Practical session week 7

Data Science course

This week we continue 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. More specifications for the report for assignment 2 will be provided in the following weeks.

This week, you will work on part 4 and 5 of the report.

Goals for week 7

- Learn to make an informed choice on the type of classifier given the feature set
- Learn to optimize hyperparameters in sklearn
- Understand the effect of preprocessing for machine learning

Preliminaries

Your code and completed tasks from weeks 5 and 6.

Tasks

1. Regression models

In the code by Yao-Jen Chang, RandomForestRegressor is used to learn the regression model, in combination with a BaggingRegressor as meta-estimator.

Find three other regression models in the sklearn documentation and compare these for the task, both in quality (RMSE) and processing time.

A comparison of the results for four regression models are part of the report for assignment 2.

2. Hyperparameter optimization

Select the model that works the best. You will now optimize the model's hyperparameters. In Sklearn there are some very simple hyper parameter tuning methods: https://scikit-learn.org/stable/modules/grid_search.html . It is also possible to use more advanced methods such as Bayesian Optimization (https://github.com/wangronin/Bayesian-Optimization/).

Make sure you are not optimizing on your test set; you will need to use cross validation on the train set (e.g using the function RandomizedSearchCV)

A description of de optimization procedure is part of the report of assignment 2.

3. The effect of pre-processing & knowledge-driven feature extraction

Last week you have added features for the query-product matching and evaluated the efficacy of each feature. You can continue with that this week. One suggestion is the following:

- You will notice that the domain has some peculiarities, such as the importance of numbers and units (both in the product descriptions and the queries). Examples are terms such as "8 ft.", "mdf 3/4". If you haven't included any feature regarding numbers and units yet, add pre-processing steps and feature extraction functions that can handle these specific aspects of the products and queries.

A description of the features you implemented, and an evaluation of the feature sets are part of the report for assignment 2.