MA333 Introduction to Big Data Analysis
Course Introduction

Zhen Zhang

Southern University of Science and Technology

Outlines

Course Info

- Semester 2024-2025 Spring
- Instructor : ZHANG, Zhen （张振）
- Office : Room M5014, School of Science
- Phone : 88018753
- Email : zhangz@sustech.edu.cn
- Office hours : Tuesday morning, 10 :00-12 :00 ; or send email to make an appointment for other time.
- Lecture : 3 credits, 3 hours per week.
- Prerequisite : Calculus (or Mathematical Analysis) ; Linear Algebra ; Probability Theory or Probability Theory and Mathematical Statistics (or other similar courses).

Grading Policy

- Homework : Approximately 6 homework assignments (including programming assignments and written problems). The written homework could be handed in after class.
- In-class quizzes : typically once every two weeks, test how well you learned about the basic concepts, including fill-in-the-blank, single and multiple choices, and simple Q & A
- Programming projects : include coding, data analytics, and reports
- One closed-book final exam
- Grading policy : assignments (30%), quizzes (15%), programming projects (20%), and the final exam (35%).

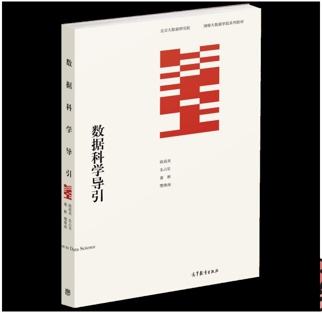
Contents

- Intended for undergraduate students who are interested in pursuing industrial work and research in big data science.
- Concise and self-contained introduction to mathematical aspect of big data science, including **theoretical analysis, algorithms and programming with python**
- Major topics :
  - Introduction to python programming and data preprocessing
  - Three fundamental problems : classification, regression, clustering
  - Model selection, dimensionality reduction
  - Hot topics : text analysis, social network analysis, neural network and deep learning, and recommender systems if time permits

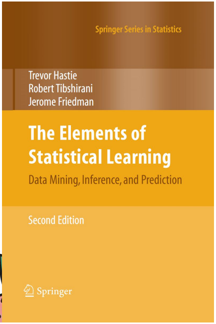
References

- 数据科学导引，欧高炎等著，高等教育出版社，2017.
- 机器学习，周志华著，清华大学出版社，2016.
- An Introduction to Statistical Learning with Applications in Python, by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani and Jonathan Taylor, Springer, 2023. https ://www.statlearning.com/
- Pattern Recognition and Machine Learning, by Christopher M. Bishop, Springer, 2006.
- The Elements of Statistical Machine Learning : Data mining, Inference and Prediction, 2nd Edition, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman, Springer, 2009. https ://hastie.su.domains/ElemStatLearn/
- Foundations of Machine Learning, by Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, MIT Press, 2018.
- Understanding Machine Learning, by Shai Shalev-Shwartz and Shai Ben-David, Cambridge University Press, 2018.
- Deep learning, by Ian Goodfellow, Yoshua Bengio, and Aaron Courville, MIT Press, 2016. https ://www.deeplearningbook.org/

References


北京大数据研究院
博雅大数据学院系列教材
数据科学导引
欧高炎
朱占星
董 彬
鄂维南
高等教育出版社


周志华 著
MACHINE LEARNING
机器学习
清华大学出版社


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

Outlines

Some Examples of Data

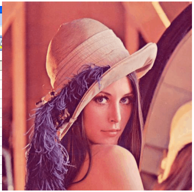
Can you give some examples of data ?

实验数据分析表

项目	加药量(ppm)			CODcr (mg/L)
	硫酸亚铁	双氧水	复合碱	
原水	/	/	/	321
1	2000	1000	2800	186
2	3000	1500	3000	157
3	4000	2000	3100	110
4	5000	2500	3500	78

实验数据仅对该批水样负责

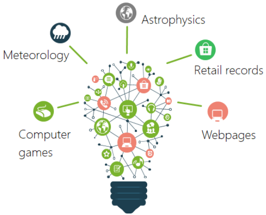
	A	B	C	D	E	F	G	H
1	姓名	语文	数学	英语	其他	平均分	总成绩	备注
2	张前伟	68	91	74	95			
3	秦士友	59	83	81	82			
4	秦洪涛	91	67	85	76			
5	秦会云	59	85	96	85			
6	任恒侠	68	74	71	95			
7	秦兴福	46	91	61	68			
8	韦春磊	89	82	53	95			
9	秦康宁	80	73	85	75			
10	王艳超	86	80	81	85			
11	秦龙霄	79	90	64	91			
12	王芹芹	59	72	62	82			
13	任梦晴	61	76	52	95			



Table, 1D signal (audio, stock price), 2D signal (image), 3D signal (video), etc.

Big Data : 5 Big "V"

- Volume : KB, MB, GB (10^9 bytes), TB, PB, EB (10^{18} bytes), ZB, YB, exponential growth (about 120%/year)
- Variety : different sources from business to industry, different types
- Veracity : Noisy data with errors and inconsistency, redundant information contained in the data, need to retrieve useful information
- Velocity : fast speed for data generation and information transfer, need for realtime processing
- Value : business values for product recommendations and trading ; social values for precision medicine, public health, traffic control, etc.



What is data science

- Retrieve information from data with the help of computational power
- Transfer the information into knowledge
- Two perspectives of data sciences :
  - Study science with the help of data : bioinformatics, astrophysics, geosciences, etc.
  - Use scientific methods to exploit data : statistics, machine learning, data mining, pattern recognition, data base, etc.

Study Science with the Help of Data

A pioneering work of data science : Kepler's Laws

开普勒：分析数据产生价值

行星	周期（年）	平均距离	周期^2/距离^3
水星	0.241	0.39	0.98
金星	0.615	0.72	1.01
地球	1.00	1.00	1.00
火星	1.88	1.52	1.01
木星	11.8	5.20	0.99
土星	29.5	9.54	1.00
天王星	84.0	19.18	1.00
海王星	165	30.06	1.00

Scientific Study of Data

- Grabbing data : business and industrial problem, professional areas
- Storing data : engineering problem, computer science, electronic engineering
- Analyzing data (**key problem**) : scientific problem, mathematics, statistics, computer science

Data Analysis

- Ordinary data types :
  - Table : classical data (could be treated as matrix)
  - Set of points : mathematical description
  - Time series : text, audio, stock prices, DNA sequences, etc.
  - Image : 2D signal (or matrix equivalently, e.g., pixels), MRI, CT, supersonic imaging
  - video : 2D in space and 1D in time (another kind of time series)
  - Webpage and newspaper : time series with spatial structure
  - Network : relational data, graph (nodes and edges)
- Basic assumption : the data are generated from an underlying model, which is unknown in practice
  - Set of points : probability distribution
  - Time series : stochastic processes, e.g., Hidden Markov Model (HMM)
  - Image : random fields, e.g., Gibbs random fields
  - Network : graphical models, Beyesian models

Difficulties

- Huge volume of data
- Extremely high dimensions
  - Curse of dimensionality : the model complexity and computational complexity become exponentially increasing with the growth of dimension
  - Solutions :
    - Make use of prior information
    - Restrict to simple models
    - Make use of special structures, e.g., sparsity, low rank, smoothness
    - Dimensionality reduction, e.g., PCA, LDA, etc.
- Complex variety of data
- Large noise : data are always contaminated with noises

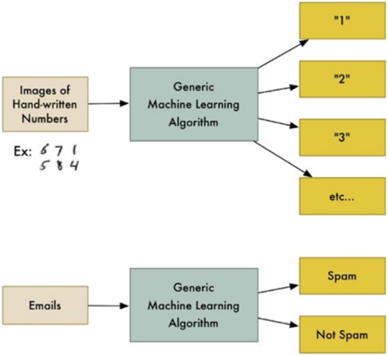
Solution - Algorithms

- Algorithms are in the interdisciplinary part of computer science and mathematics : establish mathematical models, solve it numerically, implement it in the computer languages
- Reduce the algorithmic complexity, with the help of techniques from mathematics or computer science
- Distributional and parallel computing : divide-and-conquer, e.g., MapReduce, GPU
- IEEE 2006 top 10 algorithms in data mining : C4.5, K-Means, SVM, Apriori, EM, PageRank, NaiveBayes, K-Nearest Neighbors, AdaBoost, CART

Outlines

Definition

- Artificial Intelligence (AI) : learning from experiences (data), and improve the computer program adaptively
- Mathematics : Learning the underlying model from data, and generalize the model to adapt new data



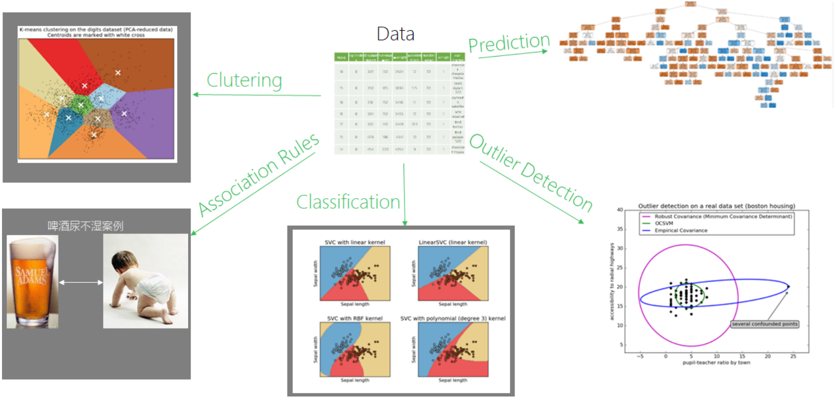
We define *machine learning* as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!).

— 《Machine Learning: A probabilistic perspective》

Related Areas

- Control theory : optimize the cost with optimal control parameters
- Information theory : entropy, optimal coding with best information
- Psycology : reference for machine learning algorithms
- Neuroscience : artificial neural network
- Biology : genetic algorithms
- Theory of Computing : study the computational complexity
- Statistics : large-sample limiting behavior, statistical learning theory
- Artificial Intelligence : symbolic computing
- Bayesian theory : conditionally probabilistic network

Applications

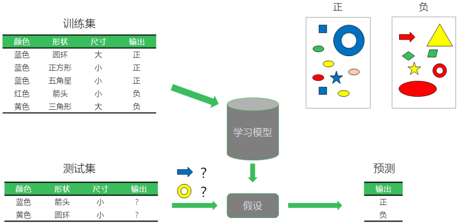


Supervised and Unsupervised Learning

- Supervised learning : classification, regression
- Unsupervised learning : density estimation, clustering, dimensionality reduction
- Semi-supervised learning : with missing data, e.g., EM ; self-supervised learning, learn the missing part of images, inpainting
- Reinforcement learning : play games, e.g., Go, StarCraft ; robotics ; auto-steering

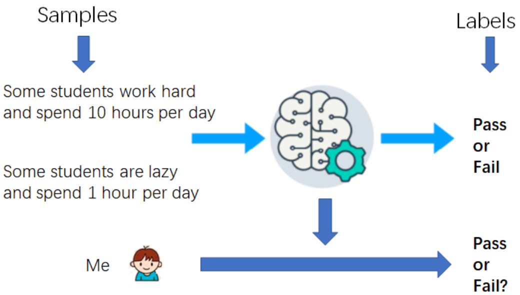
Supervised Learning

- Given labels of data : the labels could be symbols (spam or non-spam), integers (0 or 1), real numbers, etc.
- Training : find the optimal parameters (or model) to minimize the error between the prediction and target
- Classification : SVM, KNN, Desicion tree, etc.
- Regression : linear regression, CART, etc.



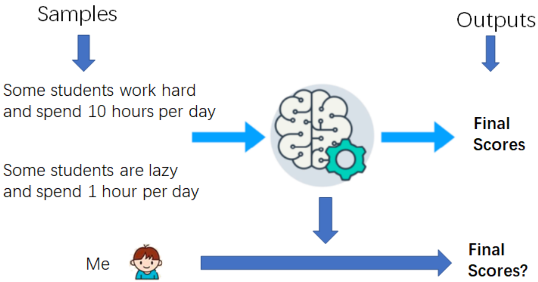
Classification

- Output is discrete
- Examples : given the study hours, in-class performance, and final grades (Pass or Fail) of past students, can you predict the final grades of the current students based on their study hours and in-class performance ?
- Applications : Credit risk evaluation, clinical prediction of tumor, classification of protein functions, etc.



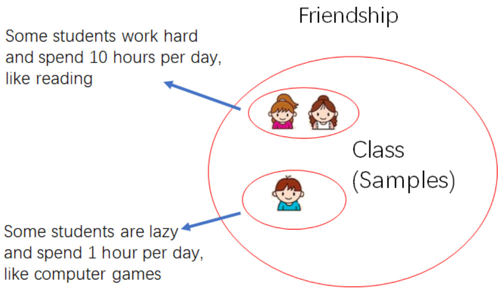
Regression

- Output is continuous
- Examples : given the study hours, in-class performance, and final scores of past students, can you predict the final scores of the current students based on their study hours and in-class performance ?
- Applications : epidemiology, finance, investment analysis, etc.



Unsupervised Learning

- No labels
- Optimize the parameters based on some natural rules, e.g., cohesion or divergence
- Clutering : K-Means, SOM

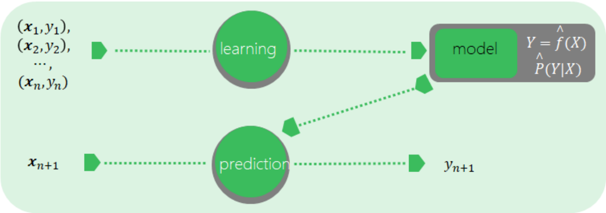


Representation of Data

- Input space $\mathcal{X} = \{\text{All possible samples}\}$; $\mathbf{x} \in \mathcal{X}$ is an input vector, also called feature, predictor, independent variable, etc. ; typically multi-dimensional ; e.g., $\mathbf{x} \in \mathbb{R}^p$ is a weight vector or coding vector
- Output space $\mathcal{Y} = \{\text{All possible results}\}$; $y \in \mathcal{Y}$ is an output vector, also called response, dependent variable, etc. ; typically one-dimensional ; e.g., $y = 0$ or 1 for classification problems, $y \in \mathbb{R}$ for regression problems.
- For supervised learning, assume that $(\mathbf{x}, y) \sim \mathcal{P}$, a joint distribution on the sample space $\mathcal{X} \times \mathcal{Y}$

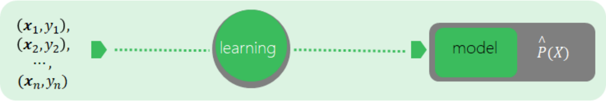
Supervised Learning

- Goal : given $\mathbf{x}$, predict what is y ; in deterministic settings, find the dependence relation $y = f(x)$; in probabilistic settings, find the conditional distribution $P(y|\mathbf{x})$ of y given $\mathbf{x}$
- Training dataset : $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} \mathcal{P}$, used to learn an approximation $\hat{f}(x)$ or $\hat{P}(y|\mathbf{x})$
- Test dataset : $\{(\mathbf{x}_j, y_j)\}_{j=n+1}^{n+m} \overset{i.i.d.}{\sim} \mathcal{P}$, used to make a prediction $\hat{y}_j = \hat{f}(\mathbf{x}_j)$ or $\hat{y}_j = \arg\max_{y_j} \hat{P}(y_j|\mathbf{x}_j)$, and verify how accurate the approximation is



Unsupervised Learning

- Goal : in probabilistic settings, find the distribution $P(\mathbf{x})$ of $\mathbf{x}$ and approximate it ; there is no y
- Training dataset : $\{\mathbf{x}_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} \mathcal{P}$, used to learn an approximation $\hat{P}(\mathbf{x})$; no test data in general



Learning Models

- Decision function (hypothesis) space :
  $\mathcal{F} = \{f_\theta | f_\theta = f_\theta(\mathbf{x}), \theta \in \Theta\}$ or $\mathcal{F} = \{P_\theta | P_\theta = P_\theta(y|\mathbf{x}), \theta \in \Theta\}$
- Loss function : a measure for the "goodness" of the prediction, $L(y, f(\mathbf{x}))$
  - 0-1 loss : $L(y, f(\mathbf{x})) = I_{y \neq f(\mathbf{x})} = 1 - I_{y = f(\mathbf{x})}$
  - Square loss : $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$
  - Absolute loss : $L(y, f(\mathbf{x})) = |y - f(\mathbf{x})|$
  - Cross-entropy loss :
    $L(y, f(\mathbf{x})) = -y \log f(\mathbf{x}) - (1 - y) \log(1 - f(\mathbf{x}))$
- Risk : in average sense,
  $R(f) = E_P[L(y, f(\mathbf{x}))] = \int_{\mathcal{X} \times \mathcal{Y}} L(y, f(\mathbf{x})) P(\mathbf{x}, y) \mathrm{d}\mathbf{x} \mathrm{d}y$
- Target of learning : choose the best f^* to minimize $R(f)$,
  $f^* = \min_f R(f)$

Risk Minimization Strategy

- Empirical risk minimization (ERM) : given training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i))$
  - By law of large number, $\lim_{n\to\infty} R_{emp}(f) = R(f)$
  - Optimization problem : $\min_{f\in F} \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i))$
- Structural risk minimization (SRM) : given training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and a complexity functional $J = J(f)$, $R_{srm}(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) + \lambda J(f)$
  - $J(f)$ measures how complex the model f is, typically the degree of complexity
  - $\lambda \geqslant 0$ is a tradeoff between the empirical risk and model complexity
  - Optimization problem : $\min_{f\in F} \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) + \lambda J(f)$

Algorithms

- Computational methods to solve the problem for f
- Numerical methods to solve the optimization problems
  - Gradient descent method, including coordinate descent, sequential minimal optimization (SMO), etc.
  - Newton's method and quasi-Newton's method
  - Combinatorial optimization
  - Genetic algorithms
  - Monte Carlo methods
  - ...

Model Assessment

Assume we have learned the model $y = \hat{f}(\mathbf{x})$, what is the error ?

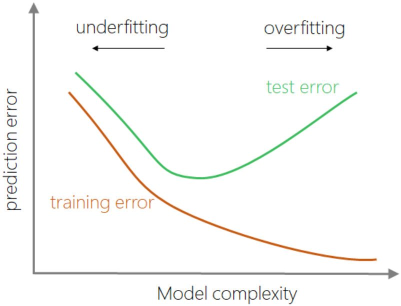
- Training error : $R_{emp}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}(\mathbf{x}_i))$, tells the difficulty of learning problem
- Test error : $e_{test}(\hat{f}) = \frac{1}{m}\sum_{j=n+1}^{n+m} L(y_j, \hat{f}(\mathbf{x}_j))$, tells the capability of prediction ; in particular, if 0-1 loss is used
  - Error rate : $e_{test}(\hat{f}) = \frac{1}{m}\sum_{j=n+1}^{n+m} I_{y_j \neq \hat{f}(\mathbf{x}_j)}$
  - Accuracy : $r_{test}(\hat{f}) = \frac{1}{m}\sum_{j=n+1}^{n+m} I_{y_j = \hat{f}(\mathbf{x}_j)}$
  - $e_{test} + r_{test} = 1$

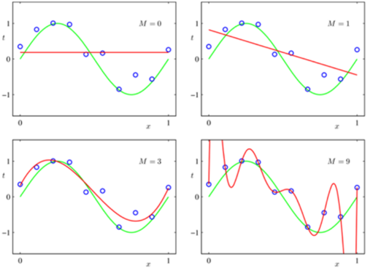
Model Assessment (Cont')

- Generalization error : $R_{exp}(\hat{f}) = E_P[L(y, \hat{f}(\mathbf{x}))] = \int_{\mathcal{X}\times\mathcal{Y}} L(y, \hat{f}(\mathbf{x}))P(\mathbf{x}, y)\mathrm{d}\mathbf{x}\mathrm{d}y$, tells the capability for predicting unknown data from the same distribution, its upper bound M defines the generalization ability
  - As $n \to \infty$, $M \to 0$
  - As F becomes larger, M increases

Overfitting

- Too many model parameters
- Better for training set, but worse for test set





fitting of degree M polynomial, green curve is the ground truth

Model Selection

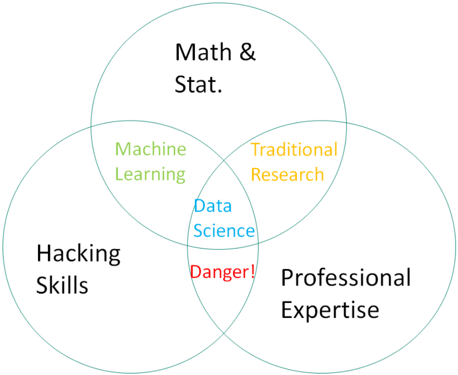
- Regularization : $\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) + \underbrace{\lambda J(f)}_{penalty}$, choose λ to minimize empirical risk and model complexity simultaneously
- Cross-validation (CV) : split the training set into training subset and validation subset, use training set to train different models repeatedly, use validation set to select the best model with the smallest (validation) error
  - Simple CV : randomly split the data into two subsets
  - K-fold CV : randomly split the data into K disjoint subsets with the same size, treat the union of $K-1$ subsets as training set, the other one as validation set, do this repeatedly and select the best model with smallest mean (validation) error
  - Leave-one-out CV : $K = n$ in the previous case

1	2	3	4	5
Train	Train	Validation	Train	Train

Outlines

Data Science VS. Other Techniques

Data science is an interdisciplinary area using mathematics, statistics, computer science and engineering, and other profession techniques



Where Math and Statistics Emerge