Intelligent System Development (知能システム開発特論)
Syllabus

Offering: Fall 2020

Course Information

Credit Hours: 1.5hours/week

Semester: Fall 2020

Meeting time and location: Tuesday 10:30, Online Meet

Course website: http://p-www.iwate-pu.ac.jp/~david/ids

Course Description

Data Science is the study of the generalizable extraction of knowledge from data. Being a data scientist requires an integrated skill set spanning mathematics, statistics, machine learning, databases and other branches of computer science along with a good understanding of the craft of problem formulation to engineer effective solutions. This lecture series will introduce students to this rapidly growing field and equip them with some of its basic principles and tools as well as its general mindset.

Students will learn concepts, techniques and tools they need to deal with various facets of data science practice, including data collection and integration, exploratory data analysis, predictive modeling, descriptive modeling, data product creation, evaluation, and effective communication.

Learning Outcomes

At the end of the 7 lectures and practices, students should be able to:

- Describe what Data Science is
- Explain in basic terms what Statistical modeling means. Identify probability distributions, fit a model to a data.
- Use Python/R to carry out basic statistical modeling and analysis
- Explain the significance of exploratory data analysis (EDA) in data science.
- Apply EDA and the Data science process in a case study.
- Apply basic machine learning algorithms
- Identify feature generation
- Identify and explain fundamental mathematical and algorithmic ingredients for data mining

Grading

- Class participation (30%)
- Assignment (30%)
- Presentation (40%)

Topics and course outline:

1. Introduction: What is Data Science?
2. Understanding a problem
3. Modeling and Data Visualization
4. Feature Generation and Feature Selection
5. Exploratory Data Analysis and the Data Science Process
6. Basic Machine Learning Algorithms
7. Text Summarization/ Text mining/ Topic Modeling with Language Model (n-grams, word2vec, neural networks, LDA)

Preparation:

RStudio, R programming　or Anaconda Python programming

Natural Language Processing, Image Processing, Tabular Processing

Kaggle Datasets

Books:

- Data Science from Scratch (Joel Grus)

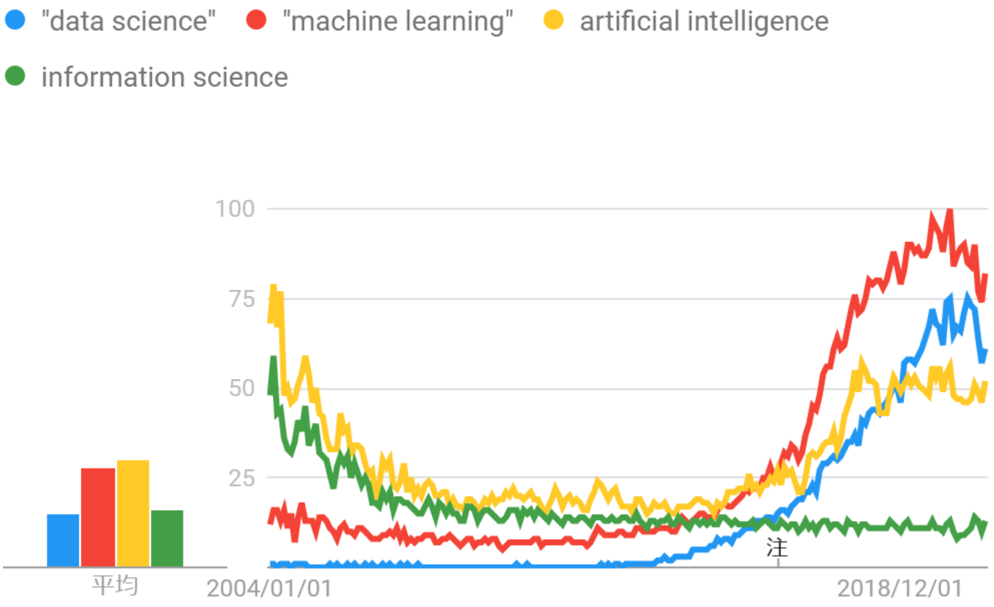
ゼロからはじめるデータサイエンス ―Pythonで学ぶ基本と実践 (Joel Grus (著), 菊池 彰 (翻訳))

- Rによるデータサイエンス（金　明哲）
- 有賀 友紀; 大橋 俊介 (2019-03-26). R と Python で学ぶ［実践的］データサイエンス＆機械学習 (Kindle の位置 No.522-523). 株式会社技術評論社

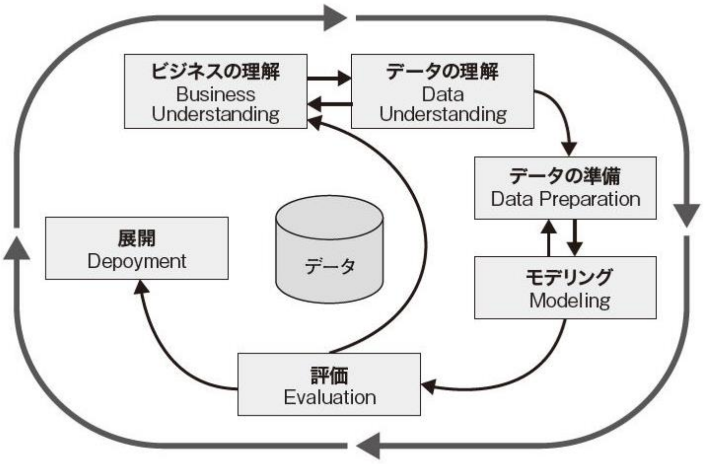
DATA SCIENCE, DATA ENGINEERING AND MACHINE LEARNING SCIENTIST

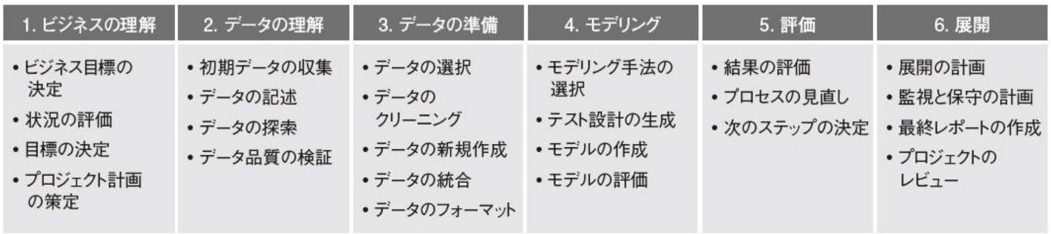
Follows the scientific method:

- Observation. Gathering of data and facts or prior knowledge
- Question: determining a good question can be very difficult and it will affect the outcome of the investigation
- Hypothesis: a scientific hypothesis must be falsifiable(反証可能性), meaning that one can identify a possible outcome of an experiment that conflicts with predictions deduced from the hypothesis; otherwise, it cannot be meaningfully tested
- Experiment: Modeling, Prediction and Testing
- Results: Comparison, Analysis
- Conclusion: future inquiries, summary



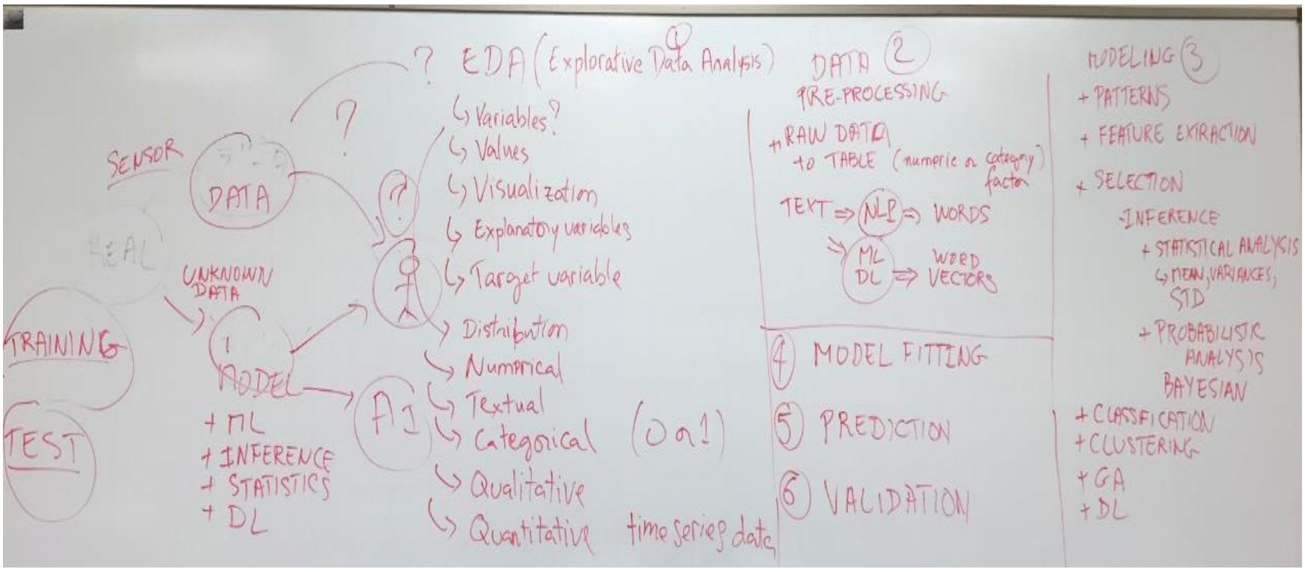
Data Science Life Cycle



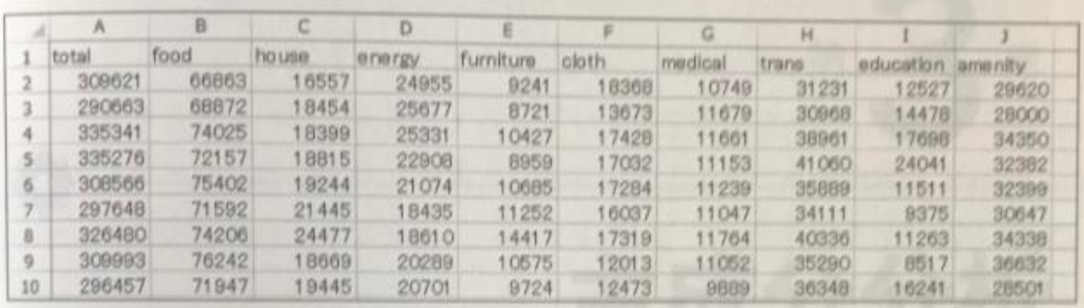


1. ビジネスの理解	2. データの理解	3. データの準備	4. モデリング	5. 評価	6. 展開
• ビジネス目標の決定 • 状況の評価 • 目標の決定 • プロジェクト計画の策定	• 初期データの収集 • データの記述 • データの探索 • データ品質の検証	• データの選択 • データのクリーニング • データの新規作成 • データの統合 • データのフォーマット	• モデリング手法の選択 • テスト設計の生成 • モデルの作成 • モデルの評価	• 結果の評価 • プロセスの見直し • 次のステップの決定	• 展開の計画 • 監視と保守の計画 • 最終レポートの作成 • プロジェクトのレビュー

図1.4　CRISP-DMで定義されている6つのフェーズ



総務統計局の家計調査（2000 年以降の時系列結果—2 人以上の世帯）のデータを以下に示す．



	A	B	C	D	E	F	G	H	I	J
1	total	food	house	energy	furniture	cloth	medical	trans	education	amenity
2	309621	66863	16557	24955	9241	18368	10749	31231	12527	29620
3	290663	68872	18454	25677	8721	13673	11679	30968	14478	28000
4	335341	74025	18399	25331	10427	17428	11661	38961	17698	34350
5	335276	72157	18815	22908	8959	17032	11153	41060	24041	32382
6	308566	75402	19244	21074	10685	17284	11239	35889	11511	32399
7	297648	71592	21445	18435	11252	16037	11047	34111	9375	30647
8	326480	74206	24477	18610	14417	17319	11764	40336	11263	34338
9	309993	76242	18669	20289	10575	12013	11052	35290	8517	36632
10	296457	71947	19445	20701	9724	12473	9889	36348	16241	28501

図 2.5　家計調査データの一部

左から，消費支出（総支出，total），食料（food），住居（house），光熱・水道（energy），家具・家事用品（furniture），被服及び履物（cloth），保健医療（medical），交通・通信（trans），教育（education），教養娯楽（amenity），その他の消費支出（others）となっている．その他の消費支出には，お小遣い，交際費，仕送りなどが含まれている．

このデータについて以下の操作を行いなさい．

① 家計の総支出に対する支出項目ごとの相関分析を行いなさい．

② 家計の総支出の線形重回帰式を決定しなさい．

https://www.kaggle.com/startupsci/titanic-data-science-solutions/notebook

https://www.kaggle.com/startupsci/titanic-data-science-solutions/data

Titanic: Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics

About this Competition

Overview

The data has been split into two groups:

- training set (train.csv)
- test set (test.csv)

The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the “ground truth”) for each passenger. Your model will be based on “features” like passengers’ gender and class. You can also use feature engineering to create new features.

The test set should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each passenger. It is your job to predict these outcomes. For each passenger in the test set, use the model you trained to predict whether or not they survived the sinking of the Titanic.

We also include **gender_submission.csv**, a set of predictions that assume all and only female passengers survive, as an example of what a submission file should look like.

Data Dictionary

VariableDefinitionKey survival Survival 0 = No, 1 = Yes pclass Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd sex Sex Age Age in years sibsp # of siblings / spouses aboard the Titanic parch # of parents / children aboard the Titanic ticket Ticket number fare Passenger fare cabin Cabin number embarked Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton

Variable Notes

pclass: A proxy for socio-economic status (SES)
1st = Upper
2nd = Middle
3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way...
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Some children travelled only with a nanny, therefore parch=0 for them.