574M: Introduction to Statistical Machine Learning

Hao Helen Zhang

Course Information

- Course homepage:
  http://www.math.arizona.edu/∼hzhang/math574m.html
- Prerequisite:
  - MATH 464, MATH 466, or equivalent
  - Programming language: R (recommended)
- Textbooks: *The Elements of Statistical Learning* (electronic version available at course website)
- Reference Books:
  1. *Principle and Theory for Data Mining and Machine Learning* by Clark, Forkoue, Zhang (2009)
  2. *Pattern Recognition and Neural Networks* by B. Ripley (1996)
  3. *Learning with Kernels* by Scholkopf and Smola (2000)
  4. *Nature of Statistical Learning Theory* by Vapnik (1998)

What is Data Mining

- the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data
- the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions
- a set of methods used in the knowledge discovery process
- the process of discovering advantageous patterns in data
- a decision support process where we look in large databases for unknown and unexpected patterns of information
- ......

DM is a process of discovering patterns and relationships in data, with an emphasis on large observational databases.

Emerging Discipline, Why Now?

Driving Forces

- explosive growth of data in a great variety of fields
  - revolution in biotechniques (microarray, GWAS, next generation sequencing)
  - internet, network, search engines, digital images, multi-media information
- Rapidly increasing computer power
  - cheaper storage devices with higher capacity
- faster communications; better database management systems

What is Big Data

Wikipedia says,

> "Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. "

NSF Big Data Initiative (2012),

> "from a scientific perspective, that scientists manage, analyze, visualize, and extract useful information from large, diverse, distributed and heterogeneous data sets so as to accelerate the progress of scientific discovery and innovation"

Wall Street Journal (04-20-2012),

> "from a business perspective, an enterprise mine all the data it collects right across its operations to unlock golden nuggets of business intelligence"

Big Data References

- *The Wall Street Journal*, Article 11-29-2012, "Big Data is on the rise, bringing big questions"
- *The Wall Street Journal*, Article 04-29-2012, "Big Data's big problem: little talent"
- *McKinsey & Company*, Report 05-2011, "Big Data: The next frontier for innovation, competition, and productivity."

Google search "Big Data" gives 0.8 billion results.

Sizes of Big Data

The sizes of modern datasets are increasing faster than ever:

Bytes=8 bits, kilobyte= 10^3 bytes, Megabyte= 10^6 bytes,
Gigabyte= 10^9 bytes, Terabyte= 10^{12} bytes,
Petabyte= 10^{15} bytes, Exabyte= 10^{18} bytes,
Zettabyte= 10^{21} bytes, Yottabyte = 10^{24} bytes, ...

- 2 Megabytes: A high resolution photograph
- 50 Terabytes: The contents of a large MASS Storage System
- 2 Petabytes: All US academic research libraries
- 5 Exabytes: All words ever spoken by human beings

Big Data Features and Challenges

Big data are featured with

- massive volume, ultra-high dimension
- complex structure, heterogeneous
- multiple-source, multiple-type data

Big Data Challenges:

- data storage, transfer [enormous memory, parallel processing].
- data search, retrieval [at a high speed]
- data cleaning, organization, visualization [present/display data in a meaningful way]
- data analysis [gain deep insight about data]

Examples

- Wal-Mart made $>$ 20 million transactions daily, and constructed an 11 terabyte database of customer transactions
- AT&T had 100 million customers and carried on the order of 300 million calls a day on its long distance network
- molecular data: DNA copy-number alteration, mRNA expression, protein expression
- images, network data, tweet data, ...

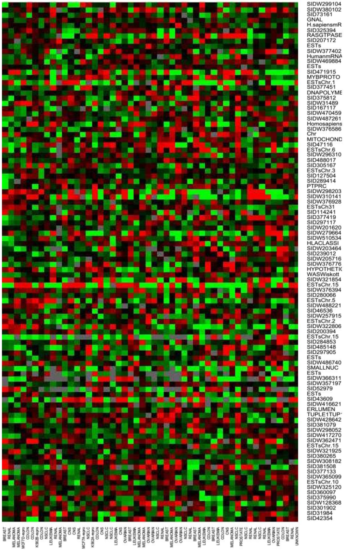
Role of Data Scientists

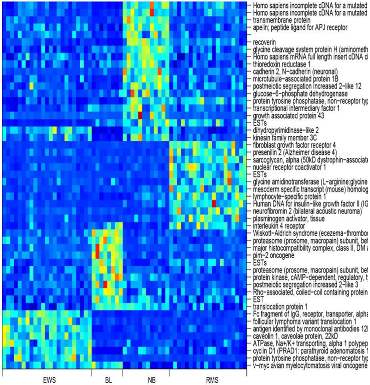
Data Science is an interdisciplinary field:

- statistics machine Learning, computer science, mathematics, pattern recognition, signal Processing

Goal: to extract useful informationn from flood of data, to find hidden patterns in data

- feasible and fast computation. Example: Hadoop system is software for **distributed** storage and processing of very large data sets on computer clusters.
- develop new tools and algorithms for data analysis
- provide theoretical foundations for learning algorithms
- help researchers gain deeper understanding of cancer


SIDW299104
SIDW380102
SID73161
GNAL
H.sapiensmR
SID325394
RASGTPASE
SID207172
ESTs
SIDW377402
HumanmRNA
SIDW469884
ESTs
SID471915
MYBPROTO
ESTsChr.1
SID377451
DNAPOLYME
SID375812
SIDW31489
SID167117
SIDW470459
SIDW487261
Homosapiens
SIDW376586
Chr
MITOCHOND
SID47116
ESTsChr.6
SIDW296310
SID488017
SID305167
ESTsChr.3
SID127504
SID289414
PTPRC
SIDW298203
SIDW310141
SIDW376928
ESTsCh31
SID114241
SID377419
SID297117
SIDW201620
SIDW279664
SIDW510534
HLACLASSI
SIDW203464
SID239012
SIDW205716
SIDW376776
HYPOTHETI(
WASWiskott
SIDW321854
ESTsChr.15
SIDW376394
SID280066
ESTsChr.5
SIDW488221
SID46536
SIDW257915
ESTsChr.2
SIDW322806
SID200394
ESTsChr.15
SID284853
SID485148
SID297905
ESTs
SIDW486740
SMALLNUC
ESTs
SIDW366311
SIDW357197
SID52979
ESTs
SID43609
SIDW416621
ERLUMEN
TUPLE1TUP
SIDW428642
SID381079
SIDW298052
SIDW417270
SIDW362471
ESTsChr.15
SIDW321925
SID380265
SIDW308182
SID381508
SID377133
SIDW365099
ESTsChr.10
SIDW325120
SID360097
SID375990
SIDW128368
SID301902
SID31984
SID42354
BREAST
RENAL
MELANOMA
MELANOMA
MCF7D-repro
COLON
COLON
K562B-repro
COLON
NSCLC
LEUKEMIA
RENAL
MELANOMA
BREAST
CNS
CNS
RENAL
MCF7A-repro
NSCLC
K562A-repro
COLON
CNS
NSCLC
NSCLC
LEUKEMIA
CNS
OVARIAN
BREAST
LEUKEMIA
MELANOMA
MELANOMA
OVARIAN
OVARIAN
NSCLC
RENAL
BREAST
MELANOMA
OVARIAN
OVARIAN
NSCLC
RENAL
BREAST
MELANOMA
LEUKEMIA
COLON
BREAST
LEUKEMIA
COLON
CNS
MELANOMA
NSCLC
PROSTATE
NSCLC
RENAL
RENAL
NSCLC
RENAL
LEUKEMIA
OVARIAN
PROSTATE
COLON
BREAST
RENAL
UNKNOWN


Homo sapiens incomplete cDNA for a mutated
Homo sapiens incomplete cDNA for a mutated
transmembrane protein
apelin; peptide ligand for APJ receptor
recoverin
glycine cleavage system protein H (aminometh
Homo sapiens mRNA full length insert cDNA cl
thioredoxin reductase 1
cadherin 2, N-cadherin (neuronal)
microtubule-associated protein 1B
postmeiotic segregation increased 2-like 12
glucose-6-phosphate dehydrogenase
protein tyrosine phosphatase, non-receptor typ
transcriptional intermediary factor 1
growth associated protein 43
ESTs
dihydropyrimidinase-like 2
kinesin family member 3C
fibroblast growth factor receptor 4
presenilin 2 (Alzheimer disease 4)
sarcoglycan, alpha (50kD dystrophin-associate
nuclear receptor coactivator 1
ESTs
glycine amidinotransferase (L-arginine:glycine
mesoderm specific transcript (mouse) homolog
lymphocyte-specific protein 1
Human DNA for insulin-like growth factor II (IG
neurofibromin 2 (bilateral acoustic neuroma)
plasminogen activator, tissue
interleukin 4 receptor
Wiskott-Aldrich syndrome (ecezema-thrombo
proteasome (prosome, macropain) subunit, be
major histocompatibility complex, class II, DM
pim-2 oncogene
ESTs
proteasome (prosome, macropain) subunit, be
protein kinase, cAMP-dependent, regulatory, t
postmeiotic segregation increased 2-like 3
Rho-associated, coiled-coil containing protein
EST
translocation protein 1
Fc fragment of IgG, receptor, transporter, alpha
follicular lymphoma variant translocation 1
antigen identified by monoclonal antibodies 12
caveolin 1, caveolae protein, 22kD
ATPase, Na+/K+ transporting, alpha 1 polypep
cyclin D1 (PRAD1: parathyroid adenomatosis 1
protein tyrosine phosphatase, non-receptor typ
v-myc avian myelocytomatosis viral oncogene
EWS
BL
NB
RMS

Applications (I)

- Business
  - Walmart data warehouse, credit card companies, bank data, stock data
- Marketing
  - Given data on age, income, etc., predict the spending capacity of each customer;
  - Discover the relationship of customers' spending behaviors;
  - recommend products (example: diaper-beer)
- Genomics
  - Human genome projects: DNA sequences; microarray data, SNP, next generation sequence

Applications (II)

- Information Retrieval
  - search engine, text mining, document search.
  - For example, given some key words in a document, determine its topic and content. (*many words in a document, and many, many documents available on the web*).
  - speech recognition, image analysis, multimedia information
- Healthfare
  - personalized medicine
  - cancer classification and treatment, given gene expressions and clinical measurements.

What is Role of Statistics

Statistics machine learning plays a central role in data mining.

- provide theoretical foundations for learning algorithms
- give useful tools to analyze an algorithm's statistical properties and performance guarantee
- help researchers gain deeper understanding of the approaches, design better algorithms, and select appropriate methods for a given problem

Examples of Learning Problems

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack, based on demographic, diet and clinical measurements (Regression and Classification)
- Predict the price of a stock in 6 months, based on company performance measures and economic data (Regression)
- Identify a handwritten ZIP code, from a digitized image (Classification)
- Identify risk factors for prostate cancer, based on clinical, genetic and demographic variables (Feature Selection)
- Identify groups of genes with similar functions from DNA microarray data (High Dimensional Clustering)
- In fraud detection, what covariates are useful in building a model to predict the probability of being a fraud order? how to handle covariate correlations? (Classification Problem)

Goals of This Course

- understand different types of data mining problems
  - supervised, unsupervised, semi-supervised learning
  - regression, classification, clustering, graphical models
  - text mining, multimedia mining, image recognition, social networks, recommendation system, network models
- learn basic statistical concepts and principles (which are favored over a black-box learning algorithm)
  - statistical inference on uncertainty, distribution, loss, risk
  - model building, evaluation, selection, prediction, and optimality; bias-variance tradefoff
  - parameter tuning, training error, test error, cross-validation,
- learn statistical and machine learning methods for big data
  - SVM, PCA, lasso, boosting, tree, random forest
- learn R software packages to analyze data
  - take into serious consideration *scalable, parallel* algorithms

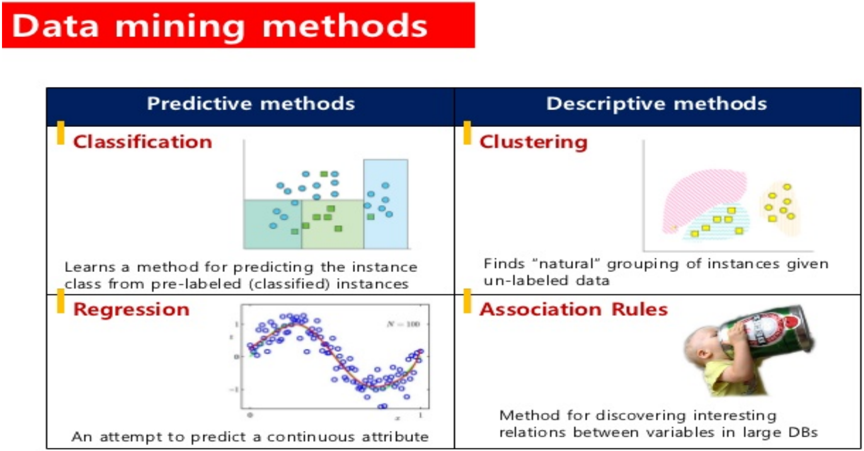
What is “Special” about this course

To combine the “art” of designing good learning algorithms and the “science” of analyzing statistical properties and performance of the approaches

- emphasize “statistical” principles behind the approaches and algorithms
- learn how to formulate a learning problem in a “statistical” framework
- understand existing techniques from a “statistical” perspective; what are the limitations and strengths? Can we do better by relaxing assumptions?

Terminologies

	Statistics	Data Mining
$\mathbf{X}$	predictor, covariate	input, feature
Y	response, outcome	output
$\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$	data points, samples	examples, training data
$\hat{f}(\mathbf{X})$	model fitting	machine learning
data analysis goals	consistency, inference confidence interval convergence rate	prediction, speed risk bound learning theory
Practical vs Theoretical Concerns	Reluctant to use methods methods without theoretical justification (even if the justification is actually meaningless)	Willing to use *ad hoc* methods if they seems to work well (though appearances may be misleading)
Computing	more and more important	heavy



Data mining methods

Predictive methods	Descriptive methods
Classification Learns a method for predicting the instance class from pre-labeled (classified) instances	**Clustering** Finds "natural" grouping of instances given un-labeled data
Regression An attempt to predict a continuous attribute	**Association Rules** Method for discovering interesting relations between variables in large DBs

Different Learning Problems

Typically, we collect data $(\mathbf{X}_i, Y_i), i = 1, \cdots, n$.

- Y_i is the *outcome* or *response* variable.
- $\mathbf{X}_i$ is the *input* or *prediction* variables.

Various Machine Learning Problems

- Supervised learning (Y observed)
- Unsupervised learning (Y unobserved)
- Semi-supervised learning (Y partially available)

Various Statistical problems

- Regression (Y quantitative, a numerical quantity)
- Classification (Y qualitative; a class label)
- Density estimation (no Y)

Supervised Learning vs Unsupervised Learning

	Supervised Learning	Unsupervised Learning
Response Y	observed	unobserved
Major Goal	predict Y, given the observed input $\mathbf{X}$	find interesting patterns in data
Examples	linear regression nonparametric regression classification	clustering density estimation dimension reduction

Example 1: Email or Spam (Textbook Page 2)

- Training data: 4, 601 email messages with known email type.
  - Input **X**: the relative frequencies of 57 of the most commonly occurring words and punctuation marks in the email message.
  - Outcome Y: -1 =email , $+$ =spam
- Goal: Design an automatic spam detector to predict whether a message is *email* or *spam*. In the future, the detector can be used to filter out the junk emails before clogging the users' mailbox.

Supervised learning, binary classification, $n > p$. (data matrix $n \times p$ "long and narrow")

Table: Average percentage of words (the largest difference between spam and email

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

Possible classification rules:

- if (george < 0.6) & (you > 1.5), then spam; otherwise email.
- if (0.2 you -0.3 george)$>$ 0, then spam; otherwise email.

Two types of decision errors:

(1) false positive: classify email to spam (filter out email)

(2) false negative: classify spam to email (email box jammed)

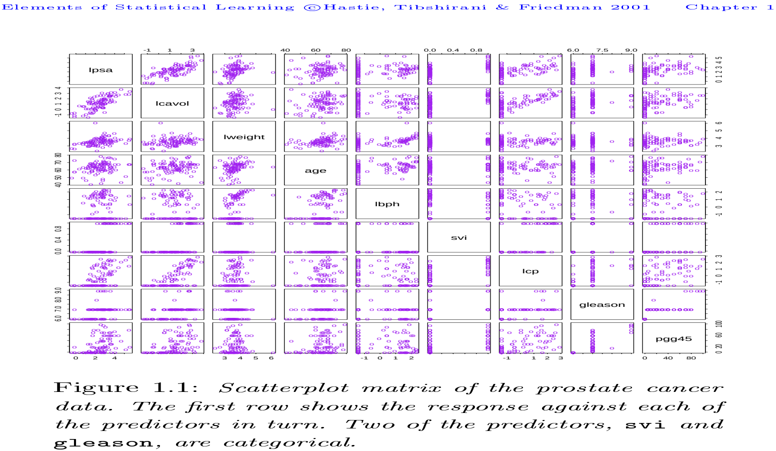
Example 2: Prostate Cancer

Read Textbook page 3

- Training data: 97 male patient with different stages of prostate cancer.
- Input **X**: eight clinical measures: log-cancer volume (*lcavol*), log prostate weight (*lweight*), age, and the other five.
- Outcome Y: the log of the level of prostate specific antigen (*lpsa*)

Questions of interest:

- What is the relationship between *lpsa* and clinical measures?
- Is the linear model sufficient? Nonlinear effect? Interactions?
- Which clinical measures are more relevant to the prediction?



Elements of Statistical Learning ©Hastie, Tibshirani & Friedman 2001 Chapter 1

Figure 1.1: *Scatterplot matrix of the prostate cancer data. The first row shows the response against each of the predictors in turn. Two of the predictors,* **svi** *and* **gleason**, *are categorical.*

Elements of Statistical Learning ©Hastie, Tibshirani & Friedman 2001 Chapter 3

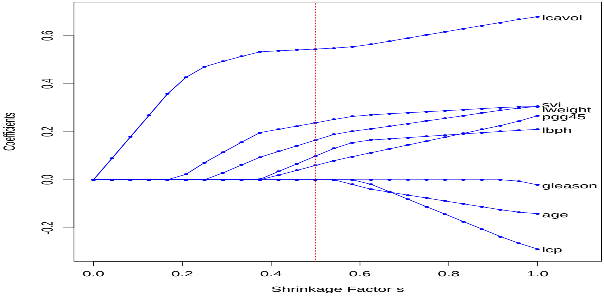


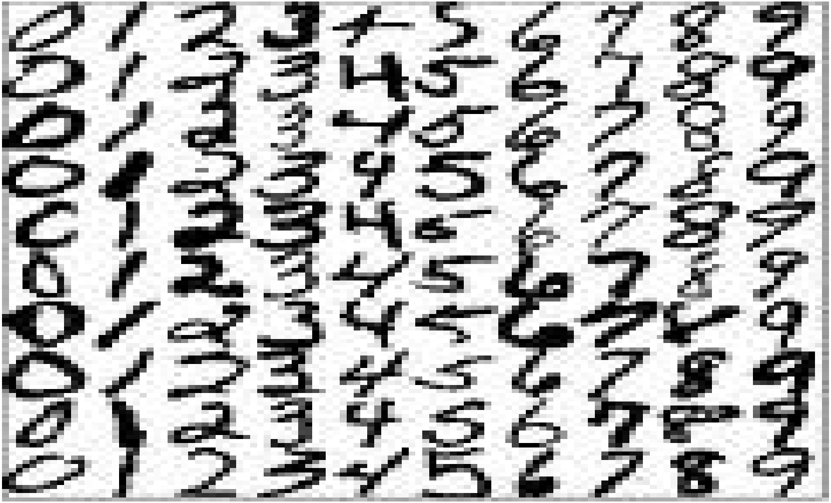
Figure 3.9: *Profiles of lasso coefficients, as tuning parameter t is varied. Coefficients are plotted versus* $s = t/\sum_1^p |\hat{\beta}_j|$*. A vertical line is drawn at* $s = 0.5$*, the value chosen by cross-validation. Compare Figure 3.7 on page 7; the lasso profiles hit zero, while those for ridge do not.*

Handwritten Digit Recognition

Handwritten Digit Recognition (Textbook page 4)

- Goal: identify single digits $0 \sim 9$ based on the images.
- Raw Data: images that are scaled segments from five digit ZIP codes.
    - Each digit has an image of 16×16 eight-bit gray scale maps
    - Pixel intensities range from 0 (black) to 255 (white)
    - Images are normalized to have approximately the same size and orientation
- Input $\mathbf{X}$ is a 16×16 matrix, or a 256 dimensional vector.
- Output $G \in \mathcal{G} = \{0, 1, ..., 9\}$.

The error rate should be very low to avoid misdirection of mail. Some objects are assigned to “do not know” category and sorted by hand.



Example 4: DNA Expression Microarray

DNA = Deoxy-ribo-nucleic acid, a basic material making up human chromosomes.

DNA Microarray Technique: for each sample from a tissue, the expression level (the amount of mRNA) of thousands of genes are measured.

- Training data: $p = 6,830$ genes (rows), $n = 64$ samples (columns) (cancer tumors) taken from two classes.
  - Input $\mathbf{X}$: the level of expression for each gene
  - Goal: discover the relationship between gene and cancer type, or find the *gene signature* of each cancer subtype

Challenge: $p >> n$ (data matrix $n \times p$ “short and fat”’)

How DNA Microarray Technique Works

A breakthrough technology in biology, facilitating the quantitative study of thousands of genes simultaneously from a single sample of cells

1. The nucleotide sequences for a few thousand genes are printed on a glass slide;
2. A target sample and a reference sample are labeled with red and green dyes, and each are hybridized with the DNA on the slide;
3. Through fluroscopy, the log (red/green) intensities of RNA hybridizing each site is measured;
4. The results are a few thousand numbers, typically ranging from -6 to 6, measuring the expression level of each gene in the target relative to the reference sample (positive values indicate higher expression in the target versus the reference, and vice versa for negative values).

Elements of Statistical Learning ©Hastie, Tibshirani & Friedman 2001 Chapte

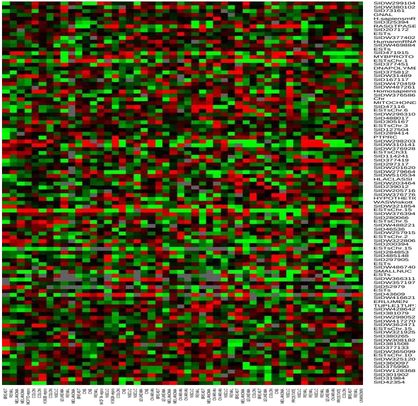


Figure 1.3: *DNA microarray data: expression matrix of 6830 genes (rows) and 64 samples (columns), for the human tumor data. Only a random sample of 100 rows are shown. The display is a heat map, ranging from bright green (negative, under expressed) to bright red (positive, over expressed). Missing values are gray. The rows and columns are displayed in a randomly chosen order.*

Microarray Data Analysis

Typical questions of interest:

- Which samples are most similar to each other, in terms of their expression profiles across genes?
- Which genes are most similar to each other, in terms of their expression profiles across sample?
- Do certain genes show very high (or low) expression for certain cancer samples?
- ....

This task could be viewed as

- a *supervised* or an *unsupervised* problem
- a *regression* or a *classification* problem

Statistical Problems in Data Mining

- Regression (linear, nonlinear, logistic regression, GLM, graphical model)
- Classification (Discriminant Analysis; LDA, QDA, SVM, trees, random forest)
- Feature Selection (Variable Selection; Sparse Analysis; LASSO, screening)
- Dimension Reduction (PCA; ICA; SDR)
- Regularization (control model complexity, aim for parsimonious and interpretable solutions)
- Clustering Analysis (k-means, association role)

Other Variations

- Semi-supervised Learning
- Mixture of the above

New Challenges to Statisticians

- **Data Complexity**: involves many variables which are often related in complex (*nonlinear*) ways.
- **Feature Selection**: many features are available but some are redundant, leading to the *feature selection* or *dimension reduction* problem.
- **Optimization**: many methods involve finding the “best” parameters values by solving complex and large (containing many parameters) optimization problems. Therefore, efficient optimization techniques are required.
- **Visualization**: much harder in the high dimensional space.

This is the so-called *curse of dimensionality*.