Week 7 (29.11 - 09.12.2019)

Part 1

Metrics for Ensemble Feature Selection with lightGBM usage

The 'early stopping' mechanism is used here, where the model with the lowest validation error over 100 runs is selected.

	Light GBM
Accuracy	0.86
Precision	0.46
Recall	0.56
F1 Score	0.51
ROC-AUC Score	0.73

Part 2

Check performance of ensemble feature selection with a regression model

We fit the model on regression data, and then convert the test labels and predicted labels to 0 (if the label is less than 40) and 1 otherwise.

Code can be found in ensembleFeatureSelectionRegression.py

	Random Forest	**Gradient Boosting**	**SVM Classifier**	**Light GBM**
Precision	0.94	0.55	0.07	0.48
Recall	0.28	0.45	0.13	0.47
F1 Score	0.43	0.49	0.09	0.47
ROC-AUC Score	0.64	0.70	0.47	0.70

Part 3

Check performance of ensemble feature selection with a 3D mordred descriptors

Without feature selection:

	Random Forest	**Gradient Boosting**	**SVM Classifier**	**Light GBM**
Accuracy	0.91	0.90	0.89	0.88
Precision	0.90	0.79	0.71	0.48
Recall	0.28	0.38	0.27	0.47

F1 Score	0.42	0.51	0.39	0.47
ROC-AUC Score	0.63	0.68	0.62	0.70

With feature selection:

	Random Forest	**Gradient Boosting**	**SVM Classifier**	**Light GBM**
Accuracy	0.91	0.91	0.89	0.87
Precision	0.90	0.75	0.75	0.43
Recall	0.28	0.40	0.23	0.43
F1 Score	0.42	0.53	0.35	0.43
ROC-AUC Score	0.63	0.69	0.61	0.67

Part 4

Feature Selection with BorutaPy

Boruta is an all relevant feature selection method, while most other are minimal optimal. This means it tries to find all features carrying information usable for prediction, rather than finding a possibly compact subset of features on which some classifier has a minimal error.

This makes it really well suited for biomedical data analysis, where we regularly collect measurements of thousands of features (genes, proteins, metabolites, microbiomes in your gut, etc), but we have absolutely no clue about which one is important in relation to our outcome variable, or where should we cut off the decreasing “importance function” of these.

You can find the relevant code in borutaFeatureSelection.py

	Random Forest	**Gradient Boosting**	**SVM Classifier**	**Light GBM**
Accuracy	0.91	0.91	0.89	0.90
Precision	0.90	0.74	0.73	0.62
Recall	0.26	0.45	0.21	0.46
F1 Score	0.41	0.56	0.32	0.53
ROC-AUC Score	0.63	0.71	0.60	0.71