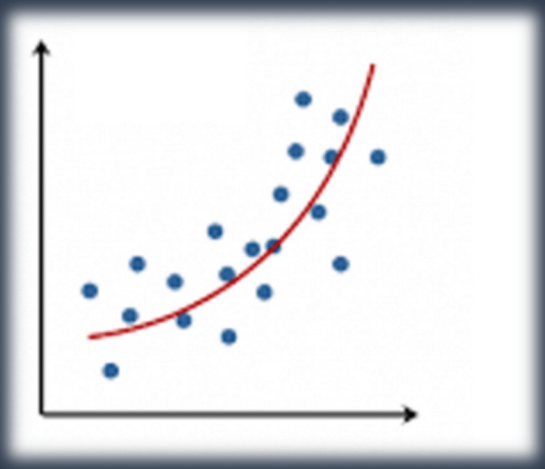
مقدمه و توضیح الگوریتم

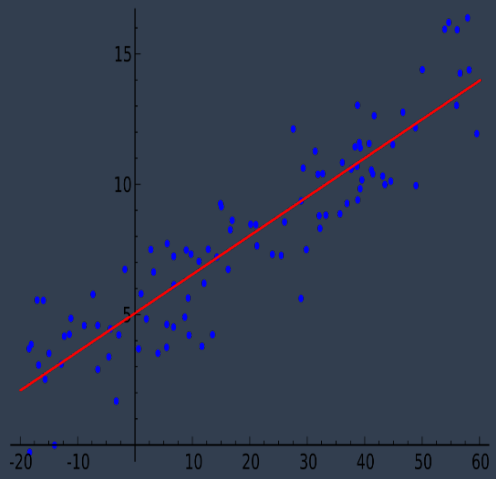
"پیش‌بینی می‌شود در ابتدای سال آینده، تعداد بیماران شناسایی شده به ۱۰۰ نفر در روز برسد"

احتمالا شما هم در این چند وقت از جملاتی که خبر از پیش‌بینی موارد ابتلا به کرونا بدهد شنیده‌اید یا در برخی موارد نمودارهایی مربوط به آن را مشاهده کرده باشید. اما دقیقا مبنای این پیش‌بینی‌ها چیست؟

پیش‌بینی کردن همیشه یکی از دغدغه‌های مهم انسان بوده است. امروزه نیز بسیاری از فعالیت‌های ما نیاز به پیش‌بینی کردن شرایط دیده می‌شود. از معاملات روز بازار بورس گرفته تا بررسی شرایط آب و هوایی در روزهای آینده. برای تخمین درست در قدم اول باید روابط بین متغیرها مشخص باشد.

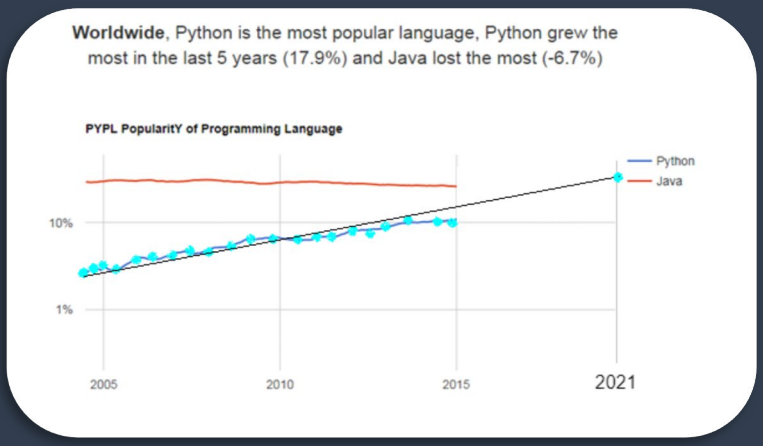
در این مینی پروژه قصد داریم با توجه به دانشی که از جبرخطی کسب کرده ایم، به مدل‌سازی روابط بین متغیر ها بپردازیم.





رگرسیون:

رگسیون 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 محبوبیت‌ها در سال‌های مختلف می‌باشد . همانطور که از شکل پیداست، نمی‌توان یک رابطه‌ی خطی یافت که داده‌های 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 پکیج‌های معروف پایتون است.

فرض کنید می‌خواهیم تعداد سوالات پرسیده شده در مورد پکیج پاندا را در سال ۲۰۲۱ پیش‌پینی 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 چندجمله‌ای از درجه دو باشد و با یافتن همچین چندجمله‌ای میتوان تخمین‌های با خطای کمتری داشت.

پس باید چندجمله‌ای 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 ایده‌آل 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:

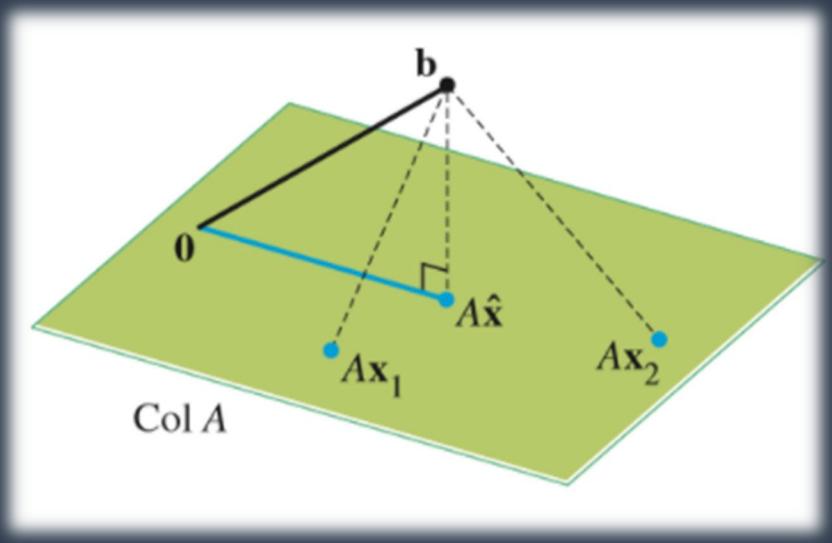
$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ 1 & t_n & t_n^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

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 ایده‌آل صدق می‌کند . برای حالت‌های دیگر باید ضرایبی را بدست بیاوریم که نزدیک ترین 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 نزدیک ترین بردار به b هستیم که در این معادله صدق کند.

تصویر b روی A نزدیک ترین بردار می‌باشد که اگر با $A\hat{x}$ آن را نمایش دهیم خواهیم داشت :



در این مثال $col\ A$ یک صفحه در فضای سه بعدی است و نزدیک‌ترین جواب $A\hat{x}$ می‌باشد.

و واضح است که $b - A\hat{x}$ به تمامی بردار های $Col\ A$ عمود است و این یعنی خواهیم داشت:

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

از معادله بالا ما به دنبال $\hat{x}$ هستیم که ضرایب چندجمله‌ای ما (در اینجا درجه ۲) می‌باشد.

برای حل دستگاه معادلات بالا می‌توانید از پروژه اول خود استفاده کنید و یا از این تابع آماده استفاده کنید.

شرح پروژه :

یک فایل از سهام گوگل در روز های متخلف در اختیار شما قرار گرفته است.

این فایل به صورت csv می‌باشد و شامل هفت ستون می‌باشد. این هفت ستون مختلف به ترتیب نشان‌دهنده تاریخ روز، شروع قیمت سهام در ابتدای روز، بالا ترین قیمت سهام در یک روز، کمترین قیمت سهام در یک روز، قیمتی 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 $regression$ تک متغیره استفاده می‌کنیم تعدادی از داده‌ها حذف شدند تا نمودار نهایی برای شما ملموس‌تر باشد)

نحوه انجام پروژه :

ابتدا فایل csv را دانلود کرده و آن را بخوانید. برای خواندن فایل csv می‌توانید از این لینک استفاده کنید (استفاده از کتابخانه و توابع دیگر بلامانع است) ستون دیتای مورد نظر ما $Open$ می‌باشد و ما به بررسی مقادیر این ستون در روزهای مختلف می‌پردازیم.

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 مرحله اول کد شما باید رگرسیون خطی را بررسی کند. برای رگرسیون خطی باید دستگاه معادله زیر حل شود که t شماره روز شماست و برای هر سطر یک واحد از سطر بالایی خود بیشتر است (روزها متوالی در نظر گرفته شده‌اند)

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

سپس برای ده روز آخر بررسی کنید که چقدر خطا داشتید و به ازای هر روز، در خروجی سه خط به فرمت زیر نمایش داده شود:

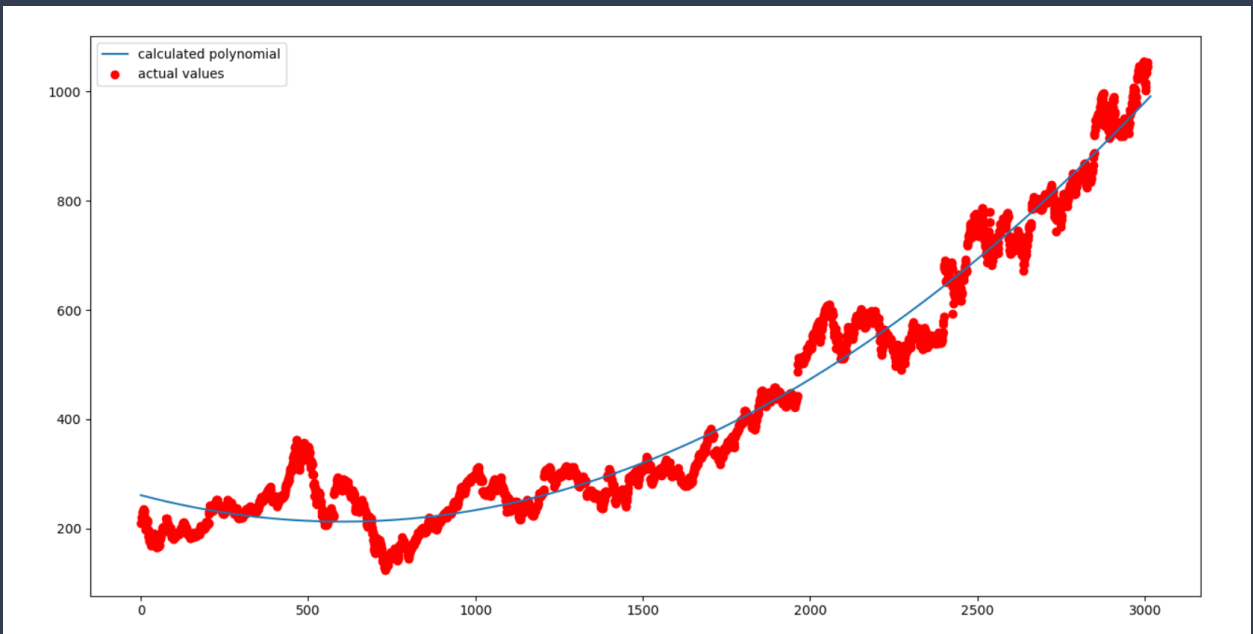
calculated value: 986

actual value: 1060.09

error: −74.09

در گام بعدی سراغ رگرسیون درجه ۲ می‌رویم و تمامی مراحل بالا (حل دستگاه معادله و محاسبه خطا) را برای این رگرسیون نیز انجام می‌دهیم.

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:



نقاط قرمز رنگ مقادیر واقعی و خط آبی رنگ مقادیر تخمین زده شده برای هر روز می‌باشد.

برای رسم نمودار نیز می‌توانید از توضیحات ویدیو استفاده کنید.

قوانین

۱. مینی پروژه به صورت انفرادی می‌باشد. تقلب‌ها توسط سامانه کوئرا چک می‌شود اما تصحیح توسط تیم تدریسیاری صورت می‌گیرد.

۲. برای خواندن فایل csv می‌توان از هر کتابخانه‌ای استفاده کرد.

۳. با توجه به مجازی بودن درس و عدم امکان تحویل حضوری، انتظار می‌رود کد پیاده‌سازی شده از توابع مختلف تشکیل شده باشد که فهم مطلب را آسان‌تر کند. وجود مستند برای توضیح کد به شدت استقبال می‌گردد.

۴. برای انجام عملیات ضرب ماتریس ها($dot()$) و ترانهاده($transpose()$) می‌توانید از توابع آماده استفاده کنید.

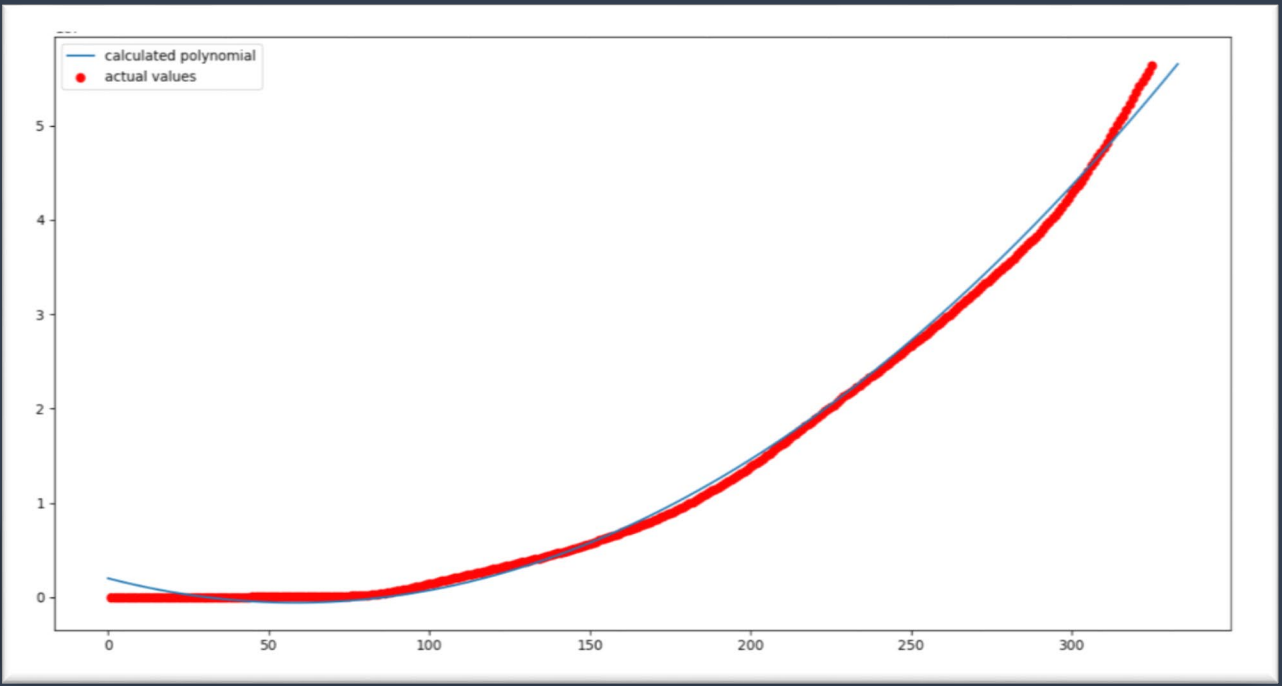
۵. فرمت فایل ارسالی بایستی به صورت $miniproject3\_STUDENTNUMBER\_Name$ باشد. به طور مثال:
$miniproject3\_9628099\_SoroushMehraban$

برای علاقه مندان

دیتای مبتلایان به کرونا را از روز اول تا به امروز می‌توانید از اینجا دریافت کنید. فایل $total\_cases.csv$ در دومین ستون خود حاوی تعداد مبتلایان جهانی کرونا می‌باشد و این آمار به صورت روزانه موجود است.

می‌توانید از رگرسیون درجه ۲ خود برای این دیتا استفاده کنید.

به طور تقریبی این نمودار حاصل می‌شود :



موفق باشید

تیم تدریس یاری جبر خطی کاربردی

آذر ۹۹