Tutorial 10 - Cross-Validation & Evaluation

AUTHOR
Victoria Hünewaldt

Tutorial 10

This tutorial will cover out-of-sample predictions, cross-validatoin and evaluation metrics. You will learn:

- how to split your data into a test & training set
- how to train your model on the training data & make predictions on the test data
- how to calculate the out-of-sample RSME
- how to apply a cross-validation procedure
- how to get a confusion matrix & evaluate your model's performance based on accuracy, precision, recall

Exercises

You will need to use the "caret" library. Create a .qmd-file and solve the tasks there. Store it in the JupyterHub folder "Session 10".

Out-of-sample prediction

Use your results from the logistic regression model from tutorial 9. Test how your model performs out-of-sample. Perform the following steps.

1. Split your data into a training and test sample (70% training, 30% testing).
2. Train your model on the training data.
3. Make predictions on the test set using the trained model.
4. Calculate the out-of-sample RMSE, i.e. the test set RMSE.

Cross-validation, confusion matrix & evaluation metrics

Now apply the cross-validation procedure using the "trainControl" command and do a 5-fold and 10-fold cross-validation. Compare the results.

To evaluate your model create a confusion matrix with the "confusionMatrix" command and find accuracy, precision and recall scores.
For the confusion matrix you will need two classes "negative" and "positive" of the predicted values as well as two classes "negative" and "positive" of the observed values from the test set. Predicted probabilities <0.5 should fall in the predicted class "negative" (0), whereas predicted probabilities >0.5 should fall in the predicted class "positive" (1).