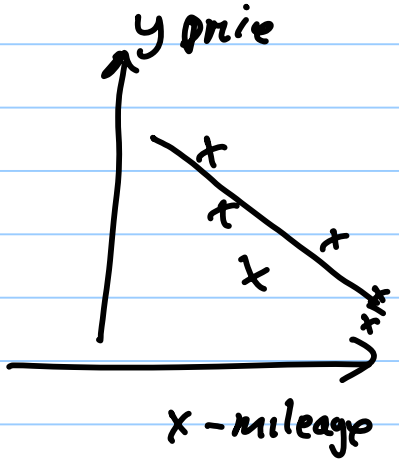
System to predict the price of a used cars

X- Inputs – car attributes – brand, year, engine capacity, mileage, and so on.

Y- Output – price of car.

Regression – predict Y from X using training data.

Fit a function – e.g. $y = wx + w_0$



regression (and classification) are supervised learning problems

Models $y = g(x|\theta)$ θ parameters.

$y \in \{0,1\}$ classification $g(.)$ discriminant function

Machine learning – optimizes for parameters θ to minimize a performance measure.

Other applications (most of statistics?)

navigation by mobile robots, autonomous car.

machine to roast coffee

- inputs ~ temperature, times, coffee beans.
  output ~ consumer satisfaction.

(9) Too often to ICS:
this course will put less emphasis on these.

(A) Unsupervised Learning → no supervisor
aim is to find regularities in the data.

Clustering → find groups / clusters of

e.g. inputs are attributes of customers.
- provides natural grouping of customers (e.g. to give $P(i|x, D)$ - default D)

Image Compression - efficient ways to store images
Image "vocabularies"

Bioinformatics. → DNA - amino acids are series of bases A, G, C, T.
alignment problem - match different sequences
difficult problem - ambiguity
clustering learns motifs which are sequences of amino acids that regularly occur. - correspond to structural/functional elements. - less ambiguous to match.

(10) (B) Reinforcement Learning.

Output of system is a sequence of actions
Want to determine the policy - the sequence of actions to reach the goal.
Game Playing - backgammon - Gerald Tesauro
Robot Navigation.

Overall, the importance of Machine Learning will increase. More and more data is becoming available. Machines are becoming better at analyzing it than humans

History. Artificial Intelligence
traditionally logical - need probabilities & statistics
Neural Networks.
- basis of neural networks lies in statistics. (what do neurons do?)
What results is a merge between C.S. (logic) & Statistics.