به نام خدا

توضیحات:

- پاسخ به تمرین‌ها باید به صورت انفرادی صورت گیرد. حداقل برخورد با پاسخ‌های مشابه، تخصیص نمره کامل منفی به طرفین خواهد بود.
- پاسخ‌های خود را به زبان فارسی و به صورت مرتب، در قالب یک فایل فشرده (zip.) با الگوی زیر در صفحه‌ی درس بارگذاری نمایید:

DM_HW[No]_[Student_number].zip

- لطفا نظم، ساختار و توالی سوالات را در پاسخ‌ها رعایت کنید.
- تمیزی و خوانایی گزارش تمرین از اهمیت بالایی برخوردار است.
- برای تمرین‌های پیاده‌سازی، علاوه بر کد، گزارش کتبی نیز ارسال کنید.
- سوالات و ابهامات خود در رابطه با بخش نوشتاری را با ایمیل های: mohamad.tavakoli7878@gmail.com و یا arminzd@aut.ac.ir و در رابطه با بخش پیاده‌سازی با ایمیل: ariamostajeran99@gmail.com مطرح کنید.

طراحی تمرین:

آقایان ذوالفقاری، توکلی و مستاجران

استاد درس:

دکتر امیرمزلقانی

بخش نوشتاری

۱. داده‌های زیر را که مربوط به سن افراد هستند، به روش نمودار جعبه‌ای[1] نمایش دهید.

25, 35, 25, 30, 13, 16, 15, 16, 19, 20, 25, 22, 25, 21, 22, 35, 20, 70, 46, 40, 33, 35, 33, 35, 52, 45, 36

۲. در ابتدا داده‌های نویزی و داده‌های پرت را تعریف کنید. سپس در رابطه با داده‌های نویزی و داده‌های پرت به سوالات زیر با دلیل پاسخ دهید.

الف) آیا داده‌های نویزی و داده‌های پرت، مطلوب هستند؟ (با مثال توضیح دهید)

ب) آیا داده‌های نویزی همیشه جزو داده‌های پرت محسوب می‌شوند؟

ج) آیا داده‌های پرت همیشه جزو داده‌های نویزی محسوب می‌شوند؟

۳. برای بردارهای داده شده، موارد خواسته شده را بدست بیاورید.

a. x = [1, 1, 1, 1] , y = [2, 2, 2, 2]
   Cosine similarity, Correlation, Euclidean distance
b. x = [0, 1, 0, 1] , y = [1, 0, 1, 0]
   Cosine similarity, Correlation, Euclidean distance, Jaccard distance
c. x = [1, 1, 0, 1, 0, 1] , y = [1, 1, 1, 0, 0, 1]
   Correlation, Manhattan distance
d. x = [2, -1, 0, 2, 0, -3] , y = [-1, 1, -1, 0, 0, -1]
   Cosine similarity, Correlation

٤. در رابطه با شباهت کسینوسی[2] و میزان همبستگی[3] به سوالات زیر پاسخ دهید.

الف) محدوده مقادیر ممکن برای شباهت کسینوسی چقدر است؟

ب) اگر شباهت کسینوسی دو شئ برابر یک باشد، آیا آنها یکسان هستند؟

ج) آیا رابطه‌ای بین شباهت کسینوسی و میزان همبستگی وجود دارد؟

د) فرض کنید میزان همبستگی بین دو متغیر، صفر است. مفهوم آن چیست؟ با توجه به تعریف متغیرهای مستقل، آیا این متغیرها مستقل هستند؟

---

[1] Box plot
[2] Cosine similarity
[3] Correlation

۵. الف) درمورد نمودار Quantile و نمودار Quantile-Quantile تحقیق کنید و توضیح دهید که هر کدام از نمودارها چه اطلاعاتی را نمایش می‌دهند، سپس دو نمودار را با هم مقایسه کنید.

ب) داده‌های زیر مربوط به سن و میزان چربی افرادی هستند که به صورت تصادفی انتخاب شده‌اند. نمودار Quantile-Quantile را برای داده‌های زیر رسم کنید و بر اساس مقادیر Q1، Median و Q3 آن را تحلیل کنید.

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

۶.

الف) یکی از راه های ارزیابی مدل بدست آمده توسط رگرسیون، استفاده از روش LOOCV[4] می‌باشد. این روش را توضیح دهید.

ب) در مسئله رگرسیون خطی زیر($Y = bX + c$)، میانگین مربعات خطا[5] را در روش LOOCV بدست آورید (X متغیر مستقل و Y متغیر وابسته است).

X	Y
1	1
3	1
1	3

[4] Leave-one-out cross-validation

[5] Mean Square Error

۷. در رابطه با رگرسیون به سوالات زیر پاسخ دهید:

الف) آیا امکان رخداد بیش‌برازش[۶] در مسائل رگرسیون خطی وجود دارد؟

ب) در حل مسائل با روش رگرسیون، آیا حذف داده های پرت[۷] ضروری است؟ راه های تشخیص داده های پرت را توضیح دهید.

ج) کم بودن مقدار انحراف باقیمانده ها[۸] از مدل رگرسیون لزوما بیان کننده خوب بودن مدل است؟ چرا؟

د) برای حل مسئله کمترین مربعات میتوان از روش مستقیم یا روش های بهینه سازی استفاده کرد. روش مستقیم حل این مسئله را با روش بهینه سازی Gradient Descent مقایسه کنید.

۸. در رابطه با کاهش بعد[۹] به سوالات زیر پاسخ دهید.

الف) انتخاب ویژگی[۱۰] و استخراج ویژگی[۱۱] دو راه مورد استفاده برای کاهش بعد می‌باشند. هر کدام از این روش ها را توضیح دهید.

ب) از روش های مورد استفاده برای استخراج ویژگی، میتوان به LDA[۱۲] و PCA[۱۳] و Autoencoder ها اشاره کرد. هر کدام از این روش ها را توضیح دهید.

۹. سایت IMDb در یک نظرسنجی، علاقه کاربران خود را به سه ژانر کمدی، درام و ترسناک مورد بررسی قرار داد. در این نظرسنجی، افراد شرکت کننده در دو دسته معمولی و یا منتقد قرار میگیرند و تعداد افراد شرکت کننده در این نظرسنجی ۱۰۰۰۰ نفر می‌باشد. با توجه به نتیجه این رای گیری به سوالات زیر پاسخ دهید.

تعداد افراد	مهارت	ژانر
1000	منتقد	کمدی
3000	معمولی	کمدی
1000	منتقد	درام
2000	معمولی	درام
2000	منتقد	ترسناک
1000	معمولی	ترسناک

[6] Overfitting
[7] Outliers
[8] Residuals
[9] Dimensionality Reduction
[10] Feature Selection
[11] Feature Extraction
[12] Linear Discriminant Analysis
[13] Principal Component Analysis

الف) مقدار Entropy را برای هر یک از متغیر های ژانر و مهارت بدست آورید.

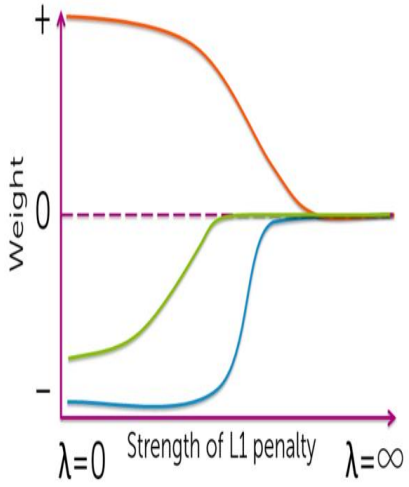
ب) مقدار Mutual Information دو متغیر مهارت و ژانر را محاسبه کنید.

۱۰. در رابطه با منظم‌سازی در رگرسیون به سوالات زیر پاسخ دهید:

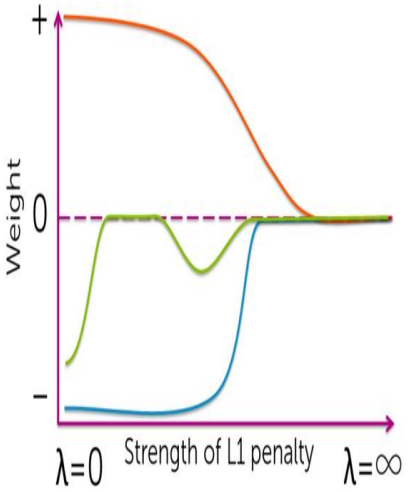
الف) فرض کنید از منظم‌سازی Ridge استفاده میکنیم. مدل را دو بار با مقادیر $\lambda = 0$ و $\lambda = 1$ آموزش میدهیم. پارامتر های بدست آمده $\boldsymbol{\theta} = \begin{bmatrix} \mathbf{1.21} \\ \mathbf{0.57} \end{bmatrix}$ و $\boldsymbol{\theta} = \begin{bmatrix} \mathbf{71.9} \\ \mathbf{44.42} \end{bmatrix}$ هستند. با دلیل توضیح دهید کدام پارامترها متعلق به $\lambda = 0$ و کدام متعلق به $\lambda = 1$ می‌باشد؟

ب) با توجه به نمودارهای زیر، توضیح دهید کدام نمودار ها نمی‌توانند متعلق به مدل Lasso باشند. چرا؟ (محور افقی مربوط به مقادیر مختلف ضریب λ - محور عمودی مربوط به مقادیر مختلف پارامتر های مدل و نمودار های رنگی نیز مربوط به ۳ پارامتر یک مدل هستند)

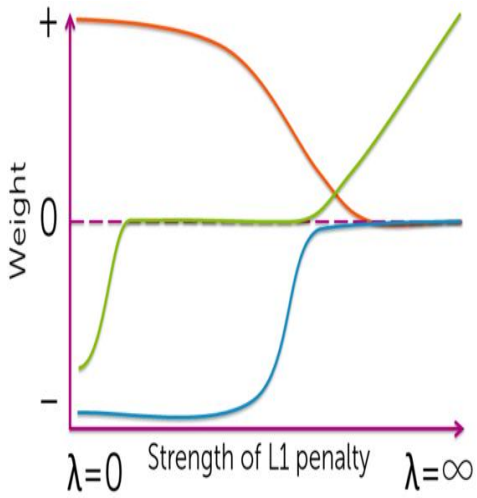
نمودار ۱



نمودار ۲



نمودار ۳



بخش پیاده‌سازی

بخش اول : پیش‌پردازش

پیش‌پردازش داده ها برای مدل‌های یادگیری ماشینی یک مهارت اصلی برای هر دانشمند داده یا مهندس یادگیری ماشین است. در یک پروژه علم داده در دنیای واقعی، پیش پردازش داده‌ها یکی از مهم ترین گام های آن است و یکی از عوامل مشترک موفقیت یک مدل است، یعنی اگر پیش پردازش داده ها و مهندسی ویژگی‌ها درست باشد، احتمال موفقیت آن مدل در مقایسه با مدلی که داده ها برای آن به خوبی پیش پردازش نشده اند، بیشتر است و نتایج بهتری تولید خواهد کرد.

تمرکز ما در این بخش بر روی کار با کتابخانه pandas میباشد.

برای انجام راحت‌تر تمرین میتوانید از Jupyter notebook و یا Google colab استفاده کنید.

۰- مجموعه داده:

مجموعه داده در نظر گرفته شده برای تمرین شما مجموعه داده IMDB-Movie-Dataset می‌باشد. این مجموعه برترین فیلم‌های IMDB و اطلاعات آن‌ها میباشد. هر فیلم ۱۲ ویژگی دارد که تعدادی از آن‌ها عبارت است از:

- Rank : رتبه فیلم
- Year : سال ساخت فیلم
- Runtime : مدت فیلم به دقیقه

برای خواندن مجموعه داده می‌توانید از کد زیر استفاده کنید.

```
import pandas as pd

df = pd.read_csv("IMDB-Movie-Data.csv")

df.head(10)
```

دستور head یک ورودی عدد صحیح گرفته و اولین سطرهای مجموعه داده را برمیگرداند. مقدار پیش فرض آن ۵ است.

۱- اهمیت دادگان از دست رفته:

یک عبارت معروف در یادگیری ماشینی وجود دارد که ممکن است آن را شنیده باشید:

Garbage in, Garbage out

اگر مجموعه داده های شما مملو از NaN و مقادیر زباله باشد ، مطمئناً مدل شما نیز نتیجه‌ی قابل قبولی ندارد. بنابراین مقابله با چنین داده هایی مهم است. در ابتدا در داده‌های خود به دنبال داده‌های NaN بگردید. برای اینکار میتوانید از تابع ()isna استفاده کنید.

یک روش برای پر کردن مقادیر از دست رفته، پرکردن آن با میانگین، میانه، واریانس آن ستون یا مقداری ثابت است. برای انجام این کار، میتوانیم ازکتابخانه‌ی sklearn و تابع Simple Imputer استفاده کنیم.

سوال ۱) در ابتدا بیابید که کدام ستون‌ها دارای مقدار NaN هستند(در مورد روش کوئری زدن در dataframe ها جست‌وجو کنید). حال سطرهایی که در آن هر دو ستون پیدا شده دارای مقدار NaN هستند را بیابید. از آنجایی که تعداد این سطرها محدود است، تمامی آنها را با استفاده از دستور dropna حذف کنید.

سوال ۲) حال یکی از ۲ ستون دارای مقدار NaN را انتخاب کرده و مقادیر NaN آن را با میانگین جایگزین کنید.

۲- تغییر مقادیر داده:

گاهی اوقات در داده‌ها خطاهایی وجود دارد که نیاز است به صورت دستی رفع شوند. همانطور که با مشاهده در دادگان مشهود است، رتبه فیلم‌ها به اشتباه شماره‌گذاری شده است.

سوال ۳) دادگان را دوباره از ابتدا بخوانید. در ابتدا با دستور sort_values مجموعه را نسبت به نمره imdb مرتب کنید، و سپس rank را دوباره و به صورت صحیح مقدار دهی کنید. برای اینکار چندین روش ساده و پیچیده وجود دارد که به دلخواه میتوانید از هرکدام استفاده کنید.

۳- نرمالسازی:

از آزمایشهای مشخصاتی ثابت شده است که مدل‌های یادگیری ماشین و یادگیری عمیق در مقایسه با مجموعه دادهای که نرمال‌سازی نشده‌اند، در یک مجموعه داده نرمال شده عملکرد بهتری دارند. هدف نرمال سازی تغییر مقادیر به یک مقیاس مشترک است. چندین راه برای این کار وجود دارد

سوال۴) با استفاده از Standard Scaler در sklearn.preprocessing اقدام به نرمال‌سازی ستون meta score کنید. دقت کنید پیش از نرمالسازی دادگان NaN را جایگزین کرده باشید. مقدار واریانس و میانگین ستون را قبل و بعد از نرمالسازی ذکر کنید.

۴- بررسی داده:

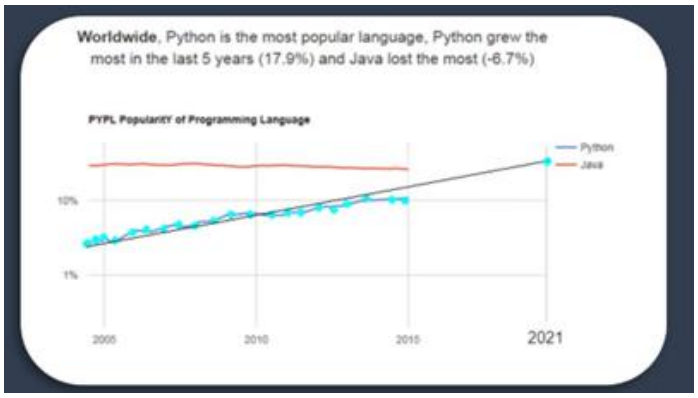
در این بخش میخواهیم یاد بگیریم که چطوری برخی از داده‌های خاص را پیدا کنیم. اطلاعات زیر را بیابید.

- تمامی فیلم‌هایی که کارگردان آن Christopher Nolan است را پیدا کنید.
- فیلم‌هایی که نمره بین ۸,۴ و ۸,۶ دارند را گزارش کنید.
- نموداری نقطه‌ای رسم کنید که نشاندهنده‌ی نمره متا اسکور بر اساس نمره IMDB باشد.
- تمامی فیلم‌هایی که قبل از ۲۰۱۳ ساخته شده و کوتاهتر از ۱۰۰ ولی بیشتر از ۸۵ دقیقه هستند را بیابید.

بخش دوم : رگرسیون

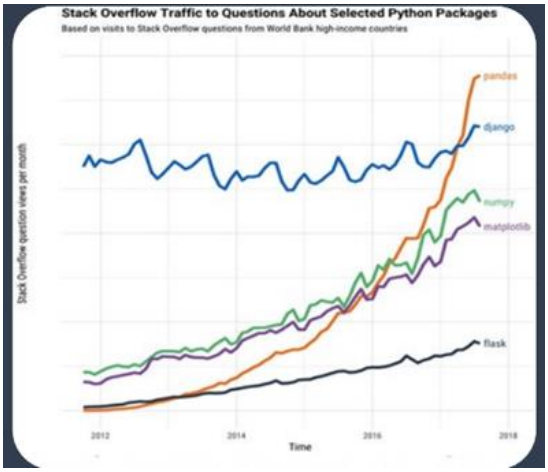
رگرسیون خطی:

در نظر بگیرید میخواهیم با داشتن میزان محبوبیت زبان پایتون بین سالهای ۲۰۰۵ تا ۲۰۱۷ محبوبیت آن را در انتهای سال ۲۰۲۱ پیشبینی کنیم.



تصویر فوق یک نمودار از پیش رسم شده برای مقایسه محبوبیت ها در سالهای مختلف میباشد . همانطور که از شکل پیداست، نمیتوان یک رابطه‌ی خطی یافت که داده های ما دقیق پیشبینی شود اما میتوان خطی را پیدا کرد که کمترین فاصله را از مجموعه نقاط داشته باشد و در این حالت امیدوار بود که تخمین ما از محبوبیت پایتون دارای کمترین خطا باشد. در اصل امیدواریم رابطه‌ی بین متغیر زمان و محبوبیت پایتون یک رابطه خطی باشد و به طور خطی پیشرفت کند.

رگرسیون غیرخطی:



عکس بالا مقایسه تعداد سوالات پرسیده شده در مورد هر یک از پکیج های معروف پایتون است. فرض کنید میخواهیم تعداد سوالات پرسیده شده در مورد پکیج پاندا را در سال ۲۰۲۱ پیشبینی کنیم. نمودار نشان میدهد که انتخاب کردن یک رابطه‌ی خطی نزدیک به داده‌ها نتیجه نزدیکی به واقعیت نخواهد داشت و هر چه بیشتر از داده‌های واقعی فاصله بگیریم خطای پیشبینی بیشتر میشود اما به نظر میرسد رابطه آن میتواند نزدیک به یک چندجمله ای از درجه دو باشد و با یافتن همچین چندجمله‌ای میتوان تخمین هایی با خطای کمتری داشت.

پس باید چندجمله ای زیر را به نحوی پیدا کنیم که به مقادیر واقعی نزدیکترین حالت را داشته باشد:

$$y(t) = b_0 + b_1 t + b_2 t^2$$

و این به این معناست که در حالت ایده آل بودن نمودار انتظار داریم که:

$$\begin{aligned} y(t_1) &= b_0 + b_1 t_1 + b_2 t_1^2 \\ y(t_2) &= b_0 + b_1 t_2 + b_2 t_2^2 \\ &\vdots \qquad\qquad \vdots \\ y(t_n) &= b_0 + b_1 t_n + b_2 t_n^2 \end{aligned}$$

و اگر معادلات بالا را به فرم ماتریسی بنویسیم:

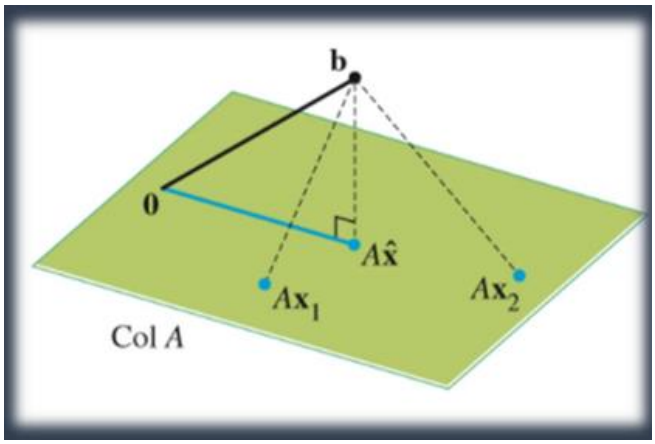
$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ 1 & t_n & t_n^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

با جایگذاری کردن زمان و تعداد سوالات، ماتریس را حل کرده و ضرایب دقیق را بدست می‌آوریم. ولی همانطور که قبلا گفته شد، این تنها برای حالت ایده آل صدق میکند . برای حالت‌های دیگر باید ضرایبی را بدست بیاوریم که نزدیک‌ترین حالت را به نمودار اصلی داشته باشند و برای این کار باید از مسئله کمترین مربعات یا *Least Square* کمک بگیریم.

مسئله کمترین مربعات:

ماتریس $Ax = b$ را در نظر بگیرید. حالتی را در نظر بگیرید که مقدار x برای پاسخ یافت نشود. در این حالت ما به دنبال نزدیک‌ترین بردار به b هستیم که در این معادله صدق کند.

تصویر b روی A نزدیک‌ترین بردار می‌باشد که اگر با بردار Ax که همان تصویر است آن را نمایش دهیم خواهیم داشت :



واضحا بردار Ax - b بر تمامی بردارهای صفحه Col A عمود است و این یعنی :

$$
\begin{aligned}
& a_j^T(b - A\hat{x}) = 0 \\
\Rightarrow\ & A^T(b - A\hat{x}) = 0 \\
\Rightarrow\ & A^T b - A^T A\hat{x} = 0 \\
\Rightarrow\ & A^T A\hat{x} = A^T b
\end{aligned}
$$

برای حل این معادله میتوانید از این تابع آماده استفاده کنید.

شرح تمرین :

یک فایل از سهام گوگل در روز های متخلف در اختیار شما قرار گرفته است. این فایل به صورت CSVمی‌باشد و شامل هفت ستون می‌باشد. این هفت ستون مختلف به ترتیب نشان‌دهنده تاریخ روز، شروع قیمت سهام در ابتدای روز، بالاترین قیمت سهام در یک روز، کمترین قیمت سهام در یک روز، قیمت سهام در انتهای روز، حجم معاملات و نام سهام می‌باشد. در این پروژه قصد داریم قیمت سهام گوگل را در ابتدای سال ۲۰۰۶ تا پایان سال ۲۰۱۷ بررسی کنیم. (چون از رگرسیون تک متغیره استفاده می‌کنیم، تعدادی از داده‌ها حذف شدند تا نمودار نهایی برای شما ملموس‌تر باشد)

نحوه انجام تمرین :

ابتدا فایل *CSV* را دانلود کرده و آن را با استفاده از کتابخانه pandas بخوانید. ستون دیتای مورد نظر Open میباشد.

به غیر از ده سطر آخر فایل، از تمامی سطرها برای بدست آوردن ضرایب معادلات استفاده کنید. سپس با توجه به ضرایبی که بدست آوردید از ده روز آخر برای بررسی خطای تخمین خود استفاده کنید و آن را نمایش دهید.

در این مرحله اول کد شما رگرسیون خطی را بررسی میکند. برای این منظور باید معادله زیر حل شود که t همان روز شماست.

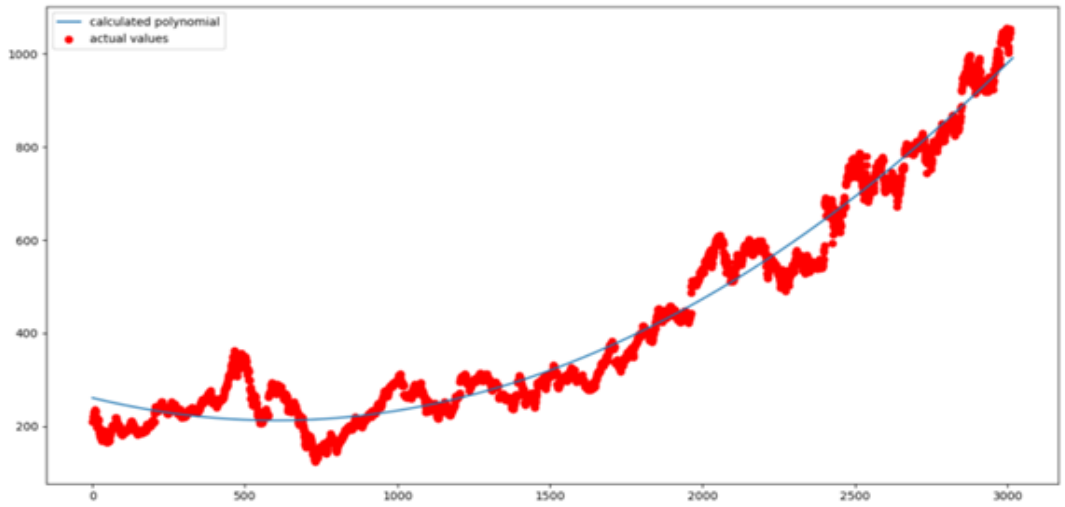
$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

سپس برای ۱۰ روز آخر خطا را بررسی کرده و به فرمت زیر آن را نشان دهید.

```
calculated value: 986
actual value: 1060.09
error: −74.09
```

در گام بعدی سراغ رگرسیون درجه ۲ رفته و تمامی مراحل بالا را برای درجه ۲ تکرار میکنیم.

حال با بررسی میزان خطاهای هر کدام تشخیص می‌دهیم که کدام یک از رگرسیون‌ها مناسب داده‌های ما می‌باشد و در آخر نمودار مقادیر تخمینی و مقادیر واقعی مربوط به رگرسیون بهتر را به شکل زیر نمایش میدهیم:



برای نمایش نمودار از کتابخانه matplotlib استفاده کنید.

نکات تکمیلی:

- از هر روشی برای انجام تمارین می‌توانید استفاده کنید. نکته اصلی استفاده از کتاب‌خانه pandas بوده و کتابخانه sklearn صرفا پیشنهاد می‌شود.
- برای قسمت پیاده‌سازی، علاوه بر کد، گزارش نیز نیاز است ارسال کنید. اگر از نوت‌بوک استفاده می‌کنید، در خود نوت‌بوک سوال‌های بخش پیاده‌سازی را پاسخ دهید و برای هر بخش، نتیجه کارتان را گزارش دهید. اگر از نوت‌بوک استفاده نمی‌کنید، سوال‌های بخش پیاده‌سازی و نتیجه را به صورت pdf همراه با بخش نوشتاری ارسال کنید.