scikit-learn、Keras、TensorFlow による 実践機械学習 第 2 版 第 1 章 前半

データ工学研究室 輪読会

AL18036 片岡 凪

芝浦工業大学 工学部 情報工学科 3 年

March 17, 2021

発表者紹介

- 片岡 凪
- 千葉県 浦安市
- 芝浦工業大学 工学部 情報工学科 3 年
- データ工学研究室（木村昌臣研究室）
- 関心：画像，XAI，自動化，効率化
- Twitter @calm_IRL
- Github KataokaNagi

目次

目次

使用ソフト

- scikit-learn
  - 多くの効率的なアルゴリズム
- TensorFlow
  - GPU による分散 NN エンジン
  - Google
- Keras
  - NN の単純化 API
  - TensorFlow などと利用

必要な予備知識

- Python
  - Numpy
  - Pandas
  - Matplotlib
- 数学
  - 解析学
  - 線形代数
  - 確率論
  - 統計学

おすすめ教材

- Python チュートリアル
  - LearnPython
  - python.org
- 機械学習
  - Cousera - Andrew Ng 機械学習講座
    - 数か月かかる
  - scikit-learn ユーザーガイド
  - Dataquest
    - 対話的教材
  - Quora
    - Q&A サイト
  - deeplearning.net

配布コード

- Jyoyter ノートブック
- https://github.com/ageron/handson-ml2

目次

第 1 章の目的

- ML の対象や意味範囲
- ML の例
  - スパムフィルタ
  - OCR
  - 商品提案
  - 音声検索

目次

ML の意味

- コンピュータがデータから学習するための科学技術
  - 学習：タスクの測定指標が向上する経験を得ること
- 各種語義
  - 訓練セット
    - 学習用のデータ例
  - 訓練インスタンス，標本
    - 個々のデータ例
  - 訓練データ
    - 経験
  - 正解率
    - 性能指標

目次

ML の概要

- ML の手順
  1. 一般的な特徴を分析
  2. 特徴をもとに検出アルゴリズムを作成
  3. プログラムをテストし、実用レベルになるまで上を繰り返す
- ML の利点
  - データごとのアルゴリズムが不要
  - 複雑な問題の解決
  - 既知のアルゴリズムがない問題の解決
  - データマイニング
    - 特徴の予想外な相関関係，トレンドの発見
    - 高速に

目次

ML の例 1/6

- 製品の自動分類
  - イメージ分類
    - CNN（Convolutional：畳み込み）などを利用
- 脳腫瘍の検出
  - セマンティックセグメンテーション
    - CNN など
- 記事の自動分類
  - テキスト分類 ∈ NLP（自然言語処理）
    - RNN（Recurrent：再帰型）
    - CNN
    - Transformer

ML の例 2/6

- 不適切発言へのフラグ付加
  - テキスト分類
- 自動要約
  - テキスト自動要約 ∈ NLP
- チャットボット、パーソナルアシスタント
  - NLU（自然言語理解）
  - Q&A モジュール

ML の例 3/6

- 次年度収益の予測
  - 回帰タスク（値の予測）
    - 線形回帰
    - 多項式回帰モデル
    - SVM 回帰
    - ランダムフォレスト回帰
    - 人工 NN
  - 過去の業績指標の利用
    - RNN
    - CNN
    - Transformer

ML の例 4/6

- 音声コマンド
  - オーディオサンプルの処理
    - 長くて複雑
    - RNN
    - CNN
    - Transformer
- クレカ詐欺の検知
  - 異常検知

ML の例 5/6

- 購入履歴による顧客分類、販売戦略
  - クラスタリング
- 高次元データセットの図示
  - データの可視化
  - 次元削除
- 購入履歴から商品提案
  - 推薦システム
    - 人工 NN など

ML の例 6/6

- ゲームのインテリジェントボット
  - アクションの選択
    - RL（Reinforced L：強化学習）
    - 全時間の報酬の最大化
    - AlphaGo

目次

ML の分類 - 概要

- ML の分類
  - 人間の関与の有無
    - 教師あり
    - 教師なし
    - 半教師あり
    - 強化学習
  - その場で少しずつ学習可能か
    - オンライン学習
    - バッチ学習
  - 既知のデータポイントから予測モデルを構築するか
    - インスタンスベース学習
    - モデルベース学習
- 上記分類を組み合わせる

1.4.1 教師あり学習 1/2

- 訓練データに正解のラベル
- 応用先
  - クラス分類
  - 回帰（regression）
    - 予測子 (predictor)（＝一連の特徴量）からターゲットの数値を予測すること
    - 語源：背の高い両親の子供が平均に帰して背が低くなる傾向にあることを統計的に示した研究より
    - 出力が複数になる問題も
  - 回帰に分類が使えるものも
  - 分類に回帰が使えるものも
    - 例：ロジスティック回帰で分類確率を導出

1.4.1 教師あり学習 2/2

- 代表的なアルゴリズム
  - k 近傍法
  - 線形回帰
  - ロジスティック回帰
  - SVM
  - 決定木，ランダムフォレスト
  - NN

1.4.1 教師なし学習 - 概要

- ラベルはない
- 代表的なアルゴリズム
  - クラスタリング
  - 異常検知、新規性検知（anomaly detection, novelty detection）
  - 可視化、次元削減（visualization, dimension reduction）
  - 相関ルール学習（association rule learning）

1.4.1 教師なし学習 - クラスタリング

- k 平均法
- DBSCAN
- 階層的クラスタ分析（HCA：Hierarchial Clustering Analysis？）
  - 分類の粒度が可変

1.4.1 教師なし学習 - 異常検知、新規性検知

- 異常検知
  - 正常なインスタンスで訓練
  - 想定内の外れ値を検知
- 新規性検知
  - ML 以外のアルゴリズムで検知できない、分類済みだと思われるインスタンスで訓練
  - 予想外の外れ値を検知
- 例
  - 1 クラス SVM
  - アイソレーションフォレスト

1.4.1 教師なし学習 - 可視化、次元削減 1/2

- 構造を保ちつつ、セマンティッククラスタを可視化
  - 片岡解釈：クラスタごとの座標などを保ちつつ、意味のある群の可視化用ラベル（次元）を増やす
  - 可視化例：クラスタ同士の重なり具合
- 特徴量抽出（Feature Extraction）
  - 情報量を保ちつつ相関する複数の特徴をまとめ（次元圧縮）、データを見やすくする
  - 種々の ML の前処理に利用
    - 速度向上
    - 計算資源の節約
    - 性能が向上することも

1.4.1 教師なし学習 - 可視化、次元削減 2/2

- 例
  - PCA（Principal Component Analysis：主成分分析）
  - カーネル PCA
  - LLE（Locally-Linear Embedding：局所線形埋め込み法）
  - t-SNE（t-distributed Stochastic Neighbor Embedding：t 分布確率的近傍埋め込み法）

1.4.1 教師なし学習 - 相関ルール学習

- 大量のデータの属性同士の興味深い関係を導く
  - 例：オムツとビール
- 例
  - ア・プリオリ
  - Eclat

1.4.1 半教師あり学習（semisupervised learning）

- 一部にラベル
- ラベルの有無に応じた重みづけ？
- 一部にラベリングして全体にラベリングすることも
- 教師あり・なしの同時使用が多い
- 例
  - DBN（deep brief network）
    - 教師なしの RBM（restricted Boltzmann machines：制限付きボルツマン
    - マシン）
    - 教師ありで微調整

1.4.1 強化学習

- 流れ
  1. エージェント（学習システム）が
  2. 環境を観察し
  3. 行動を選択して実行し
  4. 報酬 or ペナルティを得る。
  5. 報酬の高い方策（policy）を学習する。
  6. 学習した方策に従い、特別な行動を決定う
- 例
  - ロボットの歩行
  - AlphaGo
    - 自分自身とも対局

1.4.2 バッチ学習（オフライン学習）

- 事前に全ての訓練データで学習
- オフラインで大量の時間と計算資源を割く
  - 流動的なデータには不向き
  - 分割して学習できないために計算資源に余裕が必要

1.4.2 オンライン学習（差分学習 incremental L.）

- ミニバッチで段階的に訓練
- 使用済みデータの保持が不要
  - アウトオブコア（主記憶より大容量の学習システム）でも使用可能
    - 通常オフラインで使用
- 学習速度（L. rate）が重要
  - 速いと古いデータを忘れる
  - 遅いと外れ値に強くなる
    - 実用上、速さは大事？
- 性能が低下するデータの学習は打ち止める
  - 異常検知などを併用

1.4.3 インスタンスベース学習とモデルベース学習

- 汎化（generalize）による ML の分類
  - 新しいデータの予測のためのアプローチ

1.4.3 インスタンスベース学習

- 既存データの丸暗記
- 新しいデータに類似度の尺度（measure of similarity）を適用
- 例
  - k 近傍法

1.4.3 モデルベース学習

- 線形関数、超平面などにモデリング
  - モデルパラメータ θ_i（慣例）による関数 $f(\theta_i)$（モデル）を定義
    - モデルを評価する関数を定義
    - 良さを示す適応度関数 $g \circ f(\theta_i)$？（utility f.）
    - 悪さを示すコスト関数 $h \circ f(\theta_i)$？
    - θ_i を変化させて g or h を大きく or 小さくさせるよにに訓練（not 学習？）
- 3 種類の「モデル」に注意
  - 線形回帰などのモデル
  - 線形回帰などから構築したモデルアーキテクチャ
  - 訓練済みのモデル

1.4.3 線形モデルで GDP と暮らし満足度の予測を行うコード

- https://github.com/ageron/handson-ml2/blob/master/01_the_machine_learning_landscape.ipynb
- モデルベースもインスタンスベースも近い値を取る
- 必要操作
  - データの検討
  - モデルの選択
  - コスト関数による訓練
  - 新しいデータで推論（inference）

目次

問 1 ML の定義

- 片岡の解答
- テキストの解答

問 1 ML の定義

- 片岡の解答
  - コンピュータが、データからタスクの測定指標が向上する経験を得るための科学技術
- テキストの解答
  -

問 1 ML の定義

- 片岡の解答
  - コンピュータが、データからタスクの測定指標が向上する経験を得るための科学技術
- テキストの解答
  - データから学習できるシステムをつくること
    - 学習：何らかの測定手段に基づき、あるタスクを処理した成績が上がる操作

問 2 ML が発揮する問題の 4 タイプ

- 片岡の解答
  - –
  - –
  - –
  - –
- テキストの解答
  - –
  - –
  - –
  - –

問 2 ML が発揮する問題の 4 タイプ

- 片岡の解答
  - データごとのアルゴリズム実装が大変な問題
  - 複雑な問題
  - 既知のアルゴリズムがない問題
  - 予想外な相関関係とトレンドを抽出する問題（データマイニング）
- テキストの解答
  -
  -
  -
  -

問 2 ML が発揮する問題の 4 タイプ

- 片岡の解答
  - データごとのアルゴリズム実装が大変な問題
  - 複雑な問題
  - 既知のアルゴリズムがない問題
  - 予想外な相関関係とトレンドを抽出する問題（データマイニング）
- テキストの解答
  - アルゴリズムを使ったソリューションがない複雑な問題の解決
  - 思い付きの規則が延々と続くものに代わるモジュールの開発
  - 変動する環境に合わせて自分を修正できるシステムの開発
  - 人間の学習の支援（データマイニングなど）

問 3 ラベル付き訓練セットとは

- 片岡の解答
  -
- テキストの解答
  -

問 3 ラベル付き訓練セットとは

- 片岡の解答
  - 人間による分類を行った全ての訓練データ
- テキストの解答
  -

問 3 ラベル付き訓練セットとは

- 片岡の解答
  - 人間による分類を行った全ての訓練データ
- テキストの解答
  - 個々のインスタンスに問題の答えが含まれている訓練セット

問 4 教師あり学習の応用例 2 つ

- 片岡の解答
  -
  -
- テキストの解答
  -
  -

問 4 教師あり学習の応用例 2 つ

- 片岡の解答
  - 回帰
  - クラス分類
- テキストの解答
  -
  -

問 4 教師あり学習の応用例 2 つ

- 片岡の解答
  - 回帰
  - クラス分類
- テキストの解答
  - 回帰
  - 分類

問 5 教師なし学習の応用例 4 つ

- 片岡の解答
  -
  -
  -
  -
- テキストの解答
  -

問 5 教師なし学習の応用例 4 つ

- 片岡の解答
  - クラスタリング
  - 異常検知、新規性検知
  - 可視化、次元削減
  - 相関ルール学習
- テキストの解答
  -

問 5 教師なし学習の応用例 4 つ

- 片岡の解答
  - クラスタリング
  - 異常検知、新規性検知
  - 可視化、次元削減
  - 相関ルール学習
- テキストの解答
  - 同上

問 6 未知の領域を探索する、ロボットで使える ML は

- 片岡の解答
  -
- テキストの解答
  -
  -

問 6 未知の領域を探索する、ロボットで使える ML は

- 片岡の解答
  - 強化学習
- テキストの解答
  -
  -

問 6 未知の領域を探索する、ロボットで使える ML は

- 片岡の解答
  - 強化学習
- テキストの解答
  - 強化学習
  - 教師あり・なし学習だと不自然になる

問 7 顧客を分類する ML は

- 片岡の解答
  - –
- テキストの解答
  - 集団の定義が分からない場合
    - クラスタリング（教師なし学習）
  - –
    -

問 7 顧客を分類する ML は

- 片岡の解答
  - クラスタリング（教師なし学習）
- テキストの解答
  - 集団の定義が分からない場合
    - クラスタリング（教師なし学習）
  -

問 7 顧客を分類する ML は

- 片岡の解答
  - クラスタリング（教師なし学習）
- テキストの解答
  - 集団の定義が分からない場合
    - クラスタリング（教師なし学習）
  - 集団の定義が分かっている場合
    - 分類アルゴリズム（教師あり学習）

問 8 スパム検出は教師あり or なしか

- 片岡の解答
  - ■
- テキストの解答
  - ■

問 8 スパム検出は教師あり or なしか

- 片岡の解答
  - 教師あり
    - スパムの特徴は限られており、アルゴリズムを適用する必要がありそう
- テキストの解答
  -

問 8 スパム検出は教師あり or なしか

- 片岡の解答
  - 教師あり
    - スパムの特徴は限られており、アルゴリズムを適用する必要がありそう
- テキストの解答
  - 教師あり
    - ラベル（スパム or ハム）を付けたメールでアルゴリズム訓練

問 9 オンライン学習システムとは

- 片岡の解答
  -
- テキストの解答
  -
  -
  -
  -

問 9 オンライン学習システムとは

- 片岡の解答
  - ミニバッチで段階的に訓練する、計算資源の要求が比較的少ない学習
- テキストの解答
  -
  -
  -
  -

問 9 オンライン学習システムとは

- 片岡の解答
  - ミニバッチで段階的に訓練する、計算資源の要求が比較的少ない学習
- テキストの解答
  - バッチ学習システムと異なり、差分データで学習可能
  - データが変化するシステムや自律的なシステムに適用可能
  - 機敏に学習可能
  - 極端に大規模なデータを使用可能

問 10 アウトオブコア学習とは

- 片岡の解答
  -
- テキストの解答
  -
  -

問 10 アウトオブコア学習とは

- 片岡の解答
  - 主記憶より大容量のデータを扱う学習システム
- テキストの解答
  -
  -

問 10 アウトオブコア学習とは

- 片岡の解答
  - 主記憶より大容量のデータを扱う学習システム
- テキストの解答
  - 同上
  - データをミニバッチに分割し、オンライン学習のテクニックを使って学習

問 11 類似度の尺度を用いる学習は

- 片岡の解答
  -
- テキストの解答
  -
  -

問 11 類似度の尺度を用いる学習は

- 片岡の解答
  - インスタンスベース学習
- テキストの解答
  -
  -

問 11 類似度の尺度を用いる学習は

- 片岡の解答
  - インスタンスベース学習
- テキストの解答
  - 同上
  - 訓練データを丸暗記させた上で新しいインスタンスを与え、類似度の尺度から暗記したインスタンスに最も近いものを採択