1. November

OPENCAMPUS.sh

Time Series Prediction

WEEK 2

FORECASTING BASICS WITH TRENDS: AR + MA-MODELS

ATTENDANCE REGISTRATION

Online:

- **Use your full names in the zoom meetings!**
- **Only counts as attended with camera on.**

- **Organizational Matters:**
  - **Projects**

- **Session 1:**
  - **Data Preparation**
  - **AR + MA**

PROJECTS:

HTTPS://OPENCAMPUS.GITBOOK.IO/OPENCAMPUS-MACHINE-LEARNING-PROGRAM/PROJECTS/REQUIREMENTS

HTTPS://GITHUB.COM/OPENCAMPUS-SH/ML-PROJECT-TEMPLATE/TREE/MAIN

- FINANCE-1:
  DUYEN TIEU, HUBERT SZURPNICKI, IFTEKHAR, ALEXANDER NÜRENBERG

  FINANCE-2:
  JONAS MIELCK, LEIF FEDDERSEN, KINGSLEY, DAMILARE OSUNLEKE

  ENERGY:
  LIHUI HUANG, EJOEL METZ, KAI PAULSEN, SUDESH,

  WEATHER: CANCELLED

  MEDICAL: CANCELLED

  ENVIRONMENTAL:
  JOANNE, KELVIN LAMP, COSIMA, JOHANNA

  TRAFFIC: CANCELLED

SESSION 1:

MODELS FOR FORECASTING:

KEY TAKEAWAYS:

- Univariate time series models use historical data of a target variable to make predictions.
- Seasonality and trend are important effects that can be used in time series models.
- Supervised machine learning uses correlations between variables to forecast a target variable.
- Supervised machine learning can be split into classification and regression. In classification, the target variable is categorical; in regression, the target variable is numeric. Regression is most relevant for forecasting.
- The correlation coefficient is a KPI of the relationship between two variables. If the value is close to 1 or -1, the correlation is strong; if it is close to 0, it is weak. A strong correlation between an explanatory variable and the target variable is useful in supervised machine learning.
- Univariate models predict one variable. Multivariate models predict multiple variables. Most forecasting models are univariate, but some multivariate models exist.

MODEL EVALUATION FOR FORECASTING

KEY TAKEAWAYS:

- Metrics:
  - The most suitable metrics for regression problems are R squared, RMSE and MSE.
  - The R2 gives a percentage-like value.
  - The RMSE gives a value on the scale of the actuals.
  - The MSE gives a value on a scale that is difficult to interpret.
  - Metrics should be used for benchmarking different models on one and the same dataset.
- Model evaluation strategies
  - Cross-validation gives you a very reliable error estimate.
  - Adaptations are necessary to make it work for time series.
  - A train, test, and validation split can be used for benchmarking.
  - A combined strategy will give you the safest estimate:
  - Use cross-validation on the training data, validation data for model selection, and test data for a last estimate of the error.
- Overfitting models / Underfitting models
  - If your model learns too much from the training data and will not generalize into the future, it is overfitting. Overfitting is identified by a good performance on the train data but a bad performance on the test data.

THE AR-MODEL

AR+MA-FAMILY:

Name	Explanation	Chapter
AR	Autoregression	3
MA	Moving Average	4
ARMA	Combination of AR and MA models	5
ARIMA	Adding differencing (I) to the ARMA model	6
SARIMA	Adding seasonality (S) to the ARIMA model	7
SARIMAX	Adding external variables (X) to the SARIMA model *(note that external variables make the model not univariate anymore)*	8

KEY TAKEAWAYS:

- The AR model predicts the future of a variable by leveraging correlations between a variable's past and present values.
- Autocorrelation is correlation between a time series and its previous values.
- Partial autocorrelation is autocorrelation conditional on earlier lags - it prevents double counting correlations.
- The number of lags to include in the AR model can be based on theory (ACF and PACF plots) or can be determined by a grid search. A grid search consists of doing a model evaluation for each value of the hyperparameters of a model. This is an optimization method for the choice of hyperparameters.
- Yule-Walker equations are used to fit the AR model. Fitting the model means finding the coefficients of the model

THE MA-MODEL

KEY TAKEAWAYS:

- The Moving Average model predicts the future based on impulses in the past:
  - Those impulses are measured as model errors.
  - The idea behind this is that unexpected impacts can actually have a large impact on the future
  - There is no one-shot computation for the MA model coefficients, so it takes a bit longer to estimate this model compared to the AR model
- The MA model is the second building block of the SARIMA model. The AR and MA models together form the ARMA model, the topic of the next sessions
- Multistep predictions are predictions of multiple time steps into the future. This is difficult with MA models, especially when the order is low. A solution can sometimes be to do multiple one-step predictions and retrain the model every time that you receive the new data.
- The autocorrelation function and the partial autocorrelation function can help you decide whether you are looking at an AR or an MA forecast. AR forecasts see their autocorrelation exponentially decay to 0 and alternate between positive and negative. MA autocorrelation functions are characterized by many values of almost zero and a few spikes

TASKS UNTIL NEXT WEEK

- **Completion of the learning material of week 2: watch the second YouTube-playlist ;-)**

- **Complete/prepare the IPython-Notebooks:**
- **i.e. ARMA: Group-Finance-1**
- **i.e. ARIMA: Group-Finance-2**
- **i.e. SARIMA: Group-Energy**

- Chapter 5/6/7
- https://github.com/Apress/advanced-forecasting-python

- **Bring questions!**