Intro for Senior Design

Zhankun Luo

Center for Innovation through Visualization & Simulation
(CIVS)
Purdue University Northwest
Hammond, IN 46323

September 30, 2020

Agenda

- Python Basics
- Time Series Forecasting Methods
- Outlier Detection Techniques

Task

- Python Basics
  - pytorch
  - data processing library
  - measurement

Python Basics

- pytorch
  - installation: Link document: Link
  - tutorial: easy start Link LSTM examples Link Link
- data processing library
  - pandas: process data frame
  - csv: read/write csv
  - xlsread, xlswrite: read/write xls, xlsx
  - pyodbc: execute SQL query
- measurement

definition

- MAE $\equiv \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| = \frac{1}{N}\sum_{i=1}^{N}|e_i|$
- MAPE $\equiv \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right| = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{e_i}{y_i}\right|$
- RMSE $\equiv \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$
- $R^2 \equiv 1 - \frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{\sum_{t=1}^{T}(\mu - y_t)^2}$, where $\mu \equiv \frac{1}{T}\sum_{i=1}^{T} y_t$

Task

- Time Series Forecasting Methods
  - ARIMA(p, d, q)
  - decomposition of time series
    - Linear+Nonlinear
    - freq. domain
    - Trend+Season+Holiday: Prophet (Facebook)
  - machine learning
    - XGBoost
    - LightGBM (Microsoft)
    - GPR: Gaussian process regression
  - neural networks
    - LSTM, GRU
    - WavNet
    - seq2seq
    - Self-boosted: DeepAR (Amazon)
    - attention mechanism: transformer

Time Series Forecasting Methods

- ARIMA(p, d, q)
  - wikipedia: Link
  - zhihu: Link
  - implementation: Link
  - statsmodels: arima Link acf $\Rightarrow q$, pacf $\Rightarrow p$ Link Link
  - example: Link

definition

- $\left(1 - \sum_{i=1}^{p} \varphi_i L^i\right) (1 - L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$
- L: lag operator; φ_i: params of AR part; θ_i: params of MA part
- ε_t: error term, should be **white noise** (independent, identically distributed variables sampled from a normal distribution with zero mean)
  test ε_t white noise Link, test ε_t stationary Link

Time Series Forecasting Methods

- decomposition of time series
  - Linear+Nonlinear Link

definition

- $y(t) = L(t) + N(t)$, where $L(t) \equiv \frac{1}{T}\sum_{t-T+1}^{t} y(\tau), \quad N(t) \equiv y(t) - L(t)$
- select $T \Rightarrow L(t)$ satisfies Gaussian distribution $\Rightarrow$ ARIMA model $\hat{L}(t)$ to fit $L(t)$
- $\hat{L}(t), [y(t-1), ..., y(t-q)], [N(t-1), ..., N(t-p)] \Rightarrow$ (NN, nonlinear kernel) $\hat{y}(t)$

  - freq. domain
    - FD: Fourier decomposition
    - WD: wavelet decomposition
    - EMD: empirical mode decomposition
      wikipedia: Link zhihu: Link Link implementation: Link
    - VMD: variational mode decomposition
      paper: Link zhihu + implementation: Link
  - Trend+Season+Holiday: Prophet (Facebook) Link Link zhihu Link

definition

- $y(t) = g(t) + s(t) + h(t) + \varepsilon_t$, where $g(t)$: trend, $s(t)$: season, $h(t)$: holiday

Time Series Forecasting Methods

- machine learning
  - XGBoost
    - paper: Link document: Link
    - zhihu: Link Link
    - implementation: Link
    - example: Link Link Link
  - LightGBM (Microsoft)
    - paper: Link document: Link
    - zhihu: Link
    - example: Link Link
  - GPR: Gaussian process regression
    - book: Link sklearn: Link
    - zhihu: Link Link Link
    - implementation Link
    - example: Link Link

Time Series Forecasting Methods

- neural networks
  - GRU
    - wikipedia: Link document: Link
    - blog Link implementation: Link
  - WavNet (DeepMind)
    - paper: Link
    - zhihu Link example: Link Link
  - seq2seq
    - zhihu Link implementation Link Link
    - example: Link
  - attention mechanism: transformer
    - paper Link
    - example Link

Task

- Outlier Detection Techniques
  - visualization: tableau
  - Z-score
  - DBSCAN
  - isolation forest
- reading

How to Identify Outliers in your Data

A Brief Overview of Outlier Detection Techniques

Four Techniques for Outlier Detection

Outlier Detection Techniques

- visualization: tableau
  - installation: Link
  - tutorial: Link Link
- Z-score: assume Gaussian distribution

usage

- $z_i = \frac{x_i - \mu}{\sigma}$ where data x_i, $\mu \equiv \sum x_i / N$, $\sigma \equiv \sum (x_i - \mu)^2 / (N-1)$
- For abnormal value $|z_i| > z_{\mathrm{th}}$, where z_{th} should always be 2.5, 3.0 or 3.5

- DBSCAN
  - wikipedia: Link
  - sklearn: Link
  - example: Link
- isolation forest
  - description: Link Link
  - implementation: Link
  - sklearn: Link
  - example: Link