Applied Machine Learning

Time Series:
Definition, feature engineering, cross-validation

Learning goals

- What is a time series?
- Important definitions
- Feature engineering for time series
- Time series cross-validation

RUNNING EXAMPLE: THE BIKE SHARING DATASET

Hourly bike rental data from Washington D.C.

	holiday	workingday	weather	temp	feel_temp	humidity	windspeed	datetime	count
0	False	False	clear	9.84	14.395	0.81	0.00	2011-01-01 00:00:00	16
1	False	False	clear	9.02	13.635	0.80	0.00	2011-01-01 01:00:00	40
2	False	False	clear	9.02	13.635	0.80	0.00	2011-01-01 02:00:00	32
3	False	False	clear	9.84	14.395	0.75	0.00	2011-01-01 03:00:00	13
4	False	False	clear	9.84	14.395	0.75	0.00	2011-01-01 04:00:00	1

KEY DEFINITIONS

- **Time series** $\{y_t\}_{t=1}^{T}$: ordered observations indexed by time:
  - Timestamps - Specific instants in time.
  - Fixed periods - Such as the whole month of January 2017, or the whole year 2020.
  - Intervals of time - Indicated by a start and end timestamp. Periods can be thought of as special cases of intervals.
  - Experiment or elapsed time - Each timestamp is a measure of time relative to a particular start time (e.g., the diameter of a cookie baking each second since being placed in the oven), starting from 0.
- **Exogenous variable** / **covariate**: external feature $x_t^{(j)}$ influencing y_t
- **Trend**: long-term increase or decrease
- **Seasonality**: systematic, calendar-related patterns
- **Stationarity** (briefly): distribution of y_t does not change over time

TASKS

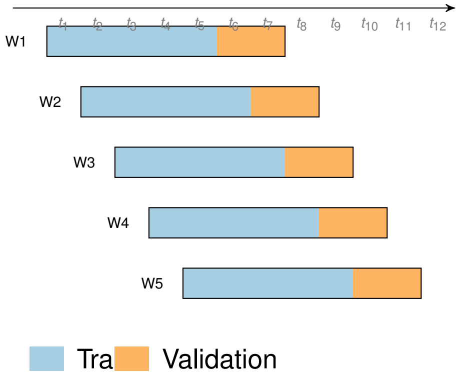
- Time Series Analysis Decompose time series into trends for understanding $\rightarrow$ time series course
- Time Series Forecasting
  Fit a model on historical data and make predictions about the future.
- Time Series Classification
  Make predictions about one complete time series, e.g., one recording of an experiment.

HOW TO TACKLE TIME SERIES?

- Statistical Models ARIMA (Autoregressive integrated moving average) and ETS (Error, Trend, Seasonality) $\rightarrow$ time series course
- Recurrent Models
  Deep learning architectures that track the state of data over time $\rightarrow$ part of the deep learning lecture
- Transformer Models
  Deep learning architecture that can treat series data in an auto-regressive manner $\rightarrow$ part of the deep learning lecture
- Feature Engineering for Machine Learning
  Make predictions about one complete time series, e.g., one recording of an experiment.

Time-Series Cross-Validation

CHRONOLOGY-AWARE SPLITS



- **Expanding window**: grow train set, roll validation forward
- **Sliding window**: fixed-width train and validation segments
- Validation window should match the application forecasting windows

AVOIDING DATA LEAKAGE

- Compute lagged / rolling features *after* defining the split or in a way it takes the split into account
- No peeking into future data when standardizing / scaling
- Requires special cross-validation methods

EVALUATION METRICS

- **MAE** $= \frac{1}{n}\sum|y_t - \hat{y}_t|$ (robust to outliers)
- **RMSE** (penalizes large errors)
- **SMAPE**: symmetric MAPE avoids division issues when y_t near 0
  - Symmetric Mean Absolute Percentage Error
  - SMAPE $= \frac{100}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{|A_t|+|F_t|}$

SIMPLE BASELINE

$$y_{t+1} = y_t$$

Feature Engineering

CALENDAR FEATURES

- Day-of-Week, Day-of-Month, Week Number, Month, Quarter
- Boolean flags: public holidays, weekend, promotion days, end-of-month

Why?

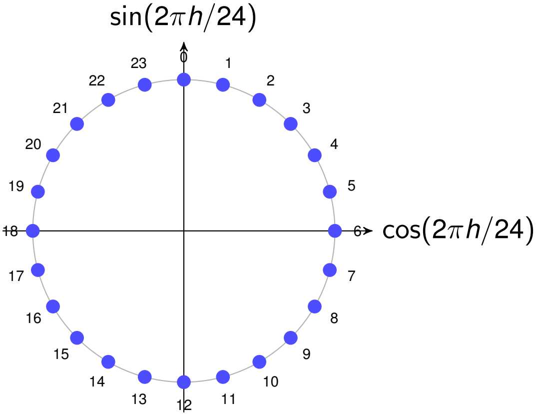
Capture human or process-driven periodicity.

CYCLIC ENCODING

$$\sin(2\pi \frac{hour}{24}), \;\; \cos(2\pi \frac{hour}{24})$$

- Converts periodic integers (hour 0-23, month 1-12) into continuous space
- Eliminates artificial jump between 23 $\rightarrow$ 0

CYCLIC ENCODING VISUALIZATION (HOUR 0-23)



LAGGED & AUTO-REGRESSIVE FEATURES

- Plain lags: $y_{t-1},\ y_{t-7},\ y_{t-14}$
- Multi-step: include future-known covariates (e.g., planned price)
- Combine with calendar features for interactions

ROLLING / EXPANDING STATISTICS

- Rolling mean / std / min / max over window
- Expanding mean: trend indicator
- Percentiles: 25th, 75th for distribution shape

INTERACTION FEATURES

- Lag $\times$ Holiday flag
- Weather $\times$ Weekend indicator
- Captures non-linear, conditional effects

Wrap-Up

KEY TAKEAWAYS

- Feature engineering often outweighs model complexity in time-series ML.
- Use chronology-aware CV to obtain honest performance estimates.
- Always benchmark against a naive or seasonal naive baseline.