نام و نام خانوادگی: پیمان هاشمی

شماره دانشجویی: 400131032

تمرین: سری اول

بخش اول

سوال 1)

الف: نادرست – زیرا ممکن است دیتای مربوط به اموزش دارای همه انواع مقادیر کلاس داده نباشد و در این صورت خطای مربوط به آموزش کاهش پیدا میکند ولی در صورتی که خطای آزمون به شدت بالا رفته به دلیل اینکه دیتا تنها بر روی یک نوع خاصی دیتا train شده است.

ب: نادرست - با کاهش تعداد داده های آموزش الگوریتم under train میشود.

ج: نادرست – در مدل هایی که کم برازش شده اند به دلیل اینکه اندازه کافی واریانس ندارد باعث میشود خطای آزمون بالا باشد در این حالت یعنی مدل پیچیدگی لازم را ندراد. در این حالت باید پیچیدگی مدل را افزایش داد تا زمانی که خطا کم شده و خطا های مربوط به واریانس از بین بروند.

د: درست - به دلیل اینکه بر خلاف MSE یا RMSE به توان 2نمیرسد و در قدر مطلق قرار دارد.

سوال 2)

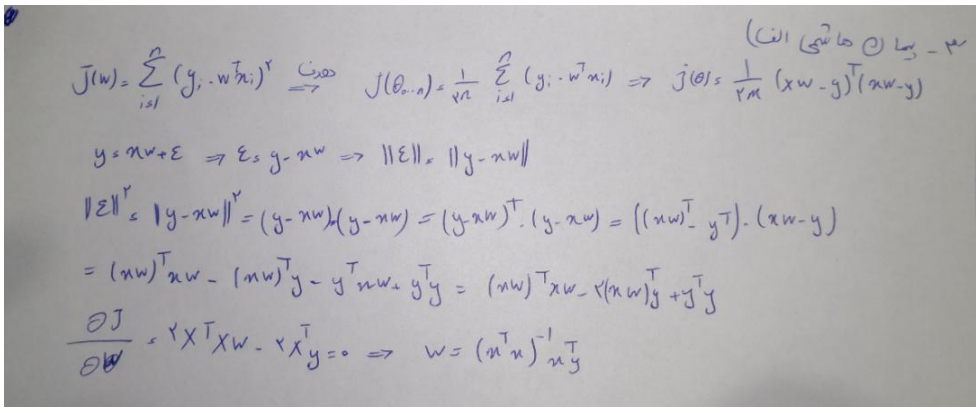
به این معنی میباشد که پیچیدگی مدل استفاده شده بسیار زیاد بوده و به اصطلاح مدل overfit شده است.

برای حل این موضوع میتوان در مزحله اول پیچیدگی مدل را کاهش داد.

باید توازن بین واریانس و بایاس برقرار باشد.

سوال 3)

الف)



ب)

یکی از مشکلات علامت است که آن را با ضرب داخلی حل میکنیم

در ضرب X*T^(X) زمانی که ستون ها بهم وابسته باشند مشکل ساز میشود که باید از PCA استفاده کنیم

زمانی تعداد ستون از تعداد داده ها بسیار بیشتر بیشتر باشد که نتواند به طور خوب پوشش بدهد

د)جمله منظم ساز به نویز حساس تر است

سوال 4)

الف: از گرادیان نزولی mini batch استفاده میکنیم زیرا در داده های بزرگ یا بیگ دیتا استفاده از معادله نرمال زمان بسیار زیادی را به خود اختصاص میدهد.

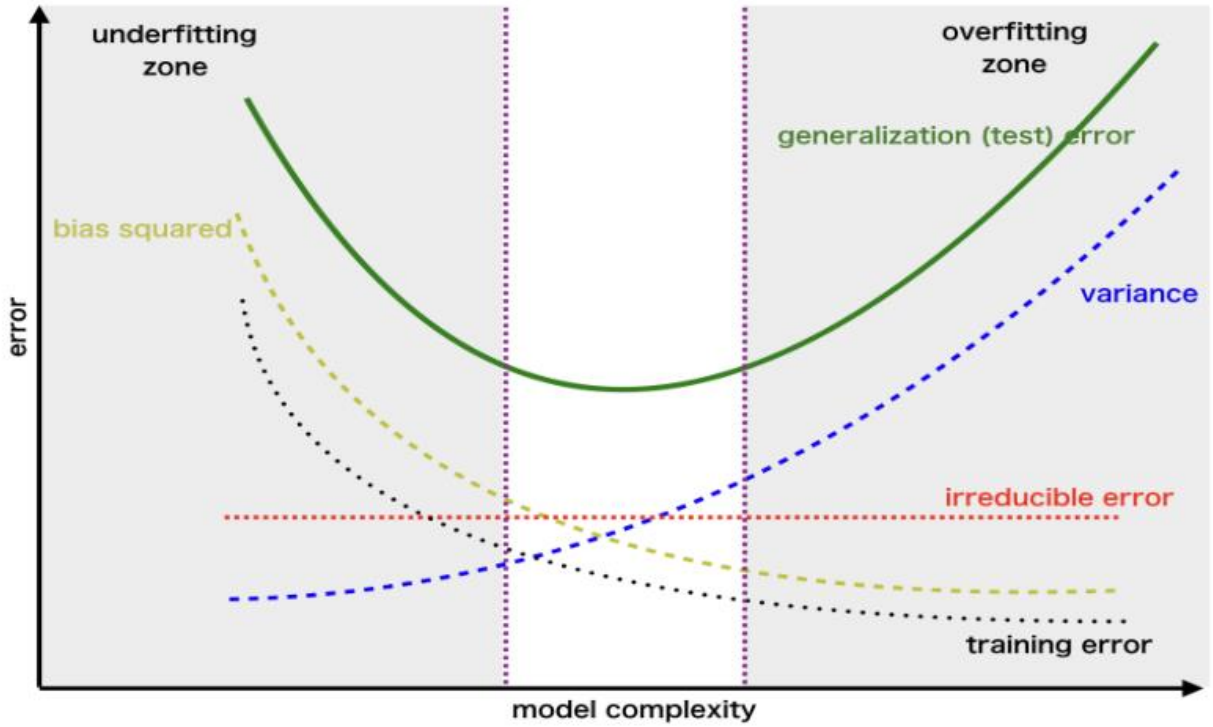
ب: در معادله نرمال تاثیری نمیگذارد و میتوان در حالت عادی از الگوریتم استفاده کرد ولی در استفاده از گرادیان نزولی سرعت بسیار کم میشود و در این حالت میتوانیم از feature scalling استفاده کنیم تا تعداد تکرار ها را کمتر کنیم.

سوال 5)

در موارد با واریانس بالا الگوریتم به نوعی درگیر بیش برازش میشود که حاصل تمرکز آن بر روی داده های نویز است ولی با اضافه کردن داده های بیشتر نیز تعداد داده های دارای نویز بالا میرود و الگوریتم قادر بیش برازش نمیباشد بنابر این تمرکز اصلی را بر روی داده های دیگر میگذارد و به نوعی از بیش برازش خارج میشود پس اضافه کردن به داده های آموزش برای واریانس بالا مناسب است.

در موارد با بایاس بالا اضافه کردن به داده های آموزش معمولا یا تغییری در بایاس ایجاد نمیکند یا باعث افزایش بایاس میشود.

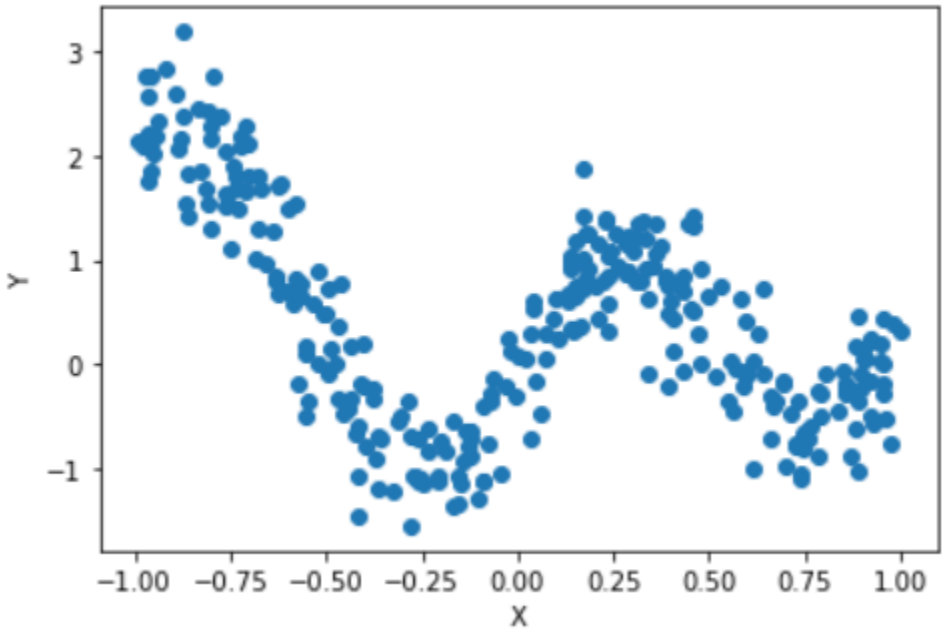
سوال 6)


underfitting
zone
overfitting
zone
generalization (test) error
bias squared
variance
irreducible error
training error
error
model complexity

بخش دو

سوال 1)

الف:

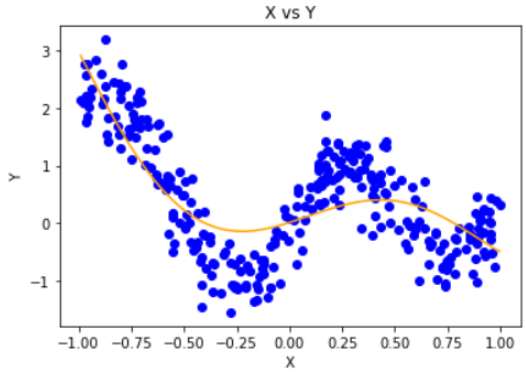


ب:

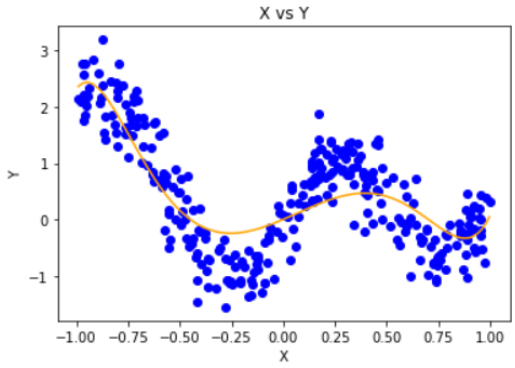
شافل کردن دیتا زمانی انجام میشود دیتاست شما بر مبنای یک ستون خاص مرتب شده است و یا زمانی که دارای high bias شده ایم از شافل کردن دیتا استفاده میکنیم تا بتوان یک توزیع مناسب برای داده ها ایجاد کنیم تا مجموعه دیتا آموزش دارای انواع دیتا باشد.

ج:

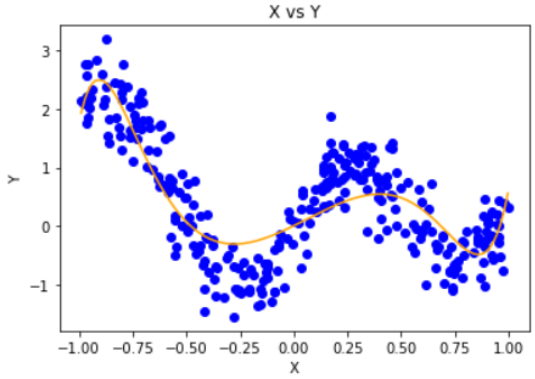
بررسی داده ها با 5 هزار بار تکرار و learning rate: 0.01



درجه 5

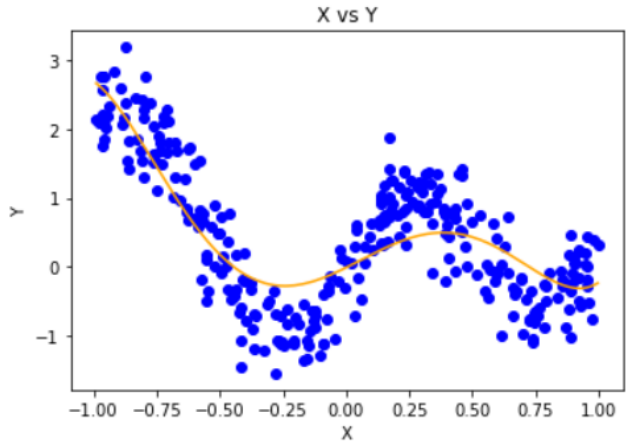


درجه 8

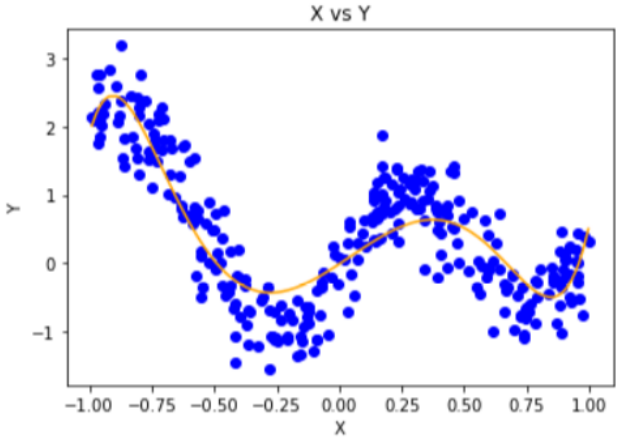


درجه 10

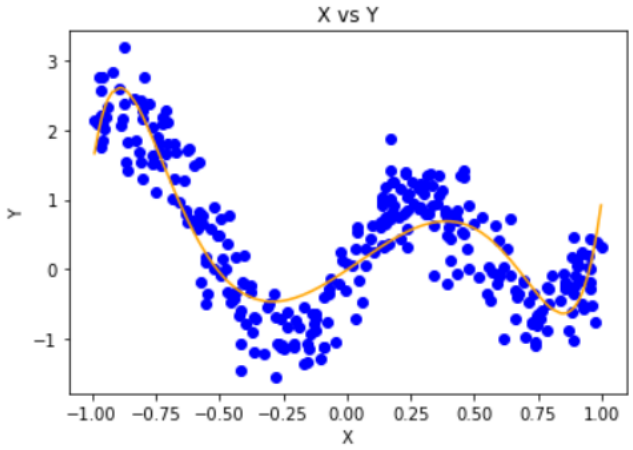
در نمودار های زیر الگوریتم بالا با تعدا تکرار 10000 بار انجام شده است



درجه 5



درجه 8



درجه 10

در درجه 10 با تعداد تکرار 10000 بار بیش برازش اتفاق افتاده است.

در این الگوریتم معادله اضلی که با آن کار میشود برابر زیر است:

Y= ax^2+ bx+c

که در آن تتا برابر a,b,c میباشد. هدف این است که ضرایب چند جمله ایی پیدا شود. در ابتدا همه ضرایب به صورت رندوم اختصاص داده میشود و بنابر منحنی رسم شده نسبت آن را با x و y میسنجدو

حال با گرادینت دیسنت سعی میکیم تا لاس فانکشن را به کمترین میزان ممکن برسانیم. در این حالت تتای بهتری بدست می آید و آن ها را در معادله میگذاریم. همین روند تا زمانی که منحنی به داده ها نزدیک شود ادامه میدهد.

سوال 2)

الف:

برای پر کردن داده های گمشده در یک فایل یک راه حل مشخص وجود ندارد و بنا بر مسئله و نوع داده، راهکار های حل این مشکل میتواند متفاوت باشد. راحت ترین راه برای مقابله با داده های گمشده حذف ردیف هایی میباشد که داده ها در آن قرار دارند. در این راه حل در صورتی که ردیف های دارای داده گمشده دارای اطلاعات حساسی باشند، این اطلاعات از بین میروند که داده آموزش را با مشکل مواجه میکند.

راه حل دیگر برای مقابله با داده های گمشده استفاده از دستور زیر میباشد. روش کار این دستور به این نحو است که از داده های بالا و پایین داده گمشده میانگین گرفته و آن را جایگزین داده گمشده میکند. این راه حل، مناسب دیتاست هایی با موارد گمشده زیاد نیست و همین طور اگر داده گمشده از ضریب همبستگی بالایی برخوردار باشد ممکن است باعث شود برازش به صورت اشتباه انجام شود.

Dataset. Interpolate()

یک روش دیگر پر کردن داده های گمشده با مقدار صفر است که برای این روش از دستور زیر استفاده میشود.

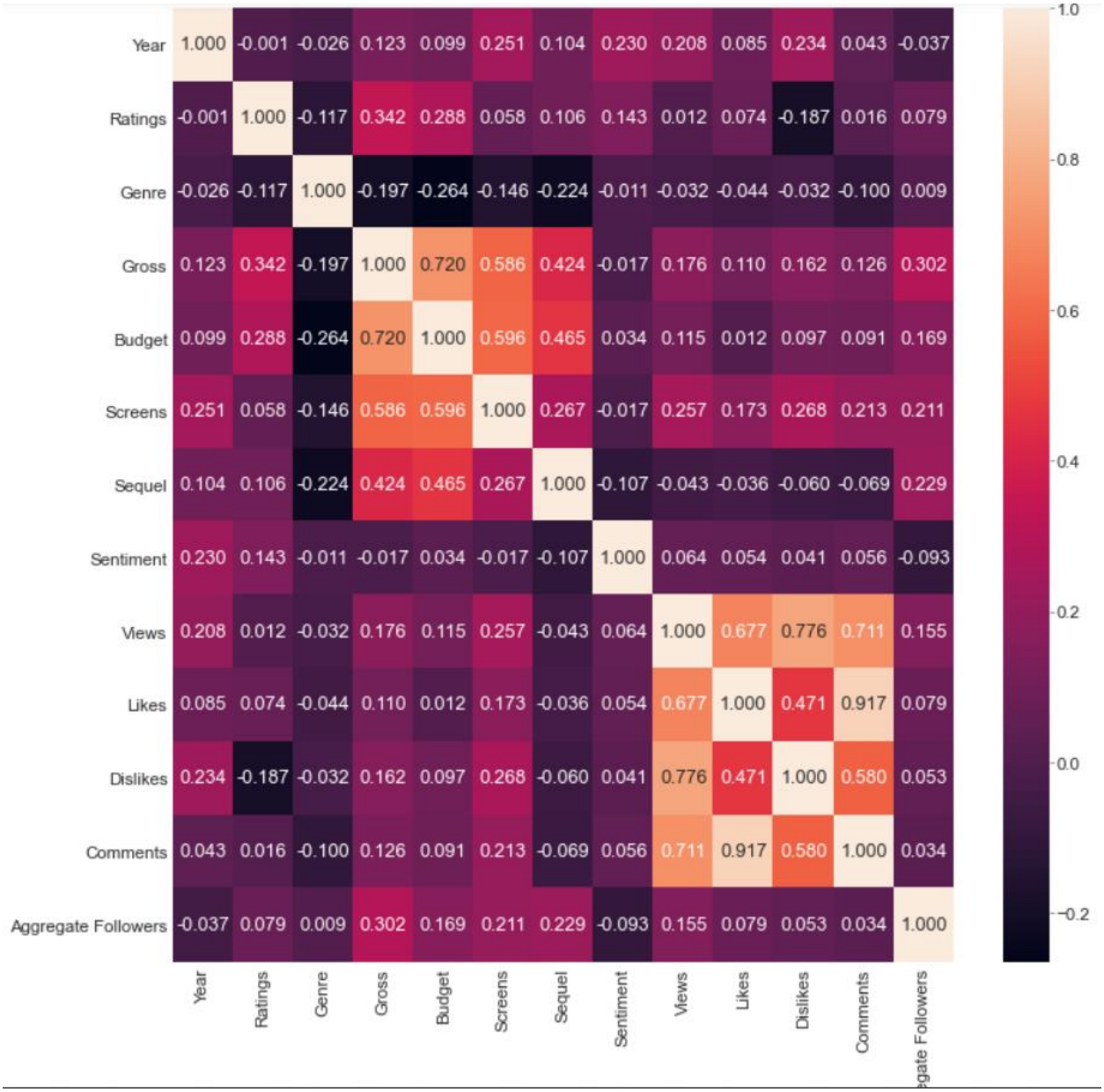
```
Dataset.fillna(0)
```

روشی دیگر پر کردن داده های گمشده با مقدار قبلی یا بعدی است که این روش نزی مانند روش های دیگر دارای نکات منفی از جمله اموزش اشتباه الگوریتم است.

یک روش دیگر برای مقابله با داده های گمشده استفاده از کتابخانه sklearn میباشد. در این کتابخانه دستوری به شکل زیر وجود دارد. با استقاده از دستور گفته شده میتوان مقادیر گمشده را با مقادیر نظیر، میانگین ، بیشترین، کمترین ، پر تکرار ترین یا عدد رندوم در آن ستون جایگزین کرد.

SimpleImputer

ب:



ج:

با توجه به نمودار و اینکه کلاس مورد هدف ما rating میباشد میتوان ان دسته از ویژگی هایی که دارای ضریب همبستگی بسیار نزدیک به عدد صفر دارند را حذف کرد که آن بسته به دیدگاه نسبت سوال دارد ولی با توجه به این که ستون های year, views, comments بین 0.02 و 0.02- است میتوان آن ها حذف کرد.

به این دلیل این ستون ها کمترین تاثیر را نسبت به rating دارند پس میشود از آن ها صرفه نظر کرد.

سوال 3)

الف: عملکرد آن به جز در مواردی که متغیر دسته‌ای مقادیر خیلی زیادی بگیرد، بسیار خوب است (معمولا از این روش برای متغیرهایی که بیش‌تر از ۱۵ مقدار متفاوت بگیرند، مناسب نیست. در برخی از مواردی که تعداد متغیرها کمتر است نیز امکان دارد گزینه مناسبی نباشد). کدبندی One-Hot ستون‌های دودویی (binary) جدیدی می‌سازد که هر یک مربوط به یکی از مقادیری هستند که متغیر به خود می‌گیرد. برای درک بهتر موضوع، مثالی در ادامه ارائه شده است.

فرض می‌شود که یک متغیر با عنوان Color در مجموعه داده وجود دارد که مقادیر آن Red ، Yellowو Greenهستند. اکنون، کلیه مقادیر این متغیر به سه ستون جدا با عنوان‌های Yellow ، Redو Green تبدیل می‌شوند. فرض کنید پنج «نمونه» در مجموعه داده وجود دارد که مقدار متغیر Color برای آن‌ها به ترتیب برابر با Green ، Yellow، Red، Redو Yellow است. اکنون، برای نمونه اول که رنگ آن قرمز است در ستون Redعدد ۱ و در ستون‌های Yellow و Green عدد صفر وارد می‌شوند. برای نمونه دوم نیز که رنگ آن قرمز است به همین صورت عمل می‌شود. برای نمونه سوم که رنگ آن زرد است در ستون Red و Green مقدار صفر و در ستون Yellow مقدار ۱ وارد می‌شود. برای متغیر چهارم که رنگ آن Green است در خانه‌های Red و Yellowمقدار صفر و در خانه Green مقدار ۱ وارد می‌شود. برای سایر نمونه‌ها نیز به همین صورت عمل می‌شود.

Interger encoding :

جایی که هر برچسب منحصر به فرد به یک عدد صحیح نگاشت می شود.

One Hot Encoding : جایی که هر برچسب به یک بردار باینری نگاشت می شود. جاسازی آموخته شده: جایی که یک نمایش توزیع شده از دسته ها آموخته میشود