ARIMA Modeling

Matt Brems

GENERAL ASSEMBLY

Learning Objectives

- Describe the purpose of the autoregressive and moving average components.
- Define hyperparameters *p*, *d*, and *q*.
- Describe AIC.
- Find the right value of *p* and *q* using AIC.
- Find the right value of *d* using the augmented Dickey-Fuller test.
- Complete a manual GridSearch.
- Fit an ARIMA model.

We have multiple approaches to work with time series data.

Linear Models

ARIMA Models

Exponential Smoothing Methods

Recurrent Neural Networks (RNNs)

Note: This is not an exhaustive list of models, but lists the most common ones!

Why ARIMA?

- Among the most common approaches to time series modeling
- Highly flexible; it can model time series with varying characteristics
    - It takes information from **both long-term trends and sudden shocks**
- Can easily be extended into more advanced models
- Tends to perform well with moderate amounts of data

Downsides of ARIMA Models

- ARIMA models are best suited for **short-term forecasts**
- They will quickly start predicting the mean if forecasts are extended past the short-term
  - Some extensions to ARIMA models can handle this better

What is an ARIMA model?

ARIMA

What is an ARIMA model?

ARIMA

Autoregressive Integrated Moving Average

RECALL: What is autocorrelation?

Autoregressive models

Autoregressive means we regress a variable on itself. We'll regress newer values on older values.

AR(p): An autoregressive model of order *p*

$$
\begin{aligned}
Y_t &= \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} \\
&= \beta_0 + \sum_{k=1}^{p} \beta_p Y_{t-p}
\end{aligned}
$$

AR(p): An autoregressive model of order *p*

$$
\begin{aligned}
Y_t &= \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} \\
&= \beta_0 + \sum_{k=1}^{p} \beta_p Y_{t-p}
\end{aligned}
$$

Purpose: An autoregressive model explains **long-term trends** in our data.

Hyperparameter: *p*, the number of previous values of *Y* to put into our model.

We'll **GridSearch** to find this value!

 |

What is an ARIMA model?

ARIMA

Autoregressive Integrated Moving Average

Moving Average Models

- A moving average model takes **previous error terms as inputs**
  - Not an actual rolling/moving average, name possibly because a zero-mean moving average process can be modeled by a weighted moving average of error terms
- Use recent forecast errors as a component in the model

MA(q): A moving average model of order *q*

$$
\begin{aligned}
Y_t &= \mu + w_1 \varepsilon_{t-1} + w_2 \varepsilon_{t-2} + \cdots + w_p \varepsilon_{t-q} \\
&= \mu + \sum_{k=1}^{q} w_q \varepsilon_{t-q}
\end{aligned}
$$

MA(q): A moving average model of order *q*

$$Y_t = \mu + w_1 \varepsilon_{t-1} + w_2 \varepsilon_{t-2} + \cdots + w_p \varepsilon_{t-q}$$
$$= \mu + \sum_{k=1}^{q} w_q \varepsilon_{t-q}$$

Purpose: A moving average model explains **sudden shocks** in our data.

Hyperparameter: *q*, the number of previous errors $\mathcal{E}$ to put into our model.

We'll **GridSearch** to find this value!

How do we GridSearch to find the best values of *p* and *q*?

- Because we're working in statsmodels, we will **manually GridSearch** values of *p* and *q* to see which gives us the **lowest AIC**
- AIC, or Akaike Information Criterion, is a common way to evaluate time series models. (AIC is an attribute in **`statsmodels`**)
- Remember that a model is a simplification of reality?
    - AIC attempts to measure how much information we lose when we simplify reality with a model

$$AIC = 2 \times [\text{\# of model parameters}] - 2 \times \log(\text{likelihood})$$

What is an ARIMA model?

ARIMA

Autoregressive Integrated Moving Average

Differencing

- Many time series models require **stationarity**
- If the data are non-stationary, model the **differenced** data
- Usually difference with order of 1 or 2
  - Extremely rare to need more than 2 differences, and over-differencing will lead to poor models

ARIMA equation

$$Y_t^{(d)} = \beta_0 + \sum_{k=1}^{p} \beta_k Y_{t-k}^{(d)} + \sum_{i=1}^{q} w_i \varepsilon_{t-i} + \varepsilon_t$$

At this point, it's helpful to see what an ARIMA model is.

ARIMA

We literally just added together the AR(p) and MA(q) models.

$$Y_t^{(d)} = \beta_0 + \sum_{k=1}^{p} \beta_k Y_{t-k}^{(d)} + \sum_{i=1}^{q} w_i \varepsilon_{t-i} + \varepsilon_t$$

Autoregressive Integrated Moving Average

ARIMA Cheat Sheet

	AR	I	MA
Stands for:	Autoregressive	Integrated	Moving Average
Summary:	Regress future values on **past values**.	Differences our Y variable.	Regress future values on **past errors**.
Looks Like:	$\beta_0 + \sum_{k=1}^{p} \beta_k Y_{t-k}^{(d)}$	$Y_t^{(d)}$	$\beta_0 + \sum_{i=1}^{q} w_i \varepsilon_{t-i} + \varepsilon_t$
Purpose:	Long-term trends.	Ensure stationarity.	Sudden shocks.
Hyperparameter:	***p***	***d***	***q***
Find good value of hyperparameter by:	GridSearch	Augmented Dickey-Fuller Test	GridSearch