Introduction to Time Series Prediction

(draft lecture notes in Advanced Data Mining)

Piotr Lipiński

Time Series

- **Definition:** A time series (x_t) is a sequence

$$x_1, x_2, \ldots, x_t, \ldots$$

  of some observations of the phenomenon under study in the successive time instants.

  - for the sake of simplicity, we assume that the observations are recorded in the regular intervals, e.g. monthly, weekly, daily, but in general they may be also irregular
  - although we assume that the sequence is infinite, because the future is infinite, we usually have a finite sequence of the historical values and we usually predict a finite sequence of future values

Time Series

- we assume that the time series (x_t) is an instance of a stochastic process (X_t) being a sequence of random variables

$$X_1, X_2, \dots, X_t, \dots$$

in such a way that

$$\begin{array}{ccccccc} x_1, & x_1', & x_1'', & \dots & \leftarrow & X_1 \\ x_2, & x_2', & x_2'', & \dots & \leftarrow & X_2 \\ \vdots & \vdots & \vdots & & & \vdots \\ x_t, & x_t', & x_t'', & \dots & \leftarrow & X_t \\ \vdots & \vdots & \vdots & & & \vdots \end{array}$$

Time Series Prediction

- **Goal:** using the historical values $x_1, x_2, \ldots, x_t$, predict the future values $x_{t+1}, x_{t+2}, \ldots, x_{t+h}$, for a given time instant t and a given time horizon h
- **Requirements:**
  - the time series (x_t) is an instance of a stochastic process (X_t)
  - the historical values allow to estimate the characteristics of the random variables $X_1, X_2, \ldots, X_t$ (e.g. their parameters)
  - the characteristic of the entire stochastic process (X_t) (especially the autocorrelation) allows to characterise the random variables $X_{t+1}, X_{t+2}, \ldots$ on the basis of the characteristics of the random variables $X_1, X_2, \ldots, X_t$
- **Remark:** in practice, even if we predict a longer time horizon $h > 1$, we may usually evaluate the entire prediction iteratively, updating it day by day, so we may usually assume $h = 1$

Time Series Prediction

- **Approach 1:** In some cases, the characteristics of the stochastic process and its random variables are quite regular and they may be discovered from the historical values.
- **Definition:** A stochastic process (X_t) is stationary if for all $t = 1, 2, \ldots$ and for all $k = 0, 1, 2, \ldots$,
  - the expected value is constant

$$\mathbb{E}(X_t) = \mathbb{E}(X_{t+k}) = \mu$$

  - the autocovariance depends only on k

$$\mathrm{Cov}(X_t, X_{t+k}) = \mathrm{Cov}(X_1, X_{k+1}) = \gamma_k$$

- In practice, for stationary time series, it usually should be possible to estimate μ and $\gamma_0, \gamma_1, \gamma_2, \ldots$ on the basis of the historical values.

Time Series Prediction

- **Approach 2:** If a time series is not stationary, some simple techniques may stationarize it.
  - detecting and removing the trend
  - detecting and removing the seasonality
  - detecting and removing other anomalies
  - differencing
- **Approach 3:** Even after removing trends/seasonalities/anomalies some time series may have the variance changing with time and then may require some different approaches.

AR(p) model

- **Autoregressive AR(p) model:**

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t$$

  where $c \in \mathbb{R}$ is a constant, $\phi_1, \phi_2, \ldots, \phi_p \in \mathbb{R}$ are autoregression parameters and $\epsilon_t \sim \mathcal{N}(0, 1)$ is the noise uncorrelated with the random variables $X_1, X_2, \ldots, X_t$
- **Random Walk:** A special case of AR(1) with $c = 0$ and $\phi_1 = 1$ is called the random walk.

MA(q) model

- **Moving Average MA(q) model:**

$$X_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}$$

where $c \in \mathbb{R}$ is a constant, $\theta_1, \theta_2, \ldots, \theta_q \in \mathbb{R}$ are moving average parameters and ϵ_t is the noise

ARMA(p, q) model

- **Autoregressive Moving Average ARMA(p, q) model:**

  $$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \ldots - \theta_q \epsilon_{t-q}$$

  where $c \in \mathbb{R}$ is a constant, $\phi_1, \phi_2, \ldots, \phi_p \in \mathbb{R}$ are autoregression parameters, $\theta_1, \theta_2, \ldots, \theta_q \in \mathbb{R}$ are moving average parameters and ϵ_t is the noise

ARIMA(p, d, q) model

- One of simple techniques of stationarizing a time series is differencing. It transforms the time series (x_t) into the time series (y_t) such that

$$y_t = x_t - x_{t-1}.$$

- **Autoregressive Integrated Moving Average ARIMA(p, d, q) model:** is ARMA(p, q) applied to the time series after differencing d times

Estimation of Parameters of the ARIMA Model

- General parameters p, d, q may be defined using
  - Akaike Information Criterion (AIC)

    $$AIC = 2n - 2\log(L),$$

    where n is the number of parameters of the model, and L is the maximum value of the likelihood function of the model
  - Bayesian Information Criterion (BIC)

    $$BIC = n\log(N) - 2\log(L),$$

    where n is the number of parameters of the model, N is the number of samples, and L is the maximum value of the likelihood function of the model
  - on the basis of the autocorrelation function, etc.

Estimation of Parameters of the ARIMA Model

- Parameters of the ARIMA model may be defined using
  - linear/non-linear least squares estimation
  - methods of moments estimation
  - Yule-Walker estimation

Yule-Walker Estimation of the AR(p) Model

For the sake of simplicity, assume that $\mu = 0$. Thus, AR(p) gives

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t. \tag{1}$$

After multiplying both sides by X_{t-1}, we have

$$X_{t-1}X_t = \sum_{i=1}^{p} \phi_i X_{t-1} X_{t-i} + X_{t-1}\epsilon_t,$$

so

$$\mathbb{E}(X_{t-1}X_t) = \sum_{i=1}^{p} \phi_i \mathbb{E}(X_{t-1}X_{t-i}) + \mathbb{E}(X_{t-1}\epsilon_t),$$

but $\mathbb{E}(X_{t-1}\epsilon_t) = 0$ because ϵ_t is uncorrelated with X_{t-1} and $\mathbb{E}(\epsilon_t) = 0$. Moreover

$$\mathbb{E}(X_{t-1}X_{t-i}) = \mathbb{E}(X_{t-1}X_{t-i}) - \mathbb{E}(X_{t-1})\mathbb{E}(X_{t-i}) = \text{Cov}(X_{t-1}, X_{t-i}) = \gamma_{i-1},$$

because $\mathbb{E}(X_{t-1}) = \mathbb{E}(X_{t-i}) = \mu = 0$, hence

$$\gamma_1 = \sum_{i=1}^{p} \phi_i \gamma_{i-1}.$$

Yule-Walker Estimation of the AR(p) Model

Therefore, defining $r_i = \frac{\gamma_i}{\gamma_0}$ for $i = 0, 1, 2, \ldots, p$, we have

$$r_1 = \sum_{i=1}^{p} \phi_i r_{i-1}. \tag{2}$$

Similarly, after multiplying both sides of (1) by X_{t-2}, we have

$$r_2 = \sum_{i=1}^{p} \phi_i r_{i-2} \tag{3}$$

(denoting $r_{-i} = r_i$), after multiplying both sides of (1) by X_{t-3}, we have

$$r_3 = \sum_{i=1}^{p} \phi_i r_{i-3}, \tag{4}$$

etc.

Yule-Walker Estimation of the AR(p) Model

Finally,

$$\begin{array}{rcrcrcccr} r_1 & = & \phi_1 r_0 & + & \phi_2 r_1 & + & \dots & + & \phi_p r_{p-1} \\ r_2 & = & \phi_1 r_1 & + & \phi_2 r_0 & + & \dots & + & \phi_p r_{p-2} \\ & \vdots & & & & & & & \\ r_p & = & \phi_1 r_{p-1} & + & \phi_2 r_{p-2} & + & \dots & + & \phi_p r_0 \end{array} \tag{5}$$

which can be rewritten as

$$\boldsymbol{r} = \boldsymbol{R}\boldsymbol{\phi},$$

where $\boldsymbol{r} = (r_1, r_2, \dots, r_p)$, $\boldsymbol{\phi} = (\phi_1, \phi_2, \dots, \phi_p)$, and

$$\boldsymbol{R} = \begin{pmatrix} r_0 & r_1 & \cdots & r_{p-1} \\ r_1 & r_0 & \cdots & r_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p-1} & r_{p-2} & \cdots & r_0 \end{pmatrix},$$

thus

$$\boldsymbol{\phi} = \boldsymbol{R}^{-1}\boldsymbol{r}. \tag{6}$$

Validation of the Time Series Prediction

- A few error measures are popular for comparing the predicted values $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_T$ with the original ones $x_1, x_2, \ldots, x_T$
  - **Mean Square Error (MSE)**:

$$MSE(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{T} \sum_{t=1}^{T} (x_t - \hat{x}_t)^2$$

  - **Mean Absolute Error (MAE)**:

$$MAE(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{T} \sum_{t=1}^{T} |x_t - \hat{x}_t|$$

  - **Mean Absolute Percentage Error (MAPE)**:

$$MAPE(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{T} \sum_{t=1}^{T} \frac{|x_t - \hat{x}_t|}{|x_t|}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_T)$ and $\hat{\boldsymbol{x}} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_T)$.

Time Series Prediction with Computational Intelligence

- see the *Introduction to Time Series Prediction* jupyter python notebook for a simple example
- linear/non-linear regression approaches
    - linear regression
    - ridge regression
    - lasso regression
    - support vector regression
- neural network approaches
    - LSTM
    - GRU
- in computational intelligence approaches, parameters are usually estimated by learning on a train dataset and the model is validated by testing on the test dataset

References

G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung.
***Time Series Analysis: Forecasting and Control*.**
Wiley, 5th edition, 2015.

S. G. Makridakis, S. C. Wheelwright, and R. J. Hyndman.
***Forecasting: Methods and Applications*.**
Wiley, 3rd edition, 1997.

S. J. Taylor.
***Modelling Financial Time Series*.**
World Scientific, 2nd edition, 2007.