Project in data intensive system

Assignment 1:

Create a Jupiter notebook with text and Python code by following the instructions and the example of "Assignment2.pdf".

📄 Assignment 1 4DV652.ipynb
26 January 2021, 7:20 PM
📄 Assignment 1 4DV652.pdf

Assignment 2:

- Train a linear regression model that maps features to expert scores based on the dataset
- Try different variants using, e.g., feature selection, combined features, removal of outlier and leverage data points, apply symmetry constraints, weighted linear regression
- Assess the accuracy of the models and select a champion variant

Software development

- Deploy a service with the champion variant that receives a record of feature values and returns an expert score
- Deploy a remote test Web client that interacts with that service
- Define a set to automatically and continuously test and deploy new variants of the regression model (devops process)

Reporting in a notebook:

- iterations over ML process steps,
- deployed client-server system (usage, architecture, design, and implementation) and
- devops process

AimoScore_WeakLink_big_scores.xls I have provided these file separately

Assignment 3:
ML

- Classify the weakest link based on the 38 predictors of a movement using a number of classification variants (different models or model parameterizations)
- Select an accuracy metric
- Test and iteratively improve the accuracy of the variants

Software development – Dependency management

- Add a new endpoint for weakest link classification to the server that implements the champion variant
- Extend the web client to use also this new endpoint
- Set up, implement, and enforce dependency management

Reporting in a fourth notebook:

- The iteration(s) over ML process steps
- Dependency management strategy, its implementation, and how you enforce it

Don't forget to give us access to the code repository as well.
20190108 scores_and_weak_links.xlsx
25 February 2020, 1:55 PM
AimoScore_WeakLink_big_scores.xls

Assignment 4:
ML

- Improved the regression and classification models using cross validation (CV)
- Test the improved accuracy of the champion variants in scoring and classification

Software development – Maintenance sprint

- Define design and coding rules

- Define and implement formatting, linting, dependency management, testing, peer review
- Apply it to the code
- Add barriers to CI/CD enforcing the above quality assurance (QA)

Reporting in a third notebook:

- CV approach and improvements in the accuracy of the models (if any)
- Major improvements to the client/server app and the DevOps/QA process
- QA rules and how you enforce them

Assignment 5

ML

- Implement cross-validation for selecting the champion approach for the two endpoints
- Challenge the current champion regression with a tree-based approach
- Challenge the current champion classification with a tree-based approach

Software development

- If applicable, implement and deploy the new champion regression and classification

Reporting in a fifth notebook:

- In a fourth notebook, document the iteration(s) over the ML process steps
- Prepare a short talk/presentation of the projects (max. 20 min) for your "customers"

Assignment 5b:

ML
•. Challenge the current champion regression with an ensemble approach
•. Challenge the current champion classification with an ensemble approach
•Software development
•. If applicable, implement and deploy the new champion regression and classification
•Reporting:
•In a next notebook, document the iteration(s) over the ML process steps

Assignment 6

ML

- Consider SVN classification as an alternative to the classification endpoint
- Implement grid optimization for selecting the champion approach for this endpoint
- Challenge the current champion classification with the SVM approach

Software development

- Consider the most accurate of each implemented classification model: logistic, k-means, Bayes, SVM, …
- Assess the classification performance, i.e., response time, of these different classifiers
- If applicable, deploy the SVM classifier as the new implementation of the classification endpoint

Reporting in a sixth notebook:

-
- Document the iteration(s) over the ML process steps
- Report classification performances of the alternative classifiers

Assignment 7:

ML

- Perform a clustering of the AIMO observations (exclude the scores)
- Interpret the clusters
- Perform a PCA of the AIMO observations
- Draw the first two principle component scores of the data points and their clusters (colored 2D scatterplot)
- Interpret the plot

Software development

-
- Maintenance sprint

Reporting in a seventh notebook:

- Document the ML steps
-
- Revise all six earlier submitted notebooks if necessary

Assignment 8:

ML:

- Go through the TensorFlow and Keras tutorials for beginners

Software development:

- Adapt the PoseNet Python port to run it in your desktop environment
- Understand the code
- Run it on your own images and videos
- Adapt the code so that the positions of joints get stored in a separate (csv, json) file

Reporting in a seventh notebook:

- the installation and usage of the adapted Python port
- the changes to the software

Assignment 9:

ML:

- Go through the Deep Learning steps
- Vary the parameters
- Suggest the optimal setup

Software development:

- Make sure you can easily retrain and use the model

Reporting in an eights notebook:

- the Deep Learning steps
- Optimum settings (what worked best) and dead ends (what did not work)

Don't forget to give us access to the code repository.

kinect_good_preprocessed.zip