【TechGym】ゼロからはじめる機械学習入門講座「Kaggle に挑戦」（テックジムオープン講座）
Numpy,Pandas ライブラリを使って機械学習のデータ処理を体験し、線形回帰の問題を解いてみよう。
ラストは学んだことを活かして、Kaggle に挑戦してみよう。

[サンプルソースの公開場所] https://github.com/techgymjp/techgym_ai
・実行環境がない場合は anaconda を install してください。

■0-A:numpy ライブラリを使う: q3Mt.py
【問題】配列[9, 2, 3, 4, 10, 6, 7, 8, 1, 5]を作成して以下を実行しましょう。
□配列、次元数、要素数の表示します。
□配列*配列(掛け算)、配列を 3 乗、配列/2(割り算)の表示をします。
□昇順、降順でソートした配列の表示をします。
□最小値、最大値、合計、積み上げ合計、積み上げ合計/合計を表示します。
■解答は Tz6s.py

■0-D:matplotlib ライブラリを使う: Kf74.py
【問題】100 個の乱数を発生させた座標と y=sin(x)関数をグラフに表示しましょう。
■解答は N6pB.py

■0-E:ヒストグラム: Xb9t.py
【問題】10 万個の乱数を発生させてヒストグラムを表示しましょう。
■解答は mY8e.py

■0-2:データフレーム: Jp9q.py
【問題】データフレームを表示してさらに、転置して表示しましょう。
■解答は X5gR.py(～22 行目まで)

■0-3:データフレーム: g5Hw.py
【問題】性別が男性の行のみを表示しましょう。
■解答は zP8u.py(17 行目)

■0-4:データフレーム: Gu4t.py
【問題】男女別に勝ちの平均回数、勝ちの最大値、勝ちの最小値を表示しましょう
■解答は Ck8N.py(～18 行目まで)

■0-9A:データフレーム: Y8c2.py
【問題】各年齢で住所地別にじゃんけんの平均勝ち回数を表示しましょう。
■解答は t2Jv.py(28 行目)

■0-11:データフレーム: C2kj.py
【問題】csv ファイルへの書き込み、読み込みをしましょう。
■解答は dJ3a.py

■0-15:データ分析: Xh86.py
【問題】以下の操作をしてみてください。
□以下のデータをデータフレームとして読み込みます。
http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data
□以下のデータの説明を表示して、データに index をつけます。
http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.names
□データの個数や型を表示する(必要であればデータ全体を表示してみましょう)
□Alcohol のヒストクラムを表示しましょう。
□Alcohol 要約統計量を表示しましょう。
□['Alcohol', 'Malic_acid', 'Ash', 'Total_phenols', 'Color_intensity']のそれぞれのデータの散布図をプロットしてデータを可視化してみましょう
■解答は T7nf.py

■0-16:データ分析:Sk4a.py
【問題】前と同じデータを使用して以下の操作をしましょう。
http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data

□説明変数を'Alcohol'で` 目的変数'Color_intensity'として、 Alcohol(アルコール度数)と Color_intensity(色の濃さ)の単回帰分析のモデルを作りましょう。
□回帰係数と切片を表示しましょう。
□'Alcohol'と'Color_intensity'の散布図を表示しましょう
□散布図の上に、計算した回帰係数と切片が示す直線を上書きして表示しましょう
□モデルの決定係数を表示しましょう。
(決定係数は 1 に近ければ良いモデルで、予測した値が実際の値に近くなります)
■解答は uU9Y.py

【チャレンジ課題】時間が余った人は挑戦してみましょう!!!
Kaggle にアカウントがない方は https://www.kaggle.com に sign in しましょう。

■0-21:データ分析:D4qt.py
【問題】Kaggle にチャレンジします。以下に住宅価格を予測するコンペティションがあります。
https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
□訓練用データ(train.csv)とテスト用データ(test.csv)を読み込みましょう
□データ型や大きさを確認して、さらに欠損値の状態を確認します。
□単回帰分析モデルを作成します。(OverallQual を説明変数、SalePrice を目的変数として)
□回帰係数と切片を表示します。
□テスト用データを使用して、SalePrice を予測します。
□kaggle の提出用で0田として、Id と SalePrice のみのデータをつくり、submission.csv という名前で保存する。
□kaggle のサイトに submit してスコアを表示させます。
■解答は d4XR.py