クレジット:
UTokyo Online Educationデータマイニング入門 2018 森　純一郎

東京大学
THE UNIVERSITY OF TOKYO

CC BY NC ND

データマイニング入門 第12回

2018年度

データ分析のフレームワークの例
データの理解 [Foster Provost: Data Science for Business, 2013]

どのような問題を解決したいのか

教師ありの問題に帰着するか、教師なしの問題に帰着するか

特徴量は正しく設計されているか

教師ありの場合

- 目的変数は正しく設計されているか

教師なしの場合

- 探索的なデータ分析の方向性（次にどこに進むべきか）が決められているか

データの準備

特徴量のためのデータを取得し、特徴ベクトルを作成し、データ行列を構成することは可能か

- 代替のデータ形式は考えられるか

教師ありの場合、訓練データと検証・テストデータの目的変数を取得し、それらをデータ行列として構成することは可能か

教師ありの場合、目的変数の値をどのように学習させるか決めらているか

学習データは、そのモデルが適用される母集団の性質に従っているか

データ分析のフレームワークの例

モデリング

選択したモデルは目的変数に対して適切か

- 回帰、分類、クラスタリング、確率推定、ランキングなど

選択したモデルの設計は適切か

- コスト関数
- 距離・類似尺度 など

選択したモデルが現実的に適用可能か

- 汎用性、モデルの理解しやすさ、モデル構築に必要な時間、実展開の時間、必要なデータ量、頑健性（欠損や外れ値対応など）、ハイパーパラメータやモデルの調整

適切な評価尺度と方法に従ってモデルの選択と評価を行ったか

データ分析のフレームワークの例

評価と展開

モデルの実展開前に妥当性をレビューできるか、モデルは理解可能か

モデルの実展開のコストと利益を考慮しているか

モデルの出力をどのように使うかについて合意がとれているか

モデルはテストデータや交差検証で評価されているか

モデルの性能評価の基準値があるか、それらと比較可能か

モデルの実展開により、実問題が実際にどのように解決されるか

学習内容（予定）

データ分析のためのPythonプログラミング
- Pythonの基本的な文法とNumpy, pandas, matplotlib, scikit-learnなどの科学技術計算のための主要なモジュール

データ分析のための数理的基礎
- 記述統計、確率分布、ベクトル・行列、固有値分解、最適化基礎など

データの前処理・加工とデータベース
- 欠損値・ノイズ・外れ値の処理、~~関係データベース~~など

テキストデータ分析の基礎
- tfidf、Bag of Words、ベクトル空間モデル、形態素解析、類似度、~~潜在意味解析~~など

ネットワーク・グラフデータ分析の基礎
- 隣接行列, 最短距離, クラスタリング係数, 中心性、コミュニティ抽出、ネットワークの数理モデル など

機械学習の基礎（教師あり学習）
- 線形回帰、ロジスティック回帰、正則化、モデル選択、ニューラルネットワーク など

機械学習の基礎（教師なし学習）
- k-means、階層化クラスタリング、EMアルゴリズム、主成分分析など

全体学習目標

データ分析の基本的なプロセスを理解する
- 前処理、管理、分析、評価

データ分析のためのプログラミング基礎を理解する
- Pythonを用いる

データ分析のための数理的基礎を理解する
- 特に、確率・統計、線形代数、解析学

代表的なデータの数理モデルと基礎的な処理を理解する
- テキストデータ、グラフデータ、~~時系列データ~~

データマイニング・機械学習の基礎的な手法を理解する

データ分析プロジェクトの基礎的な設計と実施ができる

プログラミング、記述統計、前処理

データ分析のためのプログラミング基礎を理解する
- Python, pandas, numpy, matplotlib

記述統計と可視化について理解する
- 中心傾向、ばらつき、分布、相関、可視化

前処理について理解する
- 欠損値、外れ値、標準化・スケーリング、変換

テキスト分析

学習目標

- テキスト分析の基本的な処理の流れを理解する
- テキストの基本的な前処理を理解する
  - トークナイゼーション、ストップワード、ステミングなど
- 形態素とPOSタギングについて基本を理解する
- テキストのベクトル空間モデルについて基本を理解する
- 単語の重みづけについて基本を理解する
- テキスト間の類似度について基本を理解する
- 基本的なテキスト処理の実装を理解する

ネットワーク分析

学習目標

- ネットワークデータの行列表現（隣接行列）を理解する
- 重み付き、有向ネットワークについて理解する
- ネットワークの最短経路長について理解する
- ネットワークの中心性について理解する
- 平均パス長、クラスタリング係数について理解する
- ネットワークのコミュニティ抽出について理解する
- ネットワークデータの基本的な処理の実装について理解する

教師なし学習・クラスタリング

学習目標

- クラスタリングの概念について理解する
- 特徴量ベクトルと距離・類似度について理解する
- 階層的クラスタリングについて理解する
  - 併合の方法
  - デンドログラム
- k-meansクラスタリングについて理解する
  - コスト関数
- k-means法の確率的な解釈を通して教師なし学習の統計モデルの基礎を理解する
- クラスタリングの実装について理解する

教師なし学習・主成分分析

学習目標

- 次元削減（縮約）について理解する
- 共分散行列とその固有値・固有ベクトルについて理解する
- 主成分分析について理解する
  - 射影と分散の最大化について
- 主成分分析の実装について理解する

教師あり学習・線形回帰

学習目標

- 機械学習の概念について理解する
- 教師あり学習・回帰の概念について理解する
- 線形回帰
  - コスト関数の最小化について理解する
  - パラメータの推定
    - 最急降下法について理解する
  - パラメータの解析解
    - 正規方程式について理解する
- 線形回帰の実装を理解する

教師あり学習・ロジスティック回帰

学習目標

- ロジスティック回帰について理解する
  - 2クラス分類について理解する
  - シグモイド関数について理解する
  - 決定境界について理解する
  - コスト関数（交差エントロピー）の最小化について理解する
  - パラメータの推定について理解する
- 多クラス分類について理解する
  - ソフトマックス関数

機械学習の実践

学習目標

- 非線形性、基底関数の導入について理解する
  - 多項式
- 過学習と正則化について理解する
  - L2正則化
- 特徴量エンジニアリングについて理解する
  - 特量選択、特徴量作成
- モデル評価と選択について理解する
  - 交差検証、混同行列
- BiasとVarianceについて理解する

数学について

確率・統計
- 記述統計、確率、確率変数、確率分布、大数の法則・中心極限定理、標本分布、点推定、区間推定、仮説検定など
  - 統計データ解析I・II参照

線形代数
- ベクトル、行列、内積、線形変換、固有値・固有ベクトル、対角化、行列式など
  - 前期課程 線型代数学、数学II （文科生） 、数理科学概論II（文科生）参照

解析学
- 多変数の微積分
  - 極限と連続性、偏微分、ベクトル解析、など
    - 前期課程 数学I （文科生） 、数理科学概論I（文科生）、 微積分学、ベクトル解析参照

最適化
- 凸最適化、非線形計画法、など
  - 数理手法III参照

後期課程におけるデータ分析関連講義

http://www.mi.u-tokyo.ac.jp/lectures.html

工：機械学習の数理、統計的機械学習、知能機械情報学、言語・音声情報処理、確率数理工学、システムデータ解析、応用データ解析 など

理：統計的機械学習、知能システム論、生物データマイニング論、生命情報表現論、地球物理データ解析 など

農：バイオインフォマティクス など

薬：生物統計学 など

医：医学データの統計解析、バイオインフォマティクス など

経：経済データ分析、計量経済学、ベイズ計量経済学、経営科学 など

文：社会心理学、心理学統計、社会学のためのデータ分析法 など

教育：教育政策研究方法論、教育社会学調査実習 など

教養：言語データ分析、計算社会科学、意思決定・知的システム論 など

さらに勉強するために

著作権の都合により
ここに挿入されていた画像を削除しました

書籍『統計的自然言語処理の基礎』表紙
Christopher D.Manning ・Hinrich Schutze 著
加藤 恒昭・菊井 玄一郎・林 良彦・森 辰則訳
共立出版 (2017.11)
https://www.kyoritsu-pub.co.jp/bookdetail/9784320124219

著作権の都合により
ここに挿入されていた画像を削除しました

書籍『Networks : An Introduction』表紙
M. E. J. Newman 著
Oxford Univ Pr (2010/5/20)
http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199206650.001.0001/acprof-9780199206650

著作権の都合により
ここに挿入されていた画像を削除しました

書籍『Machine Learning :
A Probabilistic Perspective』表紙
Kevin P. Murphy 著
MIT Press (2012.8)
https://mitpress.mit.edu/books/machine-learning-1

