Data Science Companion

Greg Simon, gregorygsimon@gmail.com

October 17, 2020

Abstract

A reference for basic data science tools and vocabulary, explaining essential terms and concepts, examining core ideas in major areas, and putting methods in context. Includes relevent keywords and references for further

Contents

1 General Machine Learning

1.1 Model Selection

1.1.1 Akaike information criterion (AIC)

1.1.2 Bayes information criterion (BIC)

1.1.3 References

2 Deep Learning

3 Time series & Forecasting

3.1 ARIMA

An ARIMA(p, d, q) is an *autoregressive integrated moving-average* with p autoregressive terms (AR), d differencings, and q moving average (MA) terms.

$$\phi(B)(1-B)^d Y_t = c + \theta(B)\epsilon_t$$

where

- B is the back-shift/lag operator $BY_t = Y_{t-1}$.
- $\phi(B) = (1 - \phi_1 B - \ldots - \phi_p B^p)$ is the autoregressive AR(p) component
- c is a constant
- $\theta(B) = 1 + \theta_1 B + \ldots \theta_q B^q$ is the moving average of the errors MA(q) component.
- ϵ_t is the error of the AR(p) model at time t
- The $(1-B)^d$ term induces d differencing

3.2 In `R`

`auto.arima` utilizes AIC and MLE to decide on best ARIMA parameters

3.3 In `Python`

3.4 References

https://otexts.com/fpp2/non-seasonal-arima.html

4 Natural Language Processing