INTRODUCTION

Big Data

- Widespread use of personal computers and wireless communication leads to “big data”
- We are both producers and consumers of data
- Data is not random, it has structure, e.g., customer behavior
- We need “big theory” to extract that structure from data for
  - Understanding the process
  - Making predictions for the future

Why Machine Learning

- Machine learning is programming computers to optimize a performance criterion using example data or past experience.
- E.g., there is no need to "learn" to calculate payroll
- Learning is used when:
  - Human expertise does not exist (navigating on Mars)
  - Humans are unable to explain their expertise (speech recognition)
  - Solution changes in time (routing on a computer network)
  - Solution needs to be adapted to particular cases (user biometrics)

Talk About Learning

- Learning general models from a data of particular examples
- **Data is cheap and abundant** (data warehouses, data marts); **knowledge is expensive and scarce**.
- Example in retail: Customer transactions to consumer behavior:
  People who bought "Blink" also bought "Outliers" (www.amazon.com)
- Build a model that is *a good and useful approximation* to the data.

Data Mining

- Retail: Market basket analysis, Customer relationship management (CRM)
- Finance: Credit scoring, fraud detection
- Manufacturing: Control, robotics, troubleshooting
- Medicine: Medical diagnosis
- Telecommunications: Spam filters, intrusion detection
- Bioinformatics: Motifs, alignment
- Web mining: Search engines

…

What is Machine Learning

- Optimize a performance criterion using example data or past experience.
- Role of Statistics: Inference from a sample
- Role of Computer science: Efficient algorithms to solve the optimization problem
- Representing and evaluating the model for inference

Types

- Association
- Supervised Learning
- Classification
- Regression
- Unsupervised Learning
- Reinforcement Learning

Learning Associations

Basket analysis:

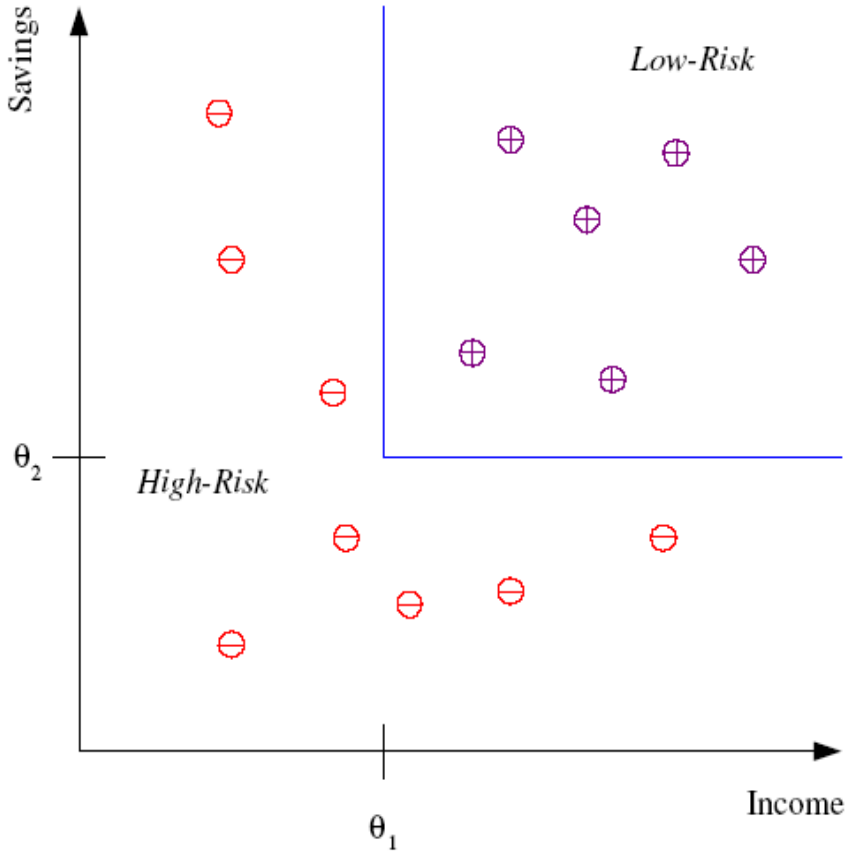
P (Y | X): probability that somebody who buys X also buys Y where X and Y are products/services.

Example: P (chips | beer) = 0.7

Classification

Example: Credit scoring

Differentiating between low-risk and high-risk customers from their income and savings



Discriminant: IF *income* > θ_1 AND *savings* > θ_2

THEN low-risk ELSE high-risk

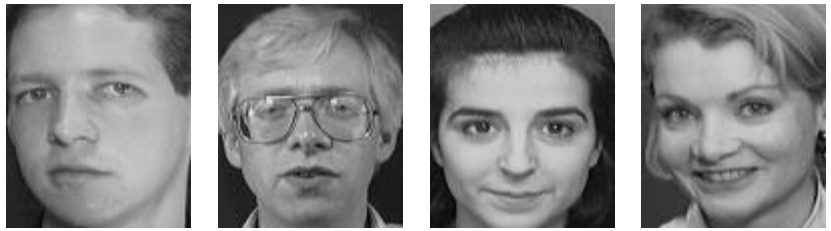
Classification: Applications

- Pattern recognition
- Face recognition: Pose, lighting, occlusion (glasses, beard), make-up, hair style
- Character recognition: Different handwriting styles.
- Speech recognition: Temporal dependency.
- Medical diagnosis: From symptoms to illnesses
- Biometrics: Recognition/authentication using physical and/or behavioral characteristics: Face, iris, signature, etc
- Outlier/novelty detection

Face Recognition

Training examples of a person

Test images



ORL dataset, AT&T Laboratories, Cambridge UK

Regression

Example: Price of a used car

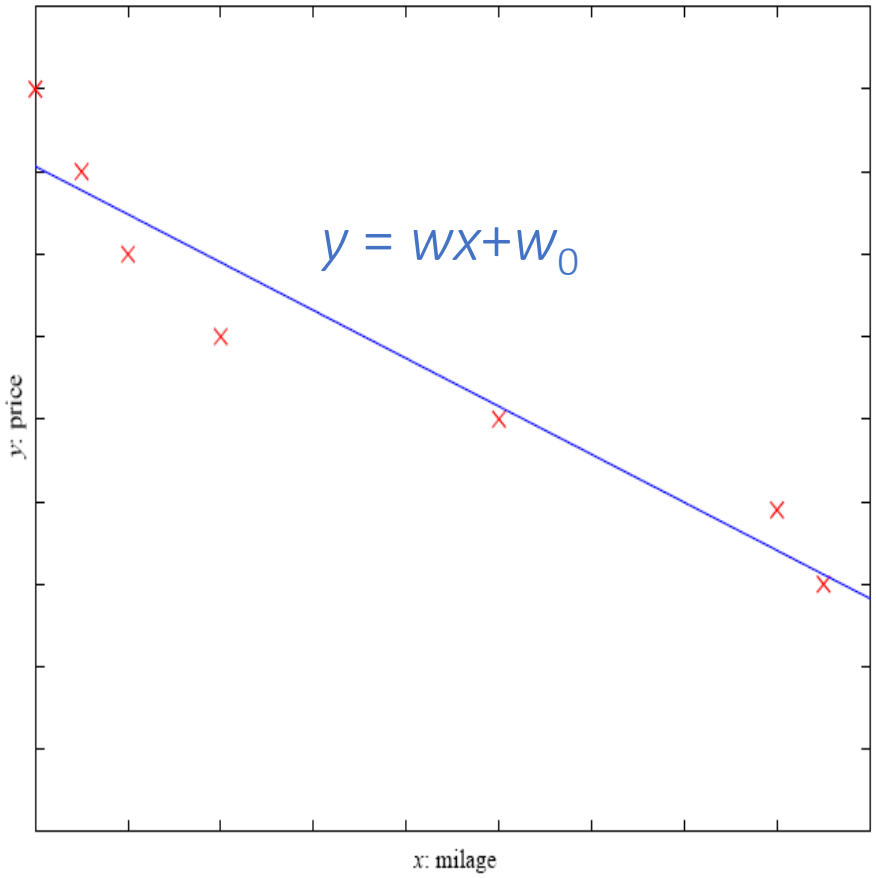
x : car attributes

y : price

$y = g(x \mid q)$

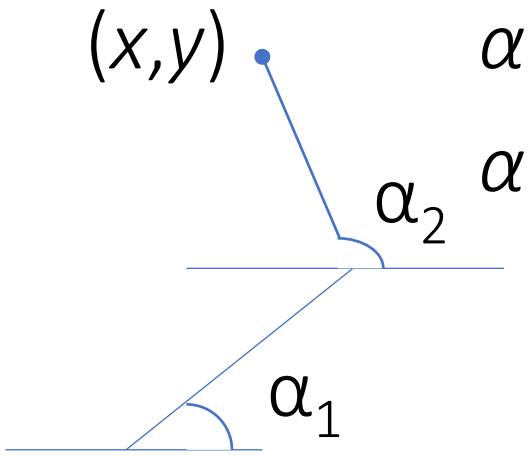
g ( ) model,

q parameters



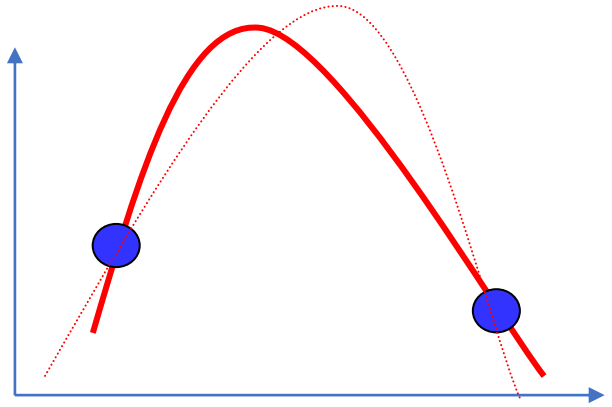
Regression Applications

- Navigating a car: Angle of the steering
- Kinematics of a robot arm



$\alpha_1 = g_1(x,y)$

$\alpha_2 = g_2(x,y)$



Response surface design

Supervised Learning: Uses

- Prediction of future cases: Use the rule to predict the output for future inputs
- Knowledge extraction: The rule is easy to understand
- Compression: The rule is simpler than the data it explains
- Outlier detection: Exceptions that are not covered by the rule, e.g., fraud

Unsupervised Learning

- Learning “what normally happens”
- No output
- Clustering: Grouping similar instances
- Example applications
- Customer segmentation in CRM
- Image compression: Color quantization
- Bioinformatics: Learning motifs

Reinforcement Learning

- Learning a policy: A sequence of outputs
- No supervised output but delayed reward
- Credit assignment problem
- Game playing
- Robot in a maze
- Multiple agents, partial observability, ...

Resources: Datasets

UCI Repository:
http://www.ics.uci.edu/~mlearn/MLRepository.html

Statlib: http://lib.stat.cmu.edu/

DATA

Types

- **Record**
  Relational records, Data matrix, e.g., numerical matrix, crosstabs, Document data: text documents: term-frequency vector, Transaction data
- **Graph and network**
  World Wide Web, Social or information networks, Molecular Structures
- **Ordered**
  Video data: sequence of images, Temporal data: time-series, Sequential Data: transaction sequences, Genetic sequence data
- **Spatial, image and multimedia**:
  Spatial data: maps, Image data, Video data:

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Characteristics

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Distribution
  - Centrality and dispersion

Data Objects

- Data sets are made up of **data objects**.
- A data object represents an **entity**.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called samples , examples, instances, data points, objects, tuples.
- Data objects are described by attributes.
  - Database rows -> data objects
  - Columns ->attributes

Attributes

- **Attribute** (or **dimensions**, **features**, **variables**): a data field, representing a characteristic or feature of a data object.
- E.g., customer_ID, name, address
- Types:
  - Nominal
  - Ordinal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attribute Types

Nominal: categories, states, or "names of things"

Hair_color = {*auburn, black, blond, brown, grey, red, white*}

marital status, occupation, ID numbers, zip codes

Binary

Nominal attribute with only 2 states (0 and 1)

Symmetric binary: both outcomes equally important

e.g., gender

Asymmetric binary: outcomes not equally important.

e.g., medical test (positive vs. negative)

Convention: assign 1 to most important outcome (e.g., HIV positive)

Ordinal

Values have a meaningful order (ranking) but magnitude between successive values is not known.

Size = {*small, medium, large*}, grades, army rankings

Numeric Data

Quantity (integer or real-valued)

Interval

- Measured on a scale of **equal-sized units**
- Values have order
  - E.g., *temperature in C˚ or F˚, calendar dates*
- No true zero-point

Ratio

- Inherent **zero-point**
- We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
  - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs Continuous

Discrete Attribute

- Has only a finite or countably infinite set of values
  - E.g., zip codes, profession, or the set of words in a collection of documents
- Sometimes, represented as integer variables
- Note: Binary attributes are a special case of discrete attributes

Continuous Attribute

- Has real numbers as attribute values
  - E.g., temperature, height, or weight
- Practically, real values can only be measured and represented using a finite number of digits
- Continuous attributes are typically represented as floating-point variables

Statistical Description

Motivation

To better understand the data: central tendency, variation and spread

Data dispersion characteristics

median, max, min, quantiles, outliers, variance, etc.

Numerical dimensions correspond to sorted intervals

Data dispersion: analyzed with multiple granularities of precision

Boxplot or quantile analysis on sorted intervals

Dispersion analysis on computed measures

Folding measures into numerical dimensions

Boxplot or quantile analysis on the transformed cube

Central Tendency

Mean (algebraic measure) (sample vs. population):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

Note: *n* is sample size and *N* is population size.

Weighted arithmetic mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Trimmed mean: chopping extreme values

Median:

Middle value if odd number of values, or average of the middle two values otherwise

Estimated by interpolation (for *grouped data*):

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Mode

Value that occurs most frequently in the data

Unimodal, bimodal, trimodal

Empirical formula:

$$mean - mode = 3 \times (mean - median)$$

Central Tendency

- Median is expensive when datasets are big
- For numeric values we can approximate median with:

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width$$

L_1: the lower boundary of the median interval

N: the number of values

$(\sum freq)_l$ is the sum of frequencies of all of the intervals that are lower than the median interval

$freq_{median}$ is the frequency of the median interval

width is the width of the median interval

Central Tendency

- Example
  - Values: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110
  - Intervals: [10,50], [51,90], [91,130]
  - $L_1 = 51$
  - N = 12
  - $(\sum freq)_l = 4$
  - $freq_{median} = 7$
  - *width* = 40

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width$$

  - median = 51 + ((12/2 - 4) / 7) * 40 = 51 + 2/7 * 40 = 51 + 11.43 = 62.43

  - Theory: median = (52+56) / 2 = 54

Central Tendency

- Example
  - Values: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 71, 72, 85, 86, 88, 90, 110
  - Intervals: [10,70], [71,100], [101,130]
  - $L_1 = 71$
  - N = 18
  - $(\sum freq)_l = 11$
  - $freq_{median} = 6$
  - *width* = 30

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width$$

  - median = 71 + ((18/2 - 11) / 6) * 30 = 71 - 2/6 * 30 = 71 - 10 = 61

  - Theory: median = (63+70) / 2 = 66.5