情報科学 3

高知工科大学　経済・マネジメント学群
2021 年度第 3 クォータ

開講日時：火曜・金曜 4 限
教室：A204
オフィスアワー： 月曜 4 限
それ以外の時間は**要予約**

担当：矢内 勇生（やない ゆうき）
研究室：A625
Email: yanai.yuki@kochi-tech.ac.jp
Website: https://yukiyanai.github.io

講義の概要と目的

この講義では機械学習 (machine learning, マシンラーニング) の基礎を学ぶ。現代はビッグデータの時代と呼ばれ、データ解析に AI（人工知能）が活用されている。コンピュータにデータを入力すると、AI がデータに隠されたパタンや特徴を見つけ出し、予測や分類を行ってくれる。たとえば、金融分野では企業・個人の信用評価を、マーケティング分野では消費者の好みに応じた商品レコメンドなどを、AI が半ば自動的に行ってくれる。これらの例で使われる AI とは、高度な機械学習や深層学習のことである。ビッグデータ時代のデータサイエンスは、機械学習によって支えられている。

そこでこの講義では、現代のデータサイエンスに欠かせない機械学習の基本的な仕組みを学び、R を使って機械学習を実行する方法を身につける。近年では高校の授業にも機械学習・データサイエンスの内容が取り入れられており、高校数学の教員を目指す学生にとっても、機械学習の基礎を理解することは有益である。

履修要件と関連科目

以下の 1, 2, 3 をすべて履修済み（単位取得済み）であることを前提に授業を行う。

1.「数学 1」または「微分積分学 1」または「経済学で使う数学」
2.「数学 4」または「線形代数学 1」
3.「統計学 2」

これらの科目を履修済みでないと、この授業の内容をあまり理解することはできないだろう。したがって、単位を取得することも難しいと思われる。

「計量経済学」を受講済みでない場合は、この授業と同時に履修することを推奨する。機械学習は計量経済学と多くの手法を共有しているので、計量経済学の内容を理解することにより、機械学習の理解も深めることができる。「情報科学 1」「情報科学 2」「プログラミング」も履修済みであることが望ましいが、必須ではない。

統計学 2 を履修済みであることを履修要件にしているので、R (RStudio) の基本操作は理解しているものとみなし、この授業では R (RStudio) の基本操作は教えない。使い方を忘れている者は、統計学 2（あるいは計量経済学）の内容をよく復習しておくこと。**R の使い方を知らない者は、この科目を履修すべきではない**。R の代わりに Python や Julia を使って課題をこなしてもよいが、授業で説明するのは R を用いる方法のみなので、他の言語を使いたい場合は自力で学習すること。

授業の方法

この授業は、情報演習室（コンピュータ教室）で講義とコンピュータ実習を織り交ぜて行う。講義と実習の時間配分は内容に応じて変わる。**実習科目なので、対面授業の録画ならびにオンライン配信は行わない**。新型コロナウィルスの感染状況によってはすべての受業をオンラインで実施することになるかもしれないが、その際の授業実施方法については決まりしだい連絡する。

実習では情報演習室に設置されているコンピュータを使うことができるが、自分のパソコンを持ち込んでもよい。自分のパソコンを持参する場合は、R (バージョン 4.0.3 を推奨[1]) と RStudio（バージョン 1.4 以降を推奨）をあらかじめインストールしておくこと。また、電源コンセントの数に限りがある（数人分しかない）ので、授業前に十分に充電してくること。情報演習室のディスプレイは HDMI ポートが空いているので、HDMI 接続用のケーブルを持参すれば、自分のパソコンの画面と併せて 2 画面で授業を受けることができる。

成績評価

成績は、以下の要素によって構成される。

- 授業への参加 [単なる出席は参加ではない]（最終成績の 30%）
- 課題の提出状況と完成度（40%）
- 期末プロジェクトの完成度（30%）

成績の目安は次のとおりである。C の条件を満たさない者は F とする。

C 機械学習の基本的な仕組みを理解している。
B C の条件に加え、機械学習を用いた予測・分類を実際に行うことができる。
A B の条件に加え、機械学習の手法を用いて社会科学のリサーチクエスチョンに答えることができる。
AA A の条件に加え、期末プロジェクトの内容が特に優れている場合。

Slack

授業時間外のコミュニケーションツールとして、Slack を使う。この授業の Slack ワークスペース は https://kut-info3-2021.slack.com であり、次のリンクから登録可能である。

https://join.slack.com/t/kut-info3-2021/signup

ただし、登録には KUT ドメインのメールアドレス（`@ugs.kochi-tech.ac.jp`） が必要である。

KUT 以外のメールアドレスでの登録を希望する場合は、以下の内容のメールを担当教員宛に送ること。

- 件名（メールのタイトル）:「情報科学 3 Slack 用メールアドレス」
- 本文に以下の内容を記載
  - 氏名
  - 学籍番号
  - Slack への登録で使いたいメールアドレス

必要事項が記載されたメールが届きしだい招待状を送る。Slack から届く招待状を確認して登録すること。

Slack における質問、回答、議論は、授業への貢献とみなし、内容に応じて参加点を加算する。授業に無関係の内容や議論を妨害するような投稿でない限り、減点はしない。

R、RStudio、R Markdown

この授業では、オープンソースの統計処理言語である R を用いてデータの収集、管理、分析を行う。また、R を使うための統合開発環境 (IDE) として、RStudio を使う。R、RStudio とも無料であり、各自のコンピュータ（Linux, Mac, Windows）にインストールすることができる。詳細については、「統計学 2」のページを参照されたい。

教科書

指定しない。予習・復習用教材は KUTLMS (moodle) で配布する。

1) 最新版は 4.1.1 だが、情報演習室のパソコンに合わせたバージョンで授業を進める。4.0.3 より新しいバージョンでも大きな問題はないはずである（ただし、4.0.4 は日本語が正しく表示できないので不可）。4.0.3 の macOS 用 (R-4.0.3.pkg) は https://cran.r-project.org/bin/macosx/base/ で、Windows 用 (R-4.0.3-win.exe) は https://cloud.r-project.org/bin/windows/base/old/4.0.3/ で入手できる。

参考書

授業内容の理解を助けると思われる本を以下に挙げる。購入する必要はない。

- 浅野正彦, 矢内勇生. 2018.『R による計量政治学』オーム社.
- Flach, P.（竹村 監訳） 2017.『機械学習』朝倉書店.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman（杉山ほか訳） 2014.『統計的学習の基礎：データマイニング・推論・予測』共立出版.
- 林賢一, 下平英寿. 2020.『R で学ぶ統計的データ解析』講談社.
- 岩沢宏和, 平松雄司. 2019.『入門 R による予測モデリング：機械学習を用いたリスク管理のために』東京図書.
- James, G., W. Witten, T. Hastie, and T. Tibshirani（落海, 首藤 訳）2018.『R による統計的学習入門』朝倉書店.
- 経済セミナー編集部（編）2020.『機械学習は経済学を変えるのか？ 経済セミナー e-Book（Kindle 版）』日本評論社.
- Lantz, B.（株式会社クイープ 監訳） 2021.『R による機械学習 [第 3 版]』翔泳社.
- Larose, Chantal D., and Daniel T. Larose（阿部, 西村 訳）2020.『Python, R で学ぶデータサイエンス』東京化学同人.
- 中谷秀洋. 2019.『わけがわかる機械学習』技術評論社.
- 櫻井豊. 2019.『機械学習ガイドブック：R と Python を使いこなす』オーム社.
- 杉山将. 2013.『イラストで学ぶ機械学習』講談社.
- 鈴木譲. 2020a.『統計的機械学習の数理 100 問 with R』共立出版.
- 鈴木譲. 2020b.『スパース推定 100 問 with R』共立出版.
- 横内大介, 青木義充. 2017.『イメージでつかむ機械学習入門：豊富なグラフ, シンプルな数学, R で理解する』技術評論社.

SA（学生アシスタント）

この授業には実習を補助してくれる SA（学生アシスタント）がいるので、実習でわからないことがあれば SA にも質問してほしい。ただし、以下の注意を守ること。

- SA への質問は、コンピュータの使い方に関するものに限定する。機械学習の内容（理論やアルゴリズム）については必ず教員に質問すること。
- 授業時間外に SA に対して授業に関する質問をすることは**禁止**する。SA が給料をもらえるのは授業時間中だけであり、授業時間外に質問に答える義務はない。
- SA には礼節をもって接すること。SA に対する暴言や暴力などは授業妨害であり、不正行為として扱う。

授業計画

授業計画は以下の通りである。ただし、授業の進捗状況に応じて変更する可能性がある。変更する際はこの講義要綱（シラバス）を更新し、授業中に案内する。

1. イントロダクション（10 月 1 日 [金] ）

まず、授業の進め方、概要、成績評価の方法などを確認する。その後、機械学習の概要を学ぶ。

予習・復習 依田高典. 2020.「経済分析ツールとしての機械学習」『経済セミナー』711: 23–27.

予習・復習 櫻井 (2019) 4 章

参考 Athey, Susan, and Guido W. Imbens. 2019. “Machine Learning Methods That Economists Should Know About.” *Annual Review of Economics* 11: 685–725.

参考 Varian, Hal R. 2014. “Big Data: New Tricks for Econometrics.” *Journal of Economic Perspectives* 28(2): 3–28.

参考 エリック・ブリニョルフソン, アンドリュー・マカフィー（村井章子 訳） 2015.『ザ・セカンド・マシン・エイジ』日経 BP.

2. 学習モデル （10 月 5 日 [火] ）

機械学習で用いられるいくつかの代表的なモデルを学ぶ。

予習・復習 櫻井 (2019) 1 章

参考 岩沢・平松 (2019) 1–2 章

参考 杉山 (2013) 1–2 章

参考 Flach (2017) 1 章

参考 Donoho, David. 2017. “50 Years of Data Science.” *Journal of Computational and Graphical Statistics* 26(4): 745–766.

3. 最小二乗学習（10 月 8 日 [金] ）

予測のための最も基本的な手法である最小二乗学習を理解する。

予習・復習 横内・青木 (2017): pp.14-51.

参考 浅野・矢内 (2018) 10–14 章

参考 鈴木 (2020a) 1 章

4. 制約付き最小二乗学習 (1)（10 月 12 日 [火]）

最小二乗学習で問題となる過剰適合に対処するための、制約付き最小二乗学習の手法を学ぶ。

予習・復習 杉山 (2013) 4 章

参考 中谷 (2019) 4 章

参考 鈴木 (2020a) 5 章

5. 制約付き最小二乗学習 (2) （10 月 15 日 [金]）

推定するパラメタの数が多いときに有効な、スパース学習の手法を学ぶ。

予習・復習 杉山 (2013) 5 章

参考 鈴木 (2020b) 1 章

6. 遅延（怠惰）学習 （10 月 19 日 [火]）

最近傍法による分類手法を学び、それが遅延（怠惰）学習 (lazy learning) と呼ばれる理由を理解する。

予習・復習 Flach (2017) 8 章
参考 有賀康顕, 中山心太, 西林孝. 2021.『仕事ではじめる機械学習 第 2 版』オライリー・ジャパン：57-61
参考 後藤正幸, 小林学 (2014) 『入門 パターン認識と機械学習』コロナ社：6 章

7. 確率的学習（10 月 22 日 [金]）

テキストデータをベイズ分類器によって分類する手法を学ぶ。

予習・復習 Larose and Larose (2020) 8 章
参考 岡田謙介, 矢内勇生 (近刊) 『ベイズ統計分析入門：実証社会科学のための統計モデリング』有斐閣: 2–3 章
参考 佐々木淳 (2021)『いちばんやさしい ベイズ統計入門』SB クリエイティブ.

8. ロジスティック回帰 (1)（10 月 26 日 [火]）

ロジスティック回帰による確率的分類の手法を学ぶ。

予習・復習 浅野・矢内 (2018) 15 章
参考 林・下平 (2020) 6 章

9. ロジスティック回帰 (2) （10 月 29 日 [金]）

引き続き、ロジスティック回帰による確率的分類の手法を学ぶ。

予習・復習 浅野・矢内 (2018) 15 章
参考 Larose and Larose (2020) 7 章
参考 林・下平 (2020) 6 章

10. 分割統治 I: 決定木 （11 月 2 日 [火]）

分割統治法 (divide-and-conquer method) による分類手法を学ぶ。特に、決定木と呼ばれる手法を理解する。

予習・復習 Lantz (2021) 5 章 [123–145]
参考 林・下平 (2020) 7 章
参考 Flach (2017) 5 章
参考 Larose and Larose (2020) 6 章

11. 分割統治 II: ルールモデル （11 月 5 日 [金]）

引き続き、分割統治法による分類手法を学ぶ。特に、分類ルール学習器を理解する。

予習・復習 Lantz (2021) 5 章 [145–161]
参考 Flach (2017) 6 章

12. クラスタリング （11 月 9 日 [火]）

教師なし分類の基本的な手法であるクラスタリングの方法を学ぶ。

予習・復習 Larose and Larose (2020) 10 章
参考 林・下平 (2020) 第 9 章
参考 鈴木顕 (2021)『機械学習アルゴリズム』共立出版：5 章

13. モデルの評価 （11 月 12 日 [金]）

機械学習モデルの性能を評価する方法を理解する。

予習・復習 Lantz (2021) 10 章

参考 有賀ほか (2021) 3 章

参考 森下光之助. (2021)『機械学習を解釈する技術：予測力と説明力を両立する実践テクニック』技術評論社.

14. モデルの改善（11 月 16 日 [火]）

機械学習モデルの性能を向上させるために取りうる手段について学ぶ。

予習・復習 Lantz (2021) 11 章

参考 Flach (2017) 10-11 章

参考 Thakur, A.（石原 訳） (2021)『Kaggle Grandmaster に学ぶ機械学習実践アプローチ』マイナビ出版.

15. 全体のまとめ （11 月 19 日 [金]）

講義全体を振り返り、機械学習について理解できた点と、さらなる学習が必要な点を明らかにする。

復習 講義全体を復習せよ

参考 Athey, Susan. 2017. “Beyond Prediction: Using Big Data for Policy Problems.” *Science* 355: 483–485. https://science.sciencemag.org/content/355/6324/483

参考 Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. “The Parable of Google Flu: Traps in Big Data Analysis.” *Science* 343: 1203–1205. https://science.sciencemag.org/content/343/6176/1203

期末プロジェクト

授業で学習した機械学習の手法を用いてデータ分析を行い、その分析過程と結果をレポートにまとめて提出しなさい。

- 研究上の疑問点や仮説を明示すること。
- データは自分で集めること（インターネット上で利用可能なデータを使ってかまわない）。
- 分析に用いたコンピュータのコードとデータセットも提出する。
- 分量は自由（必要な分だけ）：A4 用紙 4 ページ程度（文章 +と図、表、参考文献リスト）を想定しているが、厳密に考えなくてよい
- 主な分析結果は図・表にまとめる。
- 図表の内容は、すべて文章で説明する（**図・表のみのレポートは不合格**）。

レポートに関する注意

- データを自分で集める必要があるので、できるだけ早く取り組むこと。遅くとも 11 月の第 1 週までにはデータを集め始めることを強く勧める。
- レポートの主な内容は図や表ではなく、文章である。分析結果を図・表にまとめただけで文章が極端に少ないレポートは、分析内容自体が良くても高評価を得ることはない。
- 他人の文章の剽窃や盗用（コピペ）は不正行為である。この授業の成績が F になるだけでなく、今学期に履修するすべての科目が不合格になるので注意されたい。
- 締め切り後に提出されたレポートは 0 点にする。正当な理由がある場合についてはこの限りではないので、

期限までに提出できない事情が発生したら速やかに連絡すること。部活の大会や大学内外の行事など、あらかじめ日程がわかっているイベントに参加することは正当な理由にはならないので注意されたい。

- 良い成績がほしい（A 以上の成績を取りたい）なら、レポートの書き方についても勉強することを勧める。たとえば、以下にあげる参考書のどれかを一読されたい。
  - 戸田山和久. 2012.『新版論文の教室』NHK 出版.
  - 小笠原喜康. 2009.『新版大学生のためのレポート・論文術』講談社.
  - 石黒圭. 2012.『この 1 冊できちんと書ける！ 論文・レポートの基本』日本実業出版社.
  - 野田直人. 2015.『小論文・レポートの書き方：パラグラフ・ライティングとアウトラインを鍛える演習帳』人の森.
  - 倉島保美. 2012.『論理が伝わる世界標準の「書く技術」:「パラグラフ・ライティング」入門』講談社.

提出期限：2021 年 12 月 1 日（水）正午（日本時間）

以下の 3 つを、Slack のダイレクトメッセージ (DM) で担当教員に提出しなさい。ただし、ファイル名の `YourID` は自分の学籍番号に変えること。

1. データの前処理と分析に必要な R コードがすべて含まれた R マークダウンファイル
   - ファイル名：`info3_final_YourID.Rmd`
2. 1 の R マークダウンファイルを knit してできた PDF ファイル（R コードは非表示でよい）
   - ファイル名：`info3_final_YourID.pdf`
3. 分析に利用したデータ
   - 複数ある場合はすべて提出する。
   - データを提出できない事情がある場合は、11 月 22 日（月）までに連絡すること。