Universiteit Leiden

Practical session week 8

Data Science course

This week we finish working on assignment 2.

End goal for assignment 2

Deadline: April 17.

Task: Predict the relevance of search results on homedepot.com

- Information on the task: https://www.kaggle.com/c/home-depot-product-search-relevance/
- Information on the data: https://www.kaggle.com/c/home-depot-product-search-relevance/data

End product: a report containing:

1. (week 5) The task definition
2. (week 5) Data description: results of the data exploration
3. (week 5) Description of the baseline method
4. (week 6, 7) Description of the features you defined and the hyperparameter optimization
5. (week 6, 7) Results: a table with (a) Baseline results (replication of existing method); (b) Results for multiple regression models; (c) Results of the hyperparameter optimization, and (c) the results for different feature representations you experimented with
6. (week 8) A brief conclusion: which feature representation worked the best and why

This is a group assignment. I advise you to write your report in Overleaf.

This week, you will work on part 6 of the report.

Goals for week 8

- Learn to inspect the feature weights learned by an estimator
- Learn to report the results of a machine learning experiment with multiple feature sets.

Preliminaries

Your code and completed tasks from weeks 5, 6, and 7.

Tasks

1. Inspect feature weights

Most estimators in sklearn have a function to inspect the learned feature weights of the model. This function is called “coef_” for most estimators (such as LinearRegression) and feature_importances_ for tree-like estimators (such as RandomForestRegressor).

Use one of these functions to output the feature weights for your regressor (note that this requires the disabling of the BaggingRegressor meta-estimator).

In order to know which weight belongs to which feature you need to store names of the extracted features in an array. Sort the features by their weight.

A list of the top-5 most important features, and an interpretation of why they are the most important, is part of the report of assignment 2.

2. Report writing

Your report needs to contain all the parts that you have completed in weeks 5, 6, 7 and 8. The structure is:

1. Task definition
2. Data exploration
3. Baseline method
4. Features and hyperparameter optimization
5. Results (as tables):
    a) Baseline results (replication of existing method);
    b) Results for multiple regression models;
    c) Results of the hyperparameter optimization;
    d) the results for different feature sets you experimented with
6. Feature analysis and conclusion

Keep it concise; maximum 4 pages. Don't add information that is not asked for (such as your struggles with the process, or how you downloaded and processed the data). Writing concise reports is an important exercise.

This is a team assignment. I advise you to write your report in Overleaf.

Please submit your report **as a single pdf (not zipped)**, together with your code, in Brightspace.