Introduction to Machine Learning in R (NPFL054)

Easy HW – Data analysis and Logistic regression

Contact: Barbora Hladká (hladka@ufal.mff.cuni.cz)

Data

- Titanic data set
  - `https://ufal.mff.cuni.cz/~hladka/2021/docs/train.csv`
  - `https://ufal.mff.cuni.cz/~hladka/2021/docs/test.csv`
- Movie data set - `https://ufal.mff.cuni.cz/~hladka/2016/docs/mov.development.csv`

Questions

1. Load the Titanic data sets, both the train and test set, and merge them into a single data set. Explore this set graphically using tools of your choice. Create some plots highlighting the relationships among the attributes. Comment on your findings.
2. Load the Titanic `train` data set and split it into a training set and test set in 90:10 ratio. Using the training data set fit logistic regression models with `Survived` as a target binary attribute. Experiment with different subsets of the given features. Do not forget to handle the missing values using a reasonable method. Evaluate your models on the test data set using the measures Accuracy, Precision, Recall, and F-measure.
3. Load the Movie data set and split it into a train set and test set in 90:10 ratio. Using the train set fit logistic regression models with `rating` as a target categorical attribute having 5 different values. Use `one-to-all` method for multi-class classification. Experiment with different subsets of the given features. Evaluate your models on the test set using the measures Accuracy, Precision, Recall, and F-measure.

Presentation

- Create a 20 min presentation.
- Present your answers. If you want to highlight something in your R code, please do it.
- Explain your answers clearly so that your audience understands your method well.