Exercise 2

Christof Pegrisch, Simon König, Philipp Eberl

December 18, 2021

Table of Contents

Dataset 1: Solar flares

- Devided into two sections, one with more error correction
- 10 Attributes
- 1389 Instances
- No missing values
- Instances describe state of certain region of the sun
- `https://archive-beta.ics.uci.edu/ml/datasets/solar+flare`

Dataset 2: Wine quality

- Two seperate files for red and white wine
- Information about wine like citric acid or residual sugar
- 12 attributes each
- 4898 instances for white wine
- 1599 instances for red wine
- No missing values
- `https://www.kaggle.com/brendan45774/wine-quality?select=winequality-red.csv`

Dataset 3: Coronavirus

- Dataset with population and vaccination data of countries
- 10 attributes
- About 24.000 instances with different timestamps
- No missing values
- `https://www.kaggle.com/sinakaraji/covid-vaccination-vs-death/activity`

Table of Contents

What has been done

- Solar flares
  - Converting to numeric (requirement of used implementations)
  - One hot encoding (for region class, largest spot and spot distribution)
- Wine quality
  - Converting to numeric
- Coronavirus
  - Converting to numeric
  - Normalizaton by dividing through population (eg ratio of people vaccinated)
  - One hot encoding (for country names)

Table of Contents

General Idea

- Implementation of both hillclimbing and simulated annealing due to their similarities.
- Simulated annealing for:
  - Solar flares
  - Wine white
  - Wine red
- Hillclimbing for:
  - Covid
    (due to the size of the dataset and shorter runtime of hillclimbing)

Used Algorithms

- Hillclimbing (finds local maximum by searching the immediate neighborhood)
  - Starts with a random set of hyperparameters
  - Searches close neighborhood for better solution and keeps it
  - Repeats until no solution in the neighborhood is better than the current
- Simulated annealing (tries to break out of local minima using probabilities)
  - Starts with a random set of hyperparameters
  - Takes random solution from the close neighborhood
  - Keeps the new solution according to a certain probabilty dependent on the quality and a decreasing temperature value
  - Repeats until either no changes occure for a certain amount of epochs or until a maximum number of epochs is reached

Chosen Regressors

- Linear SVR
  https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVR.html#sklearn.svm.LinearSVR
- K-Neighbours Regressor
  https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html
- DecisionTree Regressor
  https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

Table of Contents

Linear SVR

Super vector regression using a linear kernel. By allowing an error in the model a support vector regressor finds the optimal hyperplane to fit the data.

- Strengths: Effective for high dimensional feature spaces, memory efficient
- Weaknesses: Not suitable for large datasets, hard to understand paramter selection (Black Box)

K-Neighbours Regressor

Uses K-nearest-neighbours algorithm to perform classification. To make a prediction the regressor takes the mean of k nearest neighbours.

- Strengths: Effective for large datasets, robust to noisy data, fast model building (Lazy learner)
- Weaknesses: Prediction is computationally costly, hard to determine which attributes contribute to regression

DecisionTree Regressor

Builds a decision tree with real values as leave nodes. In the model building process mean squared error is used to split node in sub-nodes.

- Strengths: Easy to interpret and visualise, little influence by outliers, useful in data exploration
- Weaknesses: Cannot extrapolate, tends to overfit

Table of Contents

Sklearn metrics

- Negative Mean squared error
  - Describes how close a regression line is to a specific set of points. Lower scores indicate a better fit
- R^2 score
  - R-squared explains to what extent the variance of one variable explains the variance of the second variable. Best result is 1.0, which means all of the variation can be explained by the models inputs.

Table of Contents

Linear SVR

- Covid dataset
  - Result negative MSE: -1822.8864578877833
  - Hyperparameters: (loss="epsilon_insensitive", tol=1e-5, C=10, epsilon=1e-3)
  - Result R^2: -1.6148099403159493
  - Hyperparameters: (loss="epsilon_insensitive", tol=1e-5, C=1e-4, epsilon=1)
- Red wine dataset
  - Result negative MSE: -0.651999565611231
  - Hyperparameters: (loss="epsilon_insensitive", tol=1e-5, C=1e-4, epsilon=1e-3)
  - Result R^2: 0.19967708634304884
  - Hyperparameters: (loss="epsilon_insensitive", tol=1e-3, C=1e-2, epsilon=1)

Linear SVR

- White wine dataset
  - Result negative MSE: -0.9578209077406313
  - Hyperparameters: (loss="epsilon_insensitive", tol=1e-5, C=1, epsilon=1)
  - Result R^2: 0.1956095656824933
  - Hyperparameters: (loss="epsilon_insensitive", tol=1e-5, C=1e-2, epsilon=1e-3)
- Solar flares dataset
  - Result negative MSE: -0.5772637203438242
  - Hyperparameters: (loss="epsilon_insensitive", tol=1e-1, C=10, epsilon=1e-1)
  - Result R^2: 0.01058595926564096 5
  - Hyperparameters: (loss="epsilon_insensitive", tol=1e-5, C=1e-2, epsilon=1e-1)

K-Neighbours Regressor

- Covid dataset
  - Result negative MSE: -528.5757039291524
  - Hyperparameters: (n_neighbors=4, weights="uniform", p:2)
  - Result R^2: 0.23799133614878842
  - Hyperparameters: (n_neighbors=10, weights="distance", p=1)
- Red wine dataset
  - Result negative MSE: -0.5745512186731103
  - Hyperparameters: (n_neighbors=10, weights="distance", p=1)
  - Result R^2: 0.02941273213008872 5
  - Hyperparameters: (n_neighbors=6, weights="distance", p=1)

K-Neighbours Regressor

- White wine dataset
  - Result negative MSE: -0.7485020711918213
  - Hyperparameters: (n_neighbors=7, weights="uniform", p=1)
  - Result R^2: 0.0187020649701315 6
  - Hyperparameters: (n_neighbors=6, weights="uniform", p=1)
- Solar flares dataset
  - Result negative MSE: -0.614328428953588
  - Hyperparameters: (n_neighbors=4, weights="uniform", p=2)
  - Result R^2: -0.06532347561422938
  - Hyperparameters: (n_neighbors=9, weights="uniform", p=1)

Decision Tree Regressor

- Covid dataset
  - Result negative MSE: -47.644739644785794
  - Hyperparameters: (max_depth=2, min_samples_split=5, min_samples_leaf=6)
  - Result R^2: 0.9914236829965463
  - Hyperparameters: (max_depth=4, min_samples_split=10, min_samples_leaf=2)
- Red wine dataset
  - Result negative MSE: -0.4948440103954347
  - Hyperparameters: (max_depth=6, min_samples_split=10, min_samples_leaf=10)
  - Result R^2: 0.2015905613389884
  - Hyperparameters: (max_depth=6, min_samples_split=10, min_samples_leaf=10)

Decision Tree Regressor

- White wine dataset
  - Result negative MSE: $-0.586052946055647 3$
  - Hyperparameters: (max_depth=5, min_samples_split=7, min_samples_leaf=4)
  - Result R^2: 0.240256967776180 2
  - Hyperparameters: (max_depth=4, min_samples_split=2, min_samples_leaf=10)
- Solar flares dataset
  - Result negative MSE: -0.576831949429341
  - Hyperparameters: (max_depth=9, min_samples_split=6, min_samples_leaf=10)
  - Result R^2: $-0.200573172738794 7$
  - Hyperparameters: (max_depth=4, min_samples_split=3, min_samples_leaf=9)

Table of Contents

Linear SVR

- Covid dataset
  - Result negative MSE: -676.798302552437
  - Hyperparameters: (loss="squared_epsilon_insensitive", tol=1e-05, C=1, epsilon=1)
  - Result R^2: 0.0330304336872527 6
  - Hyperparameters: (loss="squared_epsilon_insensitive", tol=1e-05, C=20, epsilon=1)
- Red wine dataset
  - Result negative MSE: -0.4336334344973999
  - Hyperparameters: (loss="squared_epsilon_insensitive", tol=1e-05, C=20, epsilon=0.001)
  - Result R^2: 0.326075686205155 8
  - Hyperparameters: (loss="squared_epsilon_insensitive", tol=1e-05, C=1, epsilon=0.001)

Linear SVR

- White wine dataset
  - Result negative MSE: -0.578150960059422
  - Hyperparameters: (loss="squared_epsilon_insensitive", tol=1e-05, C=1, epsilon=0.1)
  - Result R^2: 0.27084918178993517
  - Hyperparameters: (loss="squared_epsilon_insensitive", tol=1e-05, C=20, epsilon=0.001)
- Solar flares dataset
  - Result negative MSE: -0.5342683423739354
  - Hyperparameters: (loss="squared_epsilon_insensitive", tol=0.1, C=0.01, epsilon=0.001)
  - Result R^2: 0.15284999357211257
  - Hyperparameters: (loss="squared_epsilon_insensitive", tol=0.001, C=0.01, epsilon=0.001)

K-Neighbours Regressor

- Covid dataset
  - Result negative MSE: -183.02381848495688
  - Hyperparameters: (n_neighbors=2, weights="distance", p=2)
  - Result R^2: 0.738431897813121 7
  - Hyperparameters: (n_neighbors=2, weights="distance", p=2)
- Red wine dataset
  - Result negative MSE: -0.4736539716577819
  - Hyperparameters: (n_neighbors=9, weights="distance", p=1)
  - Result R^2: 0.2629190075467846 6
  - Hyperparameters: (n_neighbors=9, weights="distance", p=1)

K-Neighbours Regressor

- White wine dataset
  - Result negative MSE: -0.5383166244166239
  - Hyperparameters: (n_neighbors=9, weights="distance", p=1)
  - Result R^2: 0.321785064120657
  - Hyperparameters: (n_neighbors=9, weights="distance", p=1)
- Solar flares dataset
  - Result negative MSE: -0.5606531262445241
  - Hyperparameters: (n_neighbors=9, weights="uniform", p=1)
  - Result R^2: 0.1178206103915730 6
  - Hyperparameters: (n_neighbors=9, weights="uniform", p=1)

Decision Tree Regressor

- Covid dataset
  - Result negative MSE: -15.57304409288794
  - Hyperparameters: (max_depth=9, min_samples_split=8, min_samples_leaf=4)
  - Result R^2: 0.973973891378282
  - Hyperparameters: (max_depth=9, min_samples_split=9, min_samples_leaf=8)
- Red wine dataset
  - Result negative MSE: $-0.4501162754376667 5$
  - Hyperparameters: (max_depth=4, min_samples_split=4, min_samples_leaf=8)
  - Result R^2: 0.2996410728233958 6
  - Hyperparameters: (max_depth=4, min_samples_split=8, min_samples_leaf=9)

Decision Tree Regressor

- White wine dataset
  - Result negative MSE: -0.5801390660131089
  - Hyperparameters: (max_depth=5, min_samples_split=2, min_samples_leaf=8)
  - Result R^2: 0.2673631830667058
  - Hyperparameters: (max_depth=4, min_samples_split=4, min_samples_leaf=8)
- Solar flares dataset
  - Result negative MSE: -0.5529494224260217
  - Hyperparameters: (max_depth=2, min_samples_split=8, min_samples_leaf=7)
  - Result R^2: 0.1255498898193674
  - Hyperparameters: (max_depth=1, min_samples_split=9, min_samples_leaf=7)

Table of Contents

Linear SVR

- Covid dataset
  - Result negative MSE: -1343.9272236143343
  - Result R^2: -0.975865899654315
  - Hyperparameters: ('learner': SVR(C=6.1443024099986 4e-05, cache_size=512, degree=1, epsilon=471.85790078465106, gamma='auto', kernel='linear', max_iter=84056886.0, tol=0.0001977604339369954 7), 'preprocs': (), 'ex_preprocs': ())
- Red wine dataset
  - Result negative MSE: $-0.4738285409084098 7$
  - Result R^2: 0.2329833807017352
  - Hyperparameters: ('learner': SVR(C=0.021658668215710258, cache_size=512, degree=1, epsilon=0.576710701324066 8, gamma='auto', kernel='linear', max_iter=814470115.0, tol=4.057465654030927e-05), 'preprocs': (), 'ex_preprocs': ())

Linear SVR

- White wine dataset
  - Result negative MSE: -0.7742476360259494
  - Result R^2: 0.0030735412517071347
  - Hyperparameters: ('learner': SVR(C=9.231350836902164e-05, cache_size=512, degree=1, epsilon=0.00352626125030543 2, gamma='auto', kernel='linear', max_iter=450281484.0, tol=0.007903935921111719), 'preprocs': (), 'ex_preprocs': ())
- Solar flares dataset
  - Result negative MSE: $-0.620944537736228 7$
  - Result R^2: $-0.2125001034297916 7$
  - Hyperparameters: ('learner': SVR(C=0.011018759854312025, cache_size=512, degree=1, epsilon=0.472228604211841 3, gamma='auto', kernel='linear', max_iter=422105859.0, shrinking=False, tol=4.828809035580296e-05), 'preprocs': (), 'ex_preprocs': ())

K-Neighbours Regressor

- Covid dataset
  - Result negative MSE: -447.24793185054614
  - Result R^2: 0.3590363302160692 3
  - Hyperparameters:('learner': KNeighborsRegressor(metric='euclidean', n_jobs=1, n_neighbors=23), 'preprocs': (), 'ex_preprocs': ())
- Red wine dataset
  - Result negative MSE: $-0.546110501142053 7$
  - Result R^2: 0.12185830259391994
  - Hyperparameters: ('learner': KNeighborsRegressor(metric='manhattan', n_jobs=1, n_neighbors=22, p=1, weights='distance'), 'preprocs': (), 'ex_preprocs': ())

K-Neighbours Regressor

- White wine dataset
  - Result negative MSE: −0.6911960646052259
  - Result R^2: 0.108728621985633 9
  - Hyperparameters: ('learner': KNeighborsRegressor(metric='euclidean', n_jobs=1, n_neighbors=34, weights='distance'), 'preprocs': (), 'ex_preprocs': ())
- Solar flares dataset
  - Result negative MSE: −0.5189269288987464
  - Result R^2: 0.02340265519013567
  - Hyperparameters: ('learner': KNeighborsRegressor(metric='euclidean', n_jobs=1, n_neighbors=44), 'preprocs': (), 'ex_preprocs': ())

Decision Tree Regressor

- Decision Tree Regressor could not be tested for Hyperopt-sklearn since it is not yet implemented according to its github page.
  `https://github.com/hyperopt/hyperopt-sklearn`

Table of Contents

Interpretation and comparison of results

- The experiments have shown that both implemented AutoML algorithms performed very well, finding better sets of hyperparameters than they started with. They could however not reach the quality of the hyperparameter-sets of existing implementations. Especially TPOT performed extremely well, outperforming every other algorithm in almost every test.