情報理工学演習IV後半

データサイエンス

day 2: コンペティション

担当:野田 五十樹

第二回の目標

- 1. コンペティションの流れを理解する
- 2. データをダウンロードする
- 3. タスクを理解する
- 4. データを理解する
- 5. 一度提出を行う

「教師あり学習」と「教師あり学習のプロセス」
について想像できない人は、
17ページ以降の付録を参照

第二回の目標

- **1. コンペティションの流れを理解する**
- 2. データをダウンロードする
- 3. タスクを理解する
- 4. データを理解する
- 5. 一度提出を行う

コンペティションで与えられるもの

- 教師あり学習のコンペでは主に以下の二つが与えられる
  - 1. 訓練データ
    - 目標値（正解）が与えられているデータ
    - 参加者はこのデータを用いて予測モデルを作る（学習させる）
  - 2. テストデータ
    - 参加者には目標値（正解）が与えられていないデータ
    - 運営側は正解が分かっている
    - 参加者は「テストデータに対する予測」を提出
    - 提出された予測結果に対して，良さ（悪さ）を表すスコアが計算され，ランキング付けされる

コンペティションの流れ

- 1. データをダウンロードする
- 2. タスクを理解する
- 3. データを理解する
- 4. While 期日まで do:
  - I. 訓練データを用いて, 特徴設計, モデルの選択… などを行い, 予測モデルを構築
  - II. (I)で学習した予測モデルを用いてテストデータの正解を予測
  - III. (II)で予測した結果を提出 (**1日5回まで**)
  - IV. 提出結果のスコアが計算・ランキングが更新されるので, 一喜一憂しつつ次回以降の投稿に活かす

第二回の目標

- 1. コンペティションの流れを理解する
- **2. データをダウンロードする**
- 3. タスクを理解する
- 4. データを理解する
- 5. 一度提出を行う

データをダウンロードする

- ダウンロードしたdataフォルダの中に以下のファイルがあればよい
  - train.csv
  - test.csv
  - y_pred_example.txt
- 配布するサンプルプログラムを使う場合，**dataフォルダを2022フォルダの中に入れて置くこと**
- Google Colabを用いる場合は，dataフォルダを**Google Driveにアップロードしておくこと**

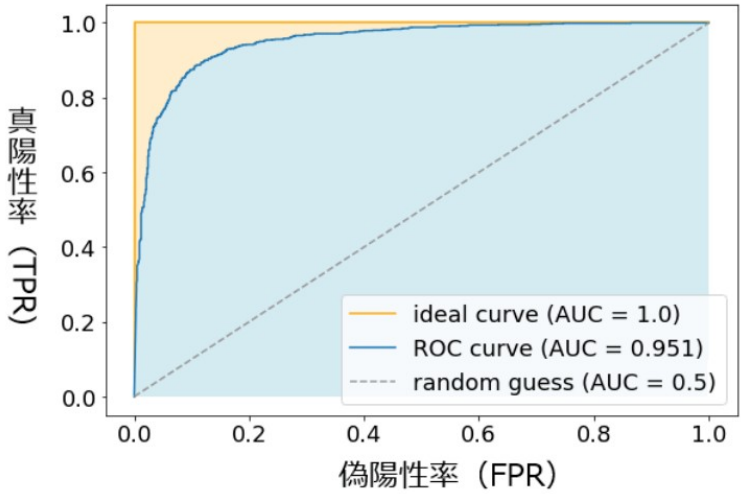
第二回の目標

- 1. コンペティションの流れを理解する
- 2. データをダウンロードする
- **3. タスクを理解する**
- 4. データを理解する
- 5. 一度提出を行う

タスクを理解する

- 患者に関する情報が与えられるので，脳卒中患者かどうかを予測する。
  - ▶ 患者1は脳卒中です。
  - ▶ 患者2は脳卒中ではありません。

- 評価指標：Area under the ROC curve (AUC)
  - ROC曲線とは縦軸に真陽性率（TPR, True Positive Rate, 陽性であるデータのうち，正しく陽性と予測したデータの割合），横軸に偽陽性率（FPR, False Positive Rate, 陰性であるデータのうち，間違えて陽性と予測したデータの割合）をとり，モデルの閾値を変化させたときの真陽性率と偽陽性率の変化をプロットして得られる．
  - AUCは，このROC曲線の下側（図中の青いの部分）の面積として定義されます．



第二回の目標

- 1. コンペティションの流れを理解する
- 2. データをダウンロードする
- 3. タスクを理解する
- **4. データを理解する**
- 5. 一度提出を行う

データについて

- 患者ついて以下の情報が与えられる:
  - id:整数値。1人ごとに割り当てられる固有の数値。
  - gender:文字列。性別を意味し、値は"Male"、"Female"、"Other"のいずれか。
  - age:整数値。年齢。
  - hypertension:整数値。高血圧患者なら1、でなければ0。
  - heart_disease:整数値。心臓病患者なら1、でなければ0。
  - ever_married:文字列。結婚経験があるかを意味し、値は"No"か"Yes"のいずれか
  - work_type:文字列。労働形態を意味し、値は“children”、“Govt_job”、“Never_worked”、"Private"、"Self-employed"のいずれか。
  - Residence_type:文字列。居住地のタイプを意味し、値は"Rural"か"Urban"のいずれか。
  - avg_glucose_level:実数値。平均血糖値。
  - bmi:実数値。ボディマス指数(Body Mass Index)。
  - smoking_status:文字列。喫煙習慣を意味し、値は"formerly smoked"、"never smoked"、"smokes"、"Unknown"のいずれか。
  - stroke:今回予測する値。整数値。脳卒中なら1、でなければ0(訓練データにのみ与えられる)
  - ▶脳卒中患者かどうかを予測(訓練データのみ,テストデータのこの値を予測する)
- 訓練データ数:3,577,テストデータ数:1,533

データについて

- train.csv：訓練データ集合，3,578行12列のCSVファイル
  - 1行目：各列（情報）の名前
  - 2-3,578行目：各行が1つのデータに対応
  - 1-11列目：患者の情報
  - 12列目：脳卒中情報（目標値，正解）
- test.csv：テストデータ集合，1,534行11列のCSVデータ
  - 1行目：各列（情報）の名前
  - 2-1,534行目：各行が1つのデータに対応
  - 1-11列目：患者の情報
  - 12列目は存在しない，12列目を予測する
- y_pred_example.txt：提出の例，1533行のテキスト
  - test.csvの予測された12列目を提出する
  - i行目の値がtest.csvのi + 1行目のデータの予測に対応させる
  - 最後に空白の行があっても良い

第二回の目標

- 1. コンペティションの流れを理解する
- 2. データをダウンロードする
- 3. タスクを理解する
- 4. データを理解する
- **5. 一度提出を行う**

提出

- y_pred_example.txtを提出してみましょう
  - 今日中に自分で予測した結果を5回提出したいという強い意思を持つ方はやらなくても結構です

- こちらからのコンペに関する説明は以上
  - 毎回サンプルプログラムの配布とその説明を行いますが，必ずしも用いる必要はない
  - ただし，毎週少なくとも一回は提出してください

- **1位を目指して頑張ってください**
  - **ただし成績は順位ではなくレポートで評価されます**

おわりに

- ●質問・意見，いつでもお気軽にどうぞ
  - ▶授業に関係あれば「コンペに関する質問」でも「プログラミン グに関する質問」でも「機械学習に関する質問」でも何でもどうぞ（基本的にはTA/TFに）
  - ▶学生同士のディスカッションも大いに推奨します
- ●アドバイス：毎週コツコツやりましょう
  - ▶一日に投稿できる最大数が決まっているので「締め切り前日に死ぬ気で頑張る」作戦はそもそも使えない可能性が高い
  - ▶コンペ締め切りはレポート締め切りの一週間前なのも注意
  - ▶毎回の取り組みでは，「どういう意図で何をしたか」を忘れないように（レポートを書く際に重要）

付録

機械学習・教師あり学習

機械学習

●機械に学習をさせる方法（に関する研究分野）

- ▶機械＝コンピュータ＝プログラム
- ▶学習＝経験から知識（技術）を得て，活用すること
- ▶経験＝データ

↓

●機械学習＝コンピュータにデータを知識に変換させ，それを活用する方法(に関する研究分野）

- ▶より技術・実用的には「データからパターンや知識を抽出し，それを未知のデータの予測や様々なタスクに活用する」技術
- ▶データマイニング・パターン認識・統計学と密接な関係

●一口に学習と言っても，データや知識の活用のさせ方で色々ある

機械学習の主な3つのタイプ

●1. 教師あり（つき）学習

- ▶ 正解のあるデータを利用して，未知のデータの正解を予測できるようにする
- ▶ 例：画像分類，音声認識，機械翻訳

●2. 教師なし学習

- ▶ データの持つパターンや知識を抽出
- ▶ 例：クラスタリング，次元削減（＝データの圧縮・要約）

●3. 強化学習

- ▶ 正解は与えられないが，行動（予測）の良し悪しを評価しうる値が得られるので，その値を元に試行錯誤的に行動を改善
- ▶ 例：ゲームAIの学習（一瞬一瞬の行動に正解はないが，最終的に勝った・負けたで良し悪しを評価できうる）

機械学習はいつ･どこでつかう?

- 前述の3分類だけでも非常に幅広い問題を含んでいる
  - ▶データがあるところではいつでもどこでも機械学習のメスが入りうる!
- 基本的には「明示的にプログラムするのが難しい問題」に対して，十全にデータがある時に使うべき
  - ▶使うべき（と現在は考えられている）例：画像分類

- 人間がどのように分類･判定しているかを**説明することは難しいし**，当然それをプログラムに**落としこむことも極めて困難**
  →機械学習!

機械学習はいつ･どこでつかう?

- 前述の3分類だけでも非常に幅広い問題を含んでいる
  - ▶データがあるところではいつでもどこでも機械学習のメスが入りうる!
- 基本的には「明示的にプログラムするのが難しい問題」に対して, 十全にデータがある時に使うべき
  - ▶使う必要のない例:北大からのメールか否かの分類

- アドレスの@以下(の接尾辞が)hokudai.ac.jpであればよい
  →文字列マッチング, **陽にプログラムできるのでそれで良い**

教師あり学習

●コンペは教師あり学習の問題なので，教師あり学習に ついてもう少し ちゃんと復習する

●1. 教師あり（つき）学習（再掲）

- ▶ 正解のあるデータを利用して，未知のデータの正解を予測できるようにする
- ▶ 例：画像分類，音声認識，機械翻訳

●教師あり学習は「入出力関係を作る」問題と言える

- ▶ **入力** ＝（正解の情報以外の）データ，**出力**（目標値）＝ 正解
- ▶ 入出力関係を作れば**未知の正解のないデータの正解**を予測可能

→数理的には「写像」を作ることと言える

教師あり学習

- 少しだけ数理的に表現すると，以下のように書ける

  - 教師あり学習
    - 入力の集合を$\mathcal{X}$，出力（目標値，正解）の集合を$\mathcal{T}$とする．
    - N個の出力（目標値，正解）のあるデータの集合（訓練集合）$\{(x_1,t_1), (x_2,t_2), \dots, (x_N,t_N)\} \in (X \times T)^N$が与えられる．
    - 訓練集合を用いて，未知のデータ$x \in \mathcal{X}$に対して，対応する出力$t \in T$をできるだけ正確に予測する写像$y: X \to T$を作る．

  - 例：RGB 1280×720の画像に写っているのが人間か否かを予測
    - $X = [0,1, \dots, 255]^{3\times1280\times720}$,T = {人間, 人間でない} = {1, 0 }
  - 例：糖尿病患者の1年後の糖尿病進行度を表す値を予測
    - $X = \mathbb{R}^{10}$（つまり10個の患者の情報を使う），$T = \mathbb{R}$

回帰問題と分類問題

- 教師あり学習は目標値の種類によって主に二種類に分けられる
  - 回帰問題：Tが連続である教師あり学習
  - 分類問題：Tが離散である教師あり学習
- 本演習で扱うコンペは分類問題

教師あり学習の主なプロセス

◆ 教師あり学習
- 入力の集合を$\mathcal{X}$, 出力（目標値, 正解）の集合を$\mathcal{T}$とする.
- N個の出力（目標値, 正解）のあるデータの集合（訓練集合）$\{(x_1,t_1), (x_2,t_2), \ldots, (x_N,t_N)\} \in (X \times T)^N$が与えられる.
- 訓練集合を用いて, 未知のデータ$x \in \mathcal{X}$に対して, 対応する出力$t \in T$をできるだけ正確に予測する写像$y: X \to T$を作る.

…これだけをみて教師あり学習を実現できるだろうか？
- Xは「どのような情報を用いるか」を表しているので, 選択の 余地があるのでは？

…10種類

- 特徴設計
- モデル/アルゴリズムの選択・設計
- モデルの評価

実際のデータ分析ではこれらを試行錯誤する必要あり
- すべてが重要
- 独立でなく密接に関係

教師あり学習の主なプロセス

◆ 教師あり学習
- 入力の集合を$\mathcal{X}$, 出力（目標値, 正解）の集合を$\mathcal{T}$とする.
- N個の出力（目標値, 正解）のあるデータの集合（訓練集合）$\{(x_1,t_1), (x_2,t_2), \ldots, (x_N,t_N)\} \in (X \times T)^N$が与えられる.
- 訓練集合を用いて, 未知のデータ$x \in \mathcal{X}$に対して, 対応する出力$t \in T$をできるだけ正確に予測する写像$y: X \to T$を作る.

…これだけをみて教師あり学習を実現できるだろうか？
- Xは「どのような情報を用いるか」を表しているので, 選択の 余地があるのでは？
  - （例：画像データは生の画像をそのまま使うのか？患者の情報 は10種類で良いのか？10種類でも色々あるのでは？）

  →特徴設計
- どうやって写像を作る？→モデル/アルゴリズムの選択・設計
- 正確性の尺度は？→モデルの評価

教師あり学習の主なプロセス

- 0. データを収集する（コンペでは考えなくてよい）
  - タスクを解くのに必要な情報を考えて集める
  - 例：病状予測をするのに，患者の中学生の時の国語のテストの点数を収集しても（十中八九）意味は無い

- 1. 特徴設計：集めたデータから特徴ベクトルを作る
  - 特徴ベクトル：各入力の数ベクトルとしての表現
  - 多くのモデル（アルゴリズム）が数ベクトルしか扱えないが，集めた情報が数値的な情報とは限らない

    →どのように数値的な情報に変換するか？を考えねばならない
  - 数値的な情報であっても，何らかの変換を行って使った方が良いことも多い
  - 使わない方が良い情報が含まれていることもある
  - 「機械学習アルゴリズムの外」にあるが，最も重要な部分

教師あり学習の主なプロセス

- 2. モデル（アルゴリズム）を選択
  - モデル：予測関数yのこと
    →モデルの選択：どのような形の予測関数yにするかを決める
  - データの形式や大きさによっては使えないモデルもある
    →プロセス0,1と密接に関係
  - 例：線形回帰？ニューラルネットワーク？決定木？SVM？
  - 大枠では同じモデルでも，細かい設定が必要なこともある（例：ニューラルネットワークの中間ユニットの数）

- 3. 学習アルゴリズムを選択，実際に学習する
  - 同じモデルでも学習方法は様々（モデルによっては自明に 定まることもある）
  - データの形式や大きさによっては使えないモデルもある
    →プロセス0,1と密接に関係

教師あり学習の主なプロセス

- 4. モデルの評価
  - 学習したモデルが「どれくらい正確なのか」を評価する
  - 「正確さ」「不正確さ」を表す尺度・指標にも色々ある
  - 例：正解率？決定係数？二乗平均誤差？絶対誤差？
  - 評価結果を以降の試行錯誤に活かす（と良い）
  - 注意1：未知のデータに対する予測の正確さを評価したい
  - 注意2：どのように評価するかは学習前に決めておく（学習後，モデルに合わせて都合よく評価はしない）

- 5. 最終的な予測モデルの選択
  - プロセス1-4（ときには0-4）を様々な条件で繰り返し行い，最終的にどの学習された予測モデルを使うかを決める
  - 基本的には4で評価した「正確さ」が最も良いものを選ぶ

教師あり学習とそのプロセスまとめ

- 教師あり学習：正解の分かっているデータを用いて，未知のデータに対する予測を行えるようにする
  - 訓練集合：正解の分かっているデータの集合
  - 数理的には「入力$\mathscr{X}$から出力$\mathscr{T}$への写像」を作ること
  - モデル＝写像y: $\mathscr{X} \to \mathscr{T}$
  - 回帰問題：$\mathscr{T} = \mathbb{R}$であるような教師あり学習の問題

- タスクが決まってデータを集めた後は以下を繰り返す
  - 1. 特徴設計
  - 2. モデル（アルゴリズム）の選択
  - 3. 学習アルゴリズムを選択し，モデルを学習
  - 4. モデルの評価（+その結果を元に次に活かす）