به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

گزارش پروژه دوم (تخمین سری زمانی)

درس یادگیری ماشین آماری

مهدی طاهراحمدی ۹۲۳۱۰۴۲

استاد: دکتر نیک آبادی

- **شرح مساله:**

مساله تحلیل یک سری زمانی مربوط به اطلاعات میزان مصرف برق بر حسب مگاوات یکی از ایالتهای کشور است که این مجموعه داده شامل 128616 نمونه میباشد که هر نمونه میزان مصرف انرژی برق در یک ساعت را مشخص میکند. این مجموعه داده تمامی ساعات بازهی زمانی ژانویه سال 2002 تا دسامبر سال2016 را در شامل میشود.

هدف ازین بررسی پیش‌بینی مصرف در سال های بعدی است وبدین منظور دو مدل آماری پارامتری و غیر پارامتری معرفی شده و نتایج و نمودار های حاصل گزارش شده است. همچنین توسط معیار MAE برای مدل های ارائه شده گزارش شده است و درباره آنها بحث و نتیجه گیری به عمل آمده است.

- **کار های گذشته:**

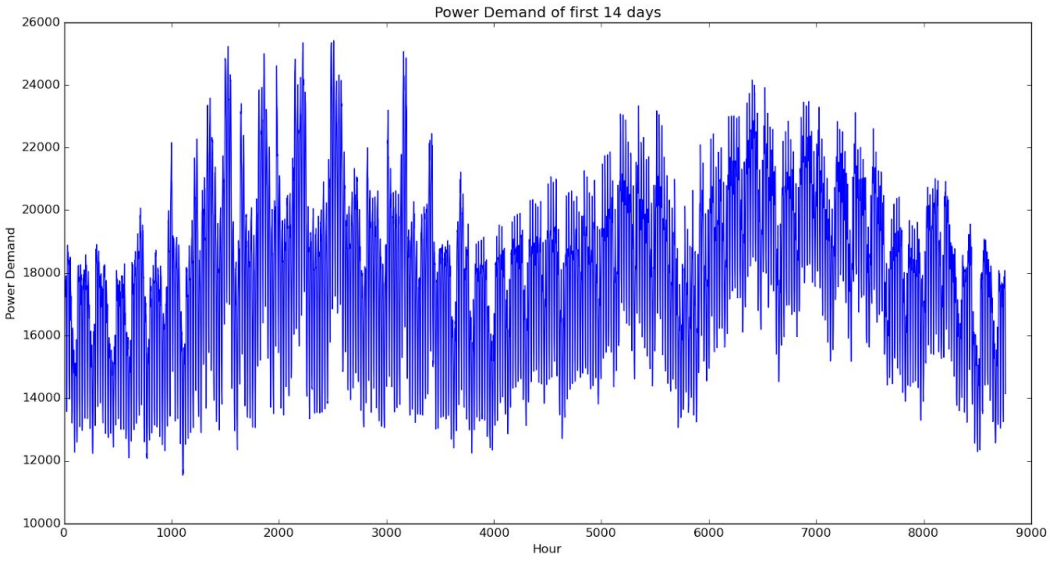
- **مدل ارائه شده، ارزیابی و گزارش نتیجه:**

در این بخش ابتدا داده هارا به دو بخش تقسیم کردیم، بخش اول ده سال برای تست و از 4 سال بعدی برای ارزیابی استفاده کردیم.

این داده ها چند ویژگی مهم داشتند که در انتخاب مدل به ما کمک میکرد.

1- با رسم نمودار داده ها در طول زمان شاهد الگوی تکرار متناوب داده ها هستیم. به این معنی که این داده های سری زمانی مانند که زنجیر مارکف با حالات جاذب و متناهی میمانند.

2- واریانس این داده ها در طول زمان وابسته به زمان نیست.



مشاهده مشود که داده ها در طول یک سال و در طول هر هفته و هر روز الگوی متناوب دارند.

- **غیر پارامتری (Empirical)**

مدل ارائه شده در این بخش بر اساس ترتیب زمانی و الگوی تناوبی مشاهده شده در داده های تست، ابتدا بازه های تناوب را روز های سال در نظر گرفتیم و داده های تست را به 14 بخش (14 سال) تقسیم کردیم وتوزیع هر بازه از هر داده نمونه را توسط یک تابع Emprical از آماره میانگین که مطابق روابط زیر تعریف میشود:

$$\begin{aligned}\widehat{F}_n(x) &= \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n} \\ &= \frac{\text{number of observations less than or equal to } x}{n} \end{aligned} \tag{8.1}$$

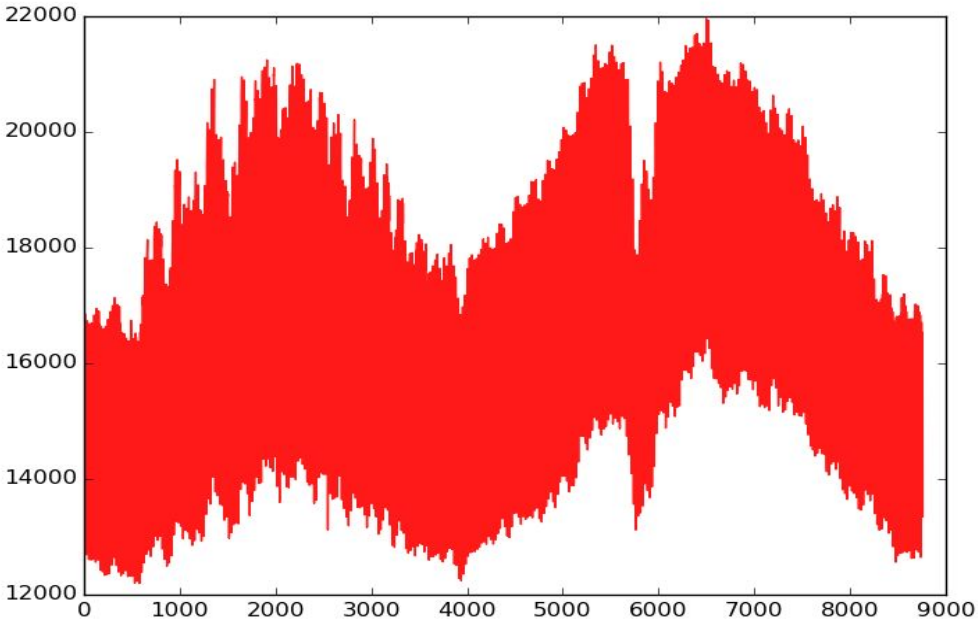
where

$$I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{if } X_i > x. \end{cases}$$

The plug-in estimator for linear functional $T(F) = \int r(x)dF(x)$ is:

$$T(\widehat{F}_n) = \int r(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i). \tag{8.2}$$

تخمین زدیم. نتیجه میانگینی از داده های هر سال داده های یک سال را تشکیل میدهند به طوری که تخمین مناسبی برای سال های بعدی بدست میدهد. نمودار زیر تخمین یک سال از داده هارا نشان میدهد.



معیار ارزیابی meas absolute error برای این مدل به شرح زیر است:

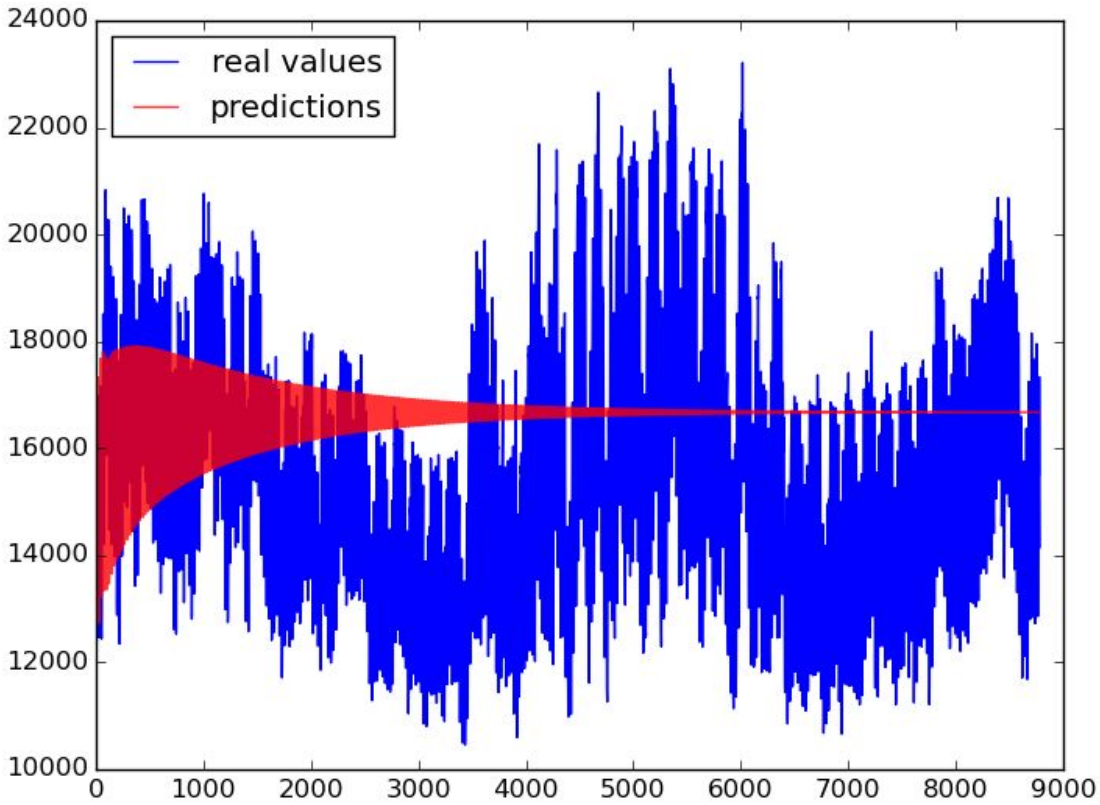
Test MAE: 2202.31586758

- **پارامتری (auto-regression)**

در این بخش برای مدل پارامتری از مدل auto-regression استفاده کردیم. این مدل همانند مدل های رگرسیون بر داده های آموزش تابع چند جمله ای را برازش میکند که پارامتر های این تابع پارامتر های مدل ما هستند.

در پیاده سازی این مدل از توابع کتابخانه ای scikit-learn استفاده شده و داده های lagged به عنوان ورودی استفاده شده. داده های lagged به منظور تعیین پارامتر های تابع و متغییر مستقل با استفاده از ورودی های بازه ای مشخص که حاوی اطلاعات بیشتری است و بازه ی قبلی نسبت به بازه پیش‌بینی استفاده میشود. در این بخش از 48 ساعت قبل به عنوان داده های لگ دار استفاده شده است.

مشاهده میشود که تابع رگرسیون تخمین زده شده نسبت به داده های اصلی چگونه است:



مشاهده میشود که پیشبینی تابع رگرسیون به مرور به یک مقدار همگرا میشود. این به این دلیل است که علیرغم تناوبی بودن داده ها، تابع رگرسیون متناوب نیست و نهایتا پس از گرفتن داده های ورودی در سری زمانی به یک مقدار مشخص میل میکند.

نتایج خروجی ضرایب مدل رگرسیون:

Coefficients:

```
[ 1.05089478e+02  1.53548592e+00 -7.06857051e-01  1.61176202e-01
 -2.05667692e-02  2.46889139e-03  2.34921094e-02 -6.23368592e-02
  4.34941830e-03  2.90889001e-02  2.40792272e-02 -1.13044649e-02
  7.60726739e-03 -3.70153848e-03  3.09666415e-02 -6.34042850e-02
 -3.02398616e-02  3.40218671e-02  4.71529539e-02 -3.28397451e-02
 -1.20212906e-03  3.42864505e-03 -3.80952226e-02  1.71738472e-01
  3.71835994e-01 -6.77680495e-01  1.81793567e-01  1.99993591e-02
 -4.44623250e-03 -2.85554055e-02  8.31130636e-03  3.70495850e-02
 -1.08416094e-02 -2.88249119e-02  1.29939927e-02 -2.22181867e-02
  5.92903803e-03 -4.03002478e-04  2.66410606e-02 -2.06061444e-02
  4.33187101e-02 -3.59451487e-02 -1.40834671e-02  1.52407275e-02
  1.29582732e-02  1.15131876e-02 -3.98603288e-02 -3.64059604e-02
  2.38543966e-01 -3.45181762e-01  2.10295265e-01 -4.31493775e-02
 -1.73215730e-02  2.38308055e-02 -3.41973641e-02  5.52949270e-03
  1.43723187e-02 -1.70939806e-02  3.11305584e-02 -1.98569899e-02
  8.99309214e-03 -1.35472106e-02 -5.38887338e-03 -1.35331049e-02
  3.53067521e-02 -1.47891258e-02 -6.66026739e-03  2.16203571e-02
 -7.96042791e-03 -1.53484208e-02  4.23323458e-02 -1.64503762e-02]
```

و معیار ارزیابی meas absolute error برای این مدل به شرح زیر است:

Test MAE: 2098.502

- **جمع بندی و نتیجه گیری**

با پیاده سازی و استفاده از دو مدل برای پیش بینی داده های یک سری زمانی بر خلاف اینکه انتظار داشتیم مدل پارامتری بهتر از مدل غیر پارامتری عمل کند. اما به نظر میرسد به دلیل انتخاب مدل پارامتری نامناسب و خاصیت تناوبی در داده های مساله، به این مقصود نرسیدیم.
همچنین همانطور که انتظار داشتیم مدل پارامتری باری تخمین یک سری زمانی نسبتا پیچیده و داده های زیاد باید مدل پیچیده و پارامتر های زیاد باشد و مدل غیر پارامتری ساده تر و بهینه تر عمل کند که مشاهده شد مدل غیر پارامتری در عین سادگی نتایج خوبی بدست داد.

- **مراجع**

[1].Larry Wasserman, All of Statistics: A Concise Course in Statistical Inference, Springer, 2010.
[2].http://www.uow.edu.au/student/qualities/statlit/module3/5.4interpret/index.html
[3].Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables Luke Keele, Oxford. ttp://www.nuffield.ox.ac.uk/politics/papers/2005/Keele%20Kelly%20LDV.pdf
[4].Time Series Analysis using Python.
https://github.com/rouseguy/TimeSeriesAnalysiswithPython
[5].Time Series Analysis in Python with statsmodels.
http://204.236.236.243/scipy2011/slides/mckinney_time_series.pdf

---