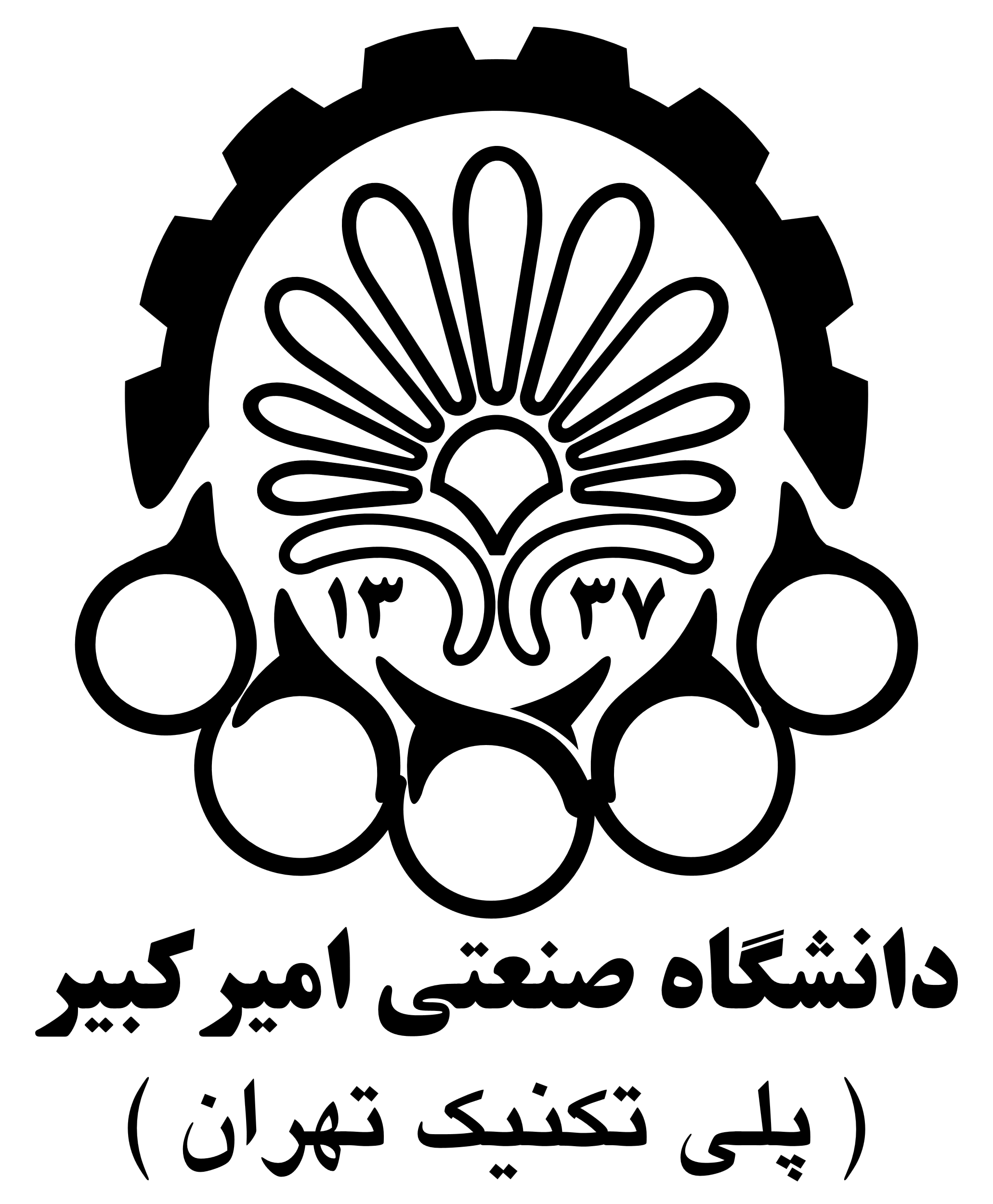
****

**دانشکده مهندسی کامپیوتر**

**تمر ین اول درس یادگیری ماشین**

**دکتر ناظرفرد**

**غلامرضا دار 400131018**

**پاییز ۱۴۰۰**

فهرست مطالب

[**بخش اول: پرسشهای تشریحی 3**](#_Toc88184521)

[سوال 1) 3](#_Toc88184522)

[سوال 2) 3](#_Toc88184523)

[سوال 3) 3](#_Toc88184524)

[سوال 4) 5](#_Toc88184525)

[سوال 5) 6](#_Toc88184526)

[سوال 6) 7](#_Toc88184527)

[**بخش دوم: پیاده سازی 8**](#_Toc88184528)

[سوال 1) 8](#_Toc88184529)

[سوال 2) 8](#_Toc88184530)

[سوال امتیازی) 9](#_Toc88184531)

# بخش اول: پرسشهای تشریحی

## سوال 1) ✅

**درستی یا نادرستی عبارات زیر را با ذکر دلیل مشخص کنید.**

**الف) در یک الگوریتم یادگیری، کاهش خطای آموزش، منجر به کاهش خطای آزمون میشود.**

نادرست. زیرا ممکن است مدل ما داده ها را حفظ کرده باشد و اصطلاحا Overfit شده باشد. همچین مدلی قابلیت Generalization خوبی ندارد و با دیدن داده جدید (Test data) نمی‎تواند حدس درستی بزند.

**ب)با کاهش تعداد داده‎های آموزش، الگوریتم بیشتر مستعد بیش برازش میشود.**

درست. هرچه تعداد داده ها کمتر باشد احتمال اینکه نتوانیم پترن های کلی داده ها را ببینیم بیشتر می‎شود این نویز ها با بقیه داده ها میانگین گرفته می‎شوند و تاثیر کمتری بر تصمیم نهایی می‎گذارند.

**ج) افزایش پیچیدگی مدل در رگرسیون همیشه سبب کاهش خطای آموزش و افزایش خطای آزمون میشود.**

نادرست. افزایش پیچیدگی همواره باعث کاهش خطای آموزش می‎شود اما همیشه باعث افزایش خطای آزمون نمی‎شود. اتفاقا برای مدلی که پیچیدگی بسیار کمی دارد، افزایش پیچیدگی می‎تواند خطای آزمون را به طرز قابل توجهی کاهش می‎دهد.

**د) از میان معیارهای , و معیار در برابر داده‎های پرت و نویزی عملکرد بهتری دارد.**

نادرست. معیار MSE به دلیل اینکه خطاها را به توان 2 می‎رساند عملکرد مناسب تری در مقابل نویز دارد. MAE با مقادیر خطای کم و زیاد به طور یکسان برخورد میکند اما MSE خطاهای زیاد را بیشتر Punish می‎کند و در نتیجه در حضور نویز و داده پرتِ زیاد عملکرد بهتری دارد.

## سوال 2) ✅

**فرض کنید که از رگرسیون چندجملهای استفاده میکنید و با رسم منحنی‎ها متوجه میشوید که اختلاف زیادی بین خطای آموزشی و خطای اعتبارسنجی وجود دارد. توضیح دهید که چه اتفاقی رخ داده است و سه راه‎حل برای حل این مشکل ارائه دهید.**

بیش‎برازش یا Overfitting رخ داده و مدل، داده‎ی آموزش را به نوعی حفظ کرده است.

* راه حل اول کم کردن پیچیدگی مدل است. در مورد رگرسیون چندجمله ای می‎توان درجه چندجمله ای را کم کرد.
* راه حل دوم افزودن دیتاست تا پترن هایی که در سمپل داده آموزش ما وجود نداشته را هم به مدل نشان دهیم. همچنین اگر داده های Train, Valid به طور همگن تقسیم نشده اند بهتر است کل داده ها را شافل کنیم و بعد به دو دسته Train, Valid تقسیم کنیم.
* راه حل سوم استفاده از Regularization است که در این صورت مقادیر مربوط به وزن ها کوچکتر میشوند که به Generalization مدل کمک میکند.

## سوال 3)

**مجموعه داده آموزشی شامل داده به فرم در اختیار داریم که دارای بعد است. تابع هزینه به صورت زیر محاسبه میشود.**

**الف)نشان دهید که در رگرسیون خطی با تابع هزینه ، بهینه به این صورت است: ❌**

**ب) دو مشکلی که استفاده از این رابطه دارد را ذکر کنید و برای هر کدام راه حلی ارائه دهید. ❌**

اضافه کردن جمله منظم ساز باعث می‎شود مقدار زیاد وزن ها جریمه شود و در نهایت مقدار همه وزن ها عددی کوچک و متعادل باشد. با این کار هیچ ویژگی (Feature) خاصی بسیار پراهمیت تر از ویژگی‎های دیگر نخواهد بود. همچنین منظم سازی مدل را تشویق می‎کند ساختار ساده تری داشته باشد که این کار باعث می‎شود مدل از Overfitting دوری کند.

**ج) اگر یک جمله منظم ساز نُرم 2 به صورت به رابطه اضافه کنیم، فرم بسته بهینه را بدست آورید. ✅**

**د) توضیح دهید اضافه کردن جمله منظم‎ساز چه مزیت‎هایی نسبت به حالت عادی دارد. ✅**

اضافه کردن جمله منظم ساز باعث می‎شود مقدار زیاد وزن ها جریمه شود و در نهایت مقدار همه وزن ها عددی کوچک و متعادل باشد. با این کار هیچ ویژگی (Feature) خاصی بسیار پراهمیت تر از ویژگی‎های دیگر نخواهد بود. همچنین منظم سازی مدل را تشویق می‎کند ساختار ساده تری داشته باشد که این کار باعث می‎شود مدل از Overfitting دوری کند.

## سوال 4) ✅

**با در نظر گرفتن الگوریتم‎های زیر برای رگرسیون خطی، به سوالات پاسخ دهید.**

1. **معادله نرمال**
2. **روشهای مبتنی بر گرادیان نزولی شامل:**
   1. Batch GD**: گرادیان‎ها بر اساس کل داده آموزش محاسبه میشوند.**
   2. Stochastic GD**: گرادیان‎ها بر اساس یک نمونه‎گیری از داده محاسبه میشوند.**
   3. Mini-Batch GD**: گرادیان‎ها با استفاده از مجموعه‎های تصادفی کوچکی از نمونه‎ها به دست می‎آیند.**

**الف) اگر یک مجموعه داده آموزش با میلیونها ویژگی در اختیار داشته باشید، از کدام الگوریتم رگرسیون خطی می توان استفاده کرد؟ دلایل انتخاب خود را شرح دهید.**

انتخاب بهتر در این مورد الگوریتم های بر پایه **Gradient** **Descent** هستند. زیرا محاسبه ترم که در روش نرمال مورد نیاز است دارای **پیچیدگی بالاتری** از اجرای الگوریتمGradient Descent حتی به تعداد Iteration های متعدد است.

**ب) فرض کنید که ویژگی های موجود در مجموعه داده آموزش در مقیاس‎های بسیار متفاوتی هستند. این ویژگی روی هر کدام از الگوریتم های مطرح شده، چه تاثیری میگذارد و راه حل شما چیست؟**

این ویژگی در مورد الگوریتم Normal Equation **تاثیری ندارد**. زیرا اساس کار معادله نرمال ضرب و اینورس ماتریس هاست و مقیاس اعداد تاثیری در این محاسبات ندارد.

اما در مورد الگوریتم Gradient Descent اینکه مقادیر ویژگی‎ها در مقیاس های تقریبا یکسان (ترجیحا -1 تا +1 ) باشد **تاثیرگذار است** زیرا در بازه های بزرگ سرعت Descend بیشتر از بازه های کوچک است و این ناهماهنگی ممکن است دچار غیربهینه شدن الگوریتم شود. راه حلی که برای این مشکل وجود دارد **Feature Scaling** نام دارد. . پس از اعمال Feature Scaling تمام مقادیر ویژگی ها در بازه تقریبا یکسان و یکنواختی قرار میگیرند که این کار سرعت الگوریتم Gradient Descent را افزایش می‎دهد. یکی از معروف ترین راه‎ها برای Feature Scaling در ادامه ذکر شده :

*(که میانگین مقادیر ویژگی i ام است و انحراف از معیار مقادیر ویژگی i ام است.)*

## سوال 5) ✅

**فرض کنید یک مدل با واریانس بالا داریم و برای بهبود عملکرد مدل تصمیم گرفته‎ایم که تعداد داده‎های آموزشی را افزایش دهیم. به نظر شما چه نتیجه‎ای حاصل خواهد شد؟ سوال قبل را در مورد مدلی که بایاس بالایی دارد نیز پاسخ دهید.**

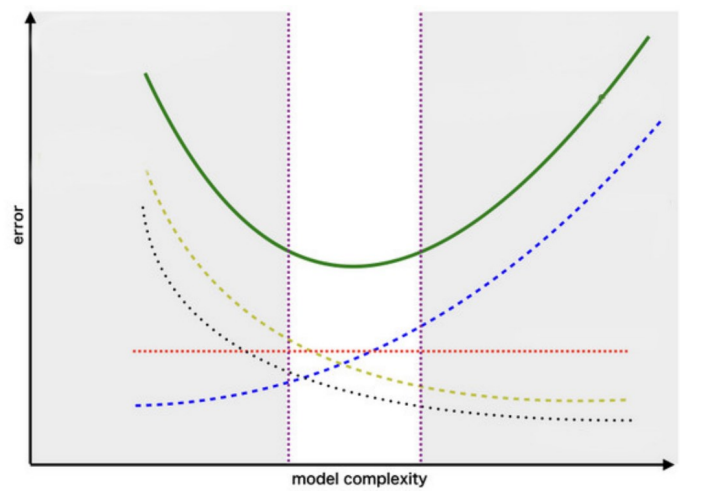
مدلی که واریانس بالایی دارد، Complexity نسبتا بالایی دارد. می‎دانیم مدل پیچیده نیاز به دیتای آموزش زیادی دارد و از این بابت **با افزایش داده ها کارایی مدل بهبود می‎یابد**. اما باید توجه کنیم که اگر واریانس مدل بسیار زیاد باشد و دیتای زیادی هم به مدل دهیم احتمالا مدل، دیتا را حفظ میکند و دچار Overfitting می‎شود. برای مقابله با بیش‎برازش و بهبود کارایی مدل میتوانیم از راه هایی مانند **Regularization** و **کاهش پیچیدگی مدل** استفاده کنیم.

مدلی که بایاس بالایی دارد،Complexity نسبتا پایینی دارد و فرض های ساده ای نسبت به توزیع داده ها میکند. افزایش تعداد داده ها در این حالت **احتمالا نمی‎تواند به مدل کمک زیادی بکند** چون اندازه ذخیره سازی اطلاعاتِ مدل پایین است و مدل پس از مدتی اشباع می‎شود. در این مورد احتمالا مدل دچار Underfitting شده و برای بهبود این شرایط راه‎ هایی مثل **اضافه کردن فیچر های بیشتر** و **پیچیده تر** و همچین **اضافه کردن تعداد پارامتر های بیشتر برای مدل** یا در کل **پیچیده کردن مدل** می‎توانند موثر باشند.

## سوال 6) ✅

**در شکل داده شده موارد زیر را مشخص کنید.**

|  |  |
| --- | --- |
| خطای آموزش ● | خطای ناشی از نویز ● |
| واریانس ● | **محدوده**Overfitting |
| مربع بایاس ● | **محدوده** Underfitting |
| خطای تست ● |  |

****

**Underfitting**

**Overfitting**

**خطای تست**

**واریانس**

**خطای آموزش**

**بایاس2**

**خطای ناشی از نویز**

# بخش دوم: پیاده سازی

## سوال 1) گزارش

**مجموعه داده dataset1.csv را که در فولدر dataset1 قرار دارد درنظر بگیرید**

**الف( داده ها را رسم کنید ✅**

**ب( شافل کردن داده ها به چه منظور انجام می شود؟ آیا مجموعه داده این سوال نیازی به این اقدام دارد؟ در صورت لزوم این مورد را بر روی مجموعه داده اعمال کنید. ✅**

مدلی که بایاس بالایی دارد،Complexity نسبتا پایینی دارد و فرض های ساده ای نسبت به توزیع داده ها میکند. افزایش تعداد داده ها در این حالت **احتمالا نمی‎تواند به مدل کمک زیادی بکند** چون اندازه ذخیره سازی اطلاعاتِ مدل پایین است و مدل پس از مدتی اشباع می‎شود. در این مورد احتمالا مدل دچار Underfitting شده و برای بهبود این شرایط راه‎ هایی مثل **اضافه کردن فیچر های بیشتر** و **پیچیده تر** و همچین **اضافه کردن تعداد پارامتر های بیشتر برای مدل** یا در کل **پیچیده کردن مدل** می‎توانند موثر باشند.

**ج( با استفاده از روش گرادیان نزولی نمودارهایی با درجه های 5 و 8 و 10 و با تعداد تکرارهای 5000 و 10000 بر روی داده ها برازش دهید و نمودارهای حاصل را رسم کنید. این عمل را برای سه معیار خطای MSE و RMSE و MAE تکرار کنید و نتایج را مقایسه کنید. همچنین در صورت مشاهده‎ی بیش‎برازش آن را گزارش کنید. ✅**

**د( برای هر یک از موارد قسمت قبل نمودار خطای آموزش و آزمون و نمودار اندازه قدم را رسم کنید. محور افقی نشان‎دهنده تکرارها و محور عمودی مقدار خطا و اندازه قدم را نشان دهد. ✅**

**هـ( قسمت قبل را با معادله‎ی نرمال و بدون در نظر گرفتن ضریب λ تکرار کنید و نتایج را مقایسه کنید. ✅**

**و( به ازای درجه 8 ،با روش معادله‎ی نرمال و با در نظر گرفتن سه مقدار مختلف برای ضریب λ نموداری بر روی داده‎ها برازش دهید و مقدار خطای RMSE را برای داده‎های آموزش و آزمون رسم کنید. تاثیر ضریب λ را بررسی نمایید. ✅**

## سوال 2)

مجموعه **داده “dataset\_CSM” که در فولدر “dataset2” قرار دارد را در نظر بگیرید. این مجموعه داده مربوط به فیلمها و امتیاز IMDB آنها میباشد که از سه منبع Twitter ،YouTube IMDB جمع‎آوری شده‎است. در این سوال میخواهیم مقدار متغیر Ratings را پیشبینی کنیم.**

**الف) در برخی از ستونهای مجموعه داده مقادیر گم‎شده وجود دارد. با استفاده از روشهای مناسب، مقادیر گمشده را پر کنید و به طور مختصر توضیح دهید که چگونه این کار را انجام دادید. همچنین اگر داده‎ها نیاز به پیش‎پردازش دیگری دارند، در این قسمت انجام دهید و در گزارش خود ذکر کنید. 🚕 اضافه کردن نرمالیزیشن**

راه های متعددی برای مقابله با مقادیر گم‎شده وجود دارد.

یکی از راه های رایج، **حذف کردن یا نادیده گرفتن سطر یا ستونی است که شامل مقدار گم‎شده است**. به این صورت که اگر ویژگی Likes دارای تعدادی مقدار گم‎شده باشد، کل ویژگی Likes را از دیتاست حذف میکنیم. همینطور که مشخص است برای این مثال خاص این کار عاقلانه ای نیست چون با حذف کردن ستون Likes با وجود اینکه مقادیر گم‎شده دیتاست را کم میکنیم اما مقدار زیادی اطلاعات مفید مخصوصا برای پیش بینی Rating که هدف این مسئله است را از دست می‎دهیم. حالت دیگر زمانی است که یک سطر دارای تعدادی مقدار گم شده باشد مثلا فیلم Ant-Man ویژگی Aggregate followers را ندارد. میتوانیم کل سطر مربوط به فیلم Ant-Man را حذف کنیم و با این کار هم تعداد مقادیر گم‎شده را کاهش داده ایم. اما این روش هم مناسب به نظر نمی‎رسد زیرا ممکن بود ما بدون نیاز به داشتن مقدار Aggregate followers برای فیلم Ant-Man و تنها با توجه به سایر ویژگی های دیتاست، پیش بینی مناسبی از Rating این فیلم داشته باشیم.

راه دیگر اصطلاحا **Imputation** نام دارد و به این معنی است که مقادیر گمشده دیتاست را به نحوی به طور مصنوعی پر بکنیم. استراتژی های بسیار زیادی برای پر کردن این مقادیر گم‎شده وجود دارد مثل جایگزینی با یک عنصر خنثی(0,1,””) یا جایگزینی با یک ویژگی آماریِ یک ویژگی یا کل دیتاست(Mean, Median, Min, Max, …) و استراتژی های دیگر. استفاده از این روش برای این دیتاست به نظر روش **معقول تری** می‎آید.

* مقادیر گم‎شده ویژگی Budget را میتوان با میانگینی از بودجه سایر فیلم ها جایگزین کرد. از آنجایی که ژانر های مختلف از نظر بودجه تفاوت زیادی با هم دارند (بودجه یک فیلم علمی تخیلی بالاتر از فیلم درام است) می‎توانیم بجای استفاده از میانگین همه فیلم ها، میانگین فیلم های هم ژانر را جایگزین کنیم.
* مقادیر گم‎شده ویژگی Screens را با میانگین سایر سطر ها جایگزین میکنیم ( شاید انتخاب بهتری هم می‎توانستیم بکنیم)
* ستون Aggregate followers برای تعداد زیادی از فیلم های به خصوص **جدید** خالی است. این اتفاق ممکن است در پیش بینی امتیاز فیلم های جدید مدل را به اشتباه بی‎اندازد به عنوان مثال جایگزین کردن میانگین این ویژگی برای اکثر فیلم های 2015 ممکن است به مدل این ایده را بدهد که فیلم های 2015 در سطح متوسط هستند و به اندازه فیلم های 2014 طرفدار ندارند. پس این ویژگی را **حذف** میکنیم.

**ب) نمودار همبستگی بین ویژگی‎ها را رسم کنید و در گزارش بیاورید. ✅**

**ج) با توجه به نمودار رسم شده آیا می‎توان یک یا چند ویژگی را حذف کرد؟ دلیل خود را ذکر کنید. ✅**

**د) با استفاده از روش گرادیان نزولی یک نمودار بر روی داده‎ها برازش دهید. (پارامترها را به گونه‎ای انتخاب کنید که به بهترین خروجی دست یابید) این عمل را یک بار با استفاده از کل ویژگی‎ها و یک بار با استفاده از ویژگی‎های منتخب تکرار کنید. نمودار خطای آموزش و آزمون و نمودار طول گام را رسم کنید. ❌❌**

## سوال امتیازی)

**دادههای مربوط به ویژگیهای شخصی و هزینهی پزشکی افراد در فولدر** dataset3 **قرار دارد. هر داده دارای 6 ویژگی ورودی و یک خروجی میباشد که هزینهی پزشکی فرد را نشان میدهد. همانطور که در داده‎های سوال دیده میشود، ویژگیهای جنیست، منطقه و سیگاری بودن از نوع** categorical **هستند. برای تبدیل این ویژگیها به ویژگیهای عددی در این سوال میخواهیم از دو روش Integer Encoding و One Hot Encoding استفاده کنیم.**

**الف) توضیح دهید که Integer Encoding و One Hot Encoding چگونه انجام می‎شوند. ✅**

هر موقع نیاز است با داده های Categorical که به صورت عددی نیستند کار کنیم، باید از یک روشEncoding استفاده کنیم و به کمک آن داده های از نوع رشته را به عدد تبدیل کنیم. دو روش معروف Integer Encoding و OHE هستند. Integer Encoding به این صورت است که مانند یک دیکشنری به هر دیتای غیر عددی یک عدد صحیح نسبت میدهیم.

مثلا Bad=0, Average=1, Good=2.

روش دیگر روشOne Hot Encoding است که بردارهایی با طول مشخص از اعداد صفر و یک را به جای هر داده‎ی غیر عددی درنظر میگیرد به طوری که در هر بردار فقط یک مولفه 1 وجود داشته باشد و سایر مولفه ها 0 هستند.

مثلا Bad=[0,0,1], Average=[0,1,0], Good=[1,0,0]

**ب)ویژگیهای جنسیت و سیگاری بودن را با استفاده از Integer Encoding و ویژگی منطقه را با استفاده از OHE به مقدار عددی تبدیل کنید (برای جنسیت :** male**=**1, female=0 **و برای سیگاری بودن :** yes=1, no=0**) ✅**

**ج) به نظر شما چرا برای ویژگی منطقه از OHE استفاده میکنیم و از Integer Encoding استفاده نمیکنیم؟ ✅**

اگر از Integer Encoding استفاده کنیم به طور ناخودآگاه رابطه کوچکتر بزرگتری بین مقادیر مختلف یک ویژگی ایجاد میکنیم. مثلا در مثال Bad=0, Average=1, Good=2 یک توالی بین مقادیر از بد تا خوب ایجاد کردیم که مدل ممکن است در پروسه یادگیری خود بخواهد از این ویژگی استفاده کند. در مثال ویژگی Region مقادیر مختلف ارتباطی به یکدیگر ندارند و هر کدام از داده ها میتواند هر عددی به خود بگیرد. با استفاده از OHE به نوعی تعدادی بردار عمود بر هم را به مقادیر ویژگی نسبت دادیم و استقلال بین مقادیر را به صورت ریاضی به مدل نشان داده ایم.

**د) رگرسیون خطی تعمیم یافته را بدون جمله‎ی منظم ساز پیاده سازی کرده، نتایج را روی داده های تست گزارش کنید. تابع basis برای ویژگی سن به صورت زیر باشد: ❌❌**

**و برای سایر ویژگی‎ها از تابع همانی استفاده کنید:**

**هـ) در این قسمت ابتدا تعداد داده‎های آموزش را 10 در نظر بگیرید و با گامهای 10 تایی، تا 1000 افزایش دهید. تغییرات خطای تست و آموزش را با افزایش داده آموزش بررسی کنید. برای مقایسه پذیر بودن خطا، از MSE برای تابع هزینه استفاده کنید. نمودار این تغییرات را در گزارش بیاورید. ❌**

**و) روش** StochasticGradientDescent **را بر روی این مجموعه داده پیاده سازی کنید و مقدار دقت و خطا را گزارش کنید. ❌❌❌**