知的システムII
データ解析　第2回

データ解析に関する技術の初歩

李　吉屹

データ解析に関する技術

- 多種多様な技術
  - 機械学習、データマイニング、コンピュータビジョン、自然言語処理、可視化等
  - 五回の演習では全然足りない！本演習では初歩的事項を紹介して、興味があったら、今後各自で勉強してください
  - データ解析プラットフォーム例
    - Kaggle (英語)
    - Signate (日本語)

- 本演習で扱う基礎技術
  - 分類
  - アンサンブル学習
  - 探索的データ解析
  - 回帰
  - クラスタリング

分類

- データを複数のクラス（グループ）に分類すること

- 異なる分類アルゴリズムの比較
  - https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

分類

- データセット
  - Breast Cancer (乳がん)
    - https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html
  - データ数: 569
  - 特徴: 30
  - 分類: 2
- コード例
  - da2_classification_ensemble.ipynbを参考してください

分類結果

方法	SVM (Default Hyperparameter)	SVM (kernel="linear", C=0.025)
精度	0.6418	**0.9403**

- [練習]
  - 他の分類器とハイパーパラメータを選んで、精度向上を試みる
  - [Option] 交差検証を調査して、使ってください

アンサンブル学習

方法	SVM (Default Hyperparameter)	SVM (kernel="linear", C=0.025)	Ensemble (Random Forest, Default Hyperparameter)
精度	0.6418	0.9403	**0.9478**

- アンサンブル学習
  - 単一の学習器ではなく、複数の学習器の結果を融合して汎化性能を高める
  - アンサンブル方法例
    - Bagging (例 : Random Forest等)
    - Boosting (例 : Adaboost等)

探索的データ解析

- 探索的データ解析 (Exploratory Data Analysis: EDA)は、統計を用いて図を作成することで、データが何を伝えているかを理解するプロセスです
- データセット
  - Boston house prices (ボストン住宅価格) データセット
    - https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html
- コード例
  - da2_eda_regression.ipynbを参考してください

回帰

- Y が連続値の時にデータに Y = f(X) というモデルを当てはめる事
  - 参考 : https://ja.wikipedia.org/wiki/%E5%9B%9E%E5%B8%B0%E5%88%86%E6%9E%90
- データセット
  - Boston house prices (ボストン住宅価格) データセット
    - https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html
  - データ数: 506
  - 特徴X: 13
  - 目標Y: MEDV (Median value of owner-occupied homes in 1000's)
- コード例
  - da2_eda_regression.ipynbを参考してください

回帰結果

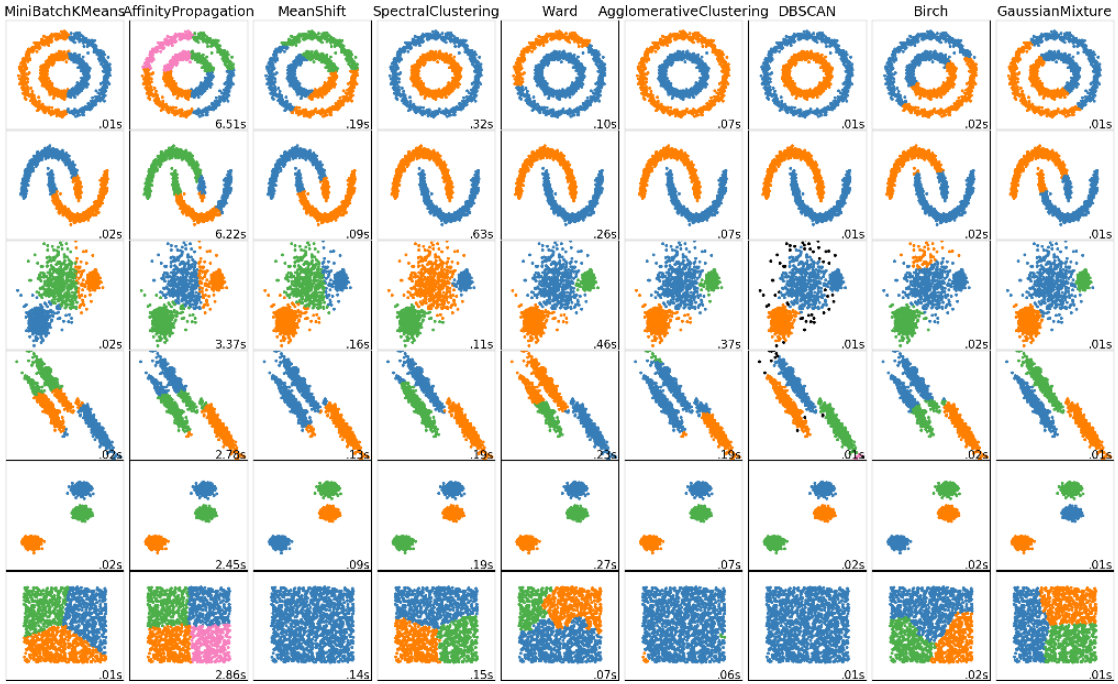
方法	Linear Regression (Ridge, Default Hyperparameter)	SVR (Default Hyperparameter)
評価指標: RMSE	**5.4930**	8.8628

RMSE: Root Mean Square Error

- [練習]
  - **他の回帰方法とハイパーパラメータを選んで、結果の向上を試みる**
  - **[Option]** Ridge回帰の特性（ヒント：目的関数、正則化等）を調査してください

クラスタリング

- 教師なし
- 異なるクラスタリングアルゴリズムの比較



参考文献：https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

クラスタリング

- [練習]
  - 下記URLのコードを実行して、クラスタリングアルゴリズムの特性を了解する
  - https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

3回目~5回目

- 3回目~4回目
  - データ解析コンテスト練習
  - 時間が少ないため、早めに始めたほうが良いです
- 5回目
  - レポート報告会

Special thanks:

- TA: 三島　大進
- TA: 小宮山　亮太

誤字・脱字などを見つけたら

- 口頭でもメールでも構いませんので、李まで連絡して下さい。