بخش اول: پرسش‌های تشریحی

سوال ۱) 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.

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

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

د) از میان معیارهای MSE و RMSE و MAE معیار MAE در برابر داده‌های پرت و نویزی عملکرد بهتری دارد.

سوال ۲) فرض کنید که از رگرسیون چندجمله‌ای استفاده می‌کنید و با رسم منحنی‌ها متوجه می‌شوید که اختلاف زیادی بین خطای آموزشی و خطای اعتبارسنجی[1] وجود دارد. توضیح دهید که چه اتفاقی رخ داده است و سه راه‌حل برای حل این مشکل ارائه دهید.

سوال ۳) مجموعه داده آموزشی شامل n داده به فرم $(x_i . y_i)$ در اختیار داریم که x_i دارای d بعد است. تابع هزینه SSE به صورت زیر محاسبه می‌شود:

$$J(w) = \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

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

$$\hat{w} = (X^T X)^{-1} X^T y$$

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

ج) اگر یک جمله منظم‌ساز نُرم ۲ به صورت $||w||^2$ به رابطه loss اضافه کنیم فرم بسته w بهینه را به دست آورید.

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

سوال ۴) با در نظر گرفتن الگوریتم‌های زیر برای رگرسیون خطی، به سوالات پاسخ دهید:

۱- معادله‌ی نرمال (Normal Equation)

۲- روش‌های مبتنی بر گرادیان نزولی شامل:

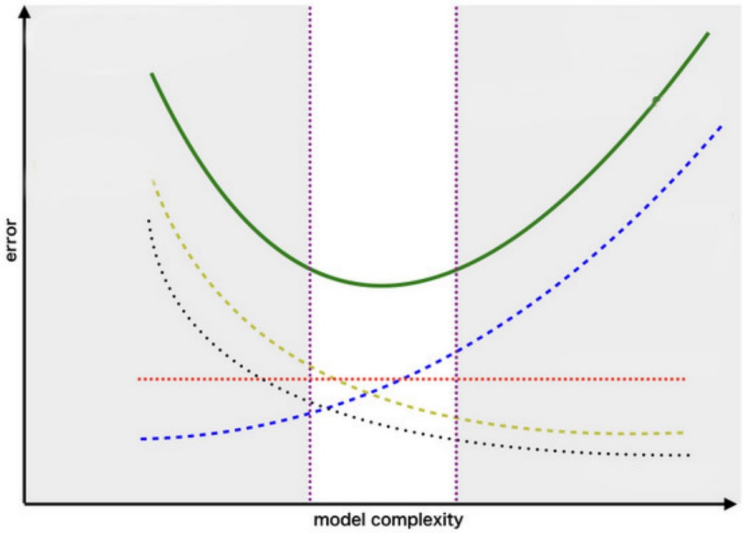
- Batch GD: گرادیان‌ها بر اساس کل داده آموزش محاسبه می‌شوند.
- Stochastic GD: گرادیان‌ها بر اساس یک نمونه‌گیری از داده محاسبه می‌شوند.
- Mini-Batch GD: گرادیان‌ها با استفاده از مجموعه‌های تصادفی کوچکی از نمونه‌ها به دست می‌آیند.

[1] validation error

الف) اگر یک مجموعه داده آموزش با میلیون‌ها ویژگی[2]در اختیار داشته باشید، از کدام الگوریتم رگرسیون خطی می توان استفاده کرد؟ دلایل انتخاب خود را شرح دهید.
ب) فرض کنید که ویژگی های موجود در مجموعه داده آموزش در مقیاس‌های[3]بسیار متفاوتی هستند .این ویژگی روی هر کدام از الگوریتم های مطرح شده، چه تاثیری می‌گذارد و راه حل شما چیست؟

سوال ۵) 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.

سوال ۶) در شکل داده شده موارد زیر را مشخص کنید.
خطای آموزش، واریانس، مربع بایاس[4]، خطای ناشی از نویز[5]، خطای تست، محدوده overfitting ، محدوده underfitting



بخش دوم: پیاده‌سازی

سوال اول:

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.
ج) با استفاده از روش گرادیان نزولی نمودارهایی با درجه های ۵ و ۸ و ۱۰ و با تعداد تکرارهای ۵۰۰۰ و ۱۰۰۰۰ 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.

---
[2] feature
[3] scale
[4] Bias squared
[5] irreducible error

د) برای هر یک از موارد قسمت قبل نمودار خطای آموزش و آزمون و نمودار اندازه قدم را رسم کنید. محور افقی نشان‌دهنده تکرارها[6] و محور عمودی مقدار خطا و اندازه قدم را نشان دهد.

ه) قسمت قبل را با معادله‌ی نرمال و بدون در نظر گرفتن ضریب λ تکرار کنید و نتایج را مقایسه کنید.

و) به ازای درجه ۸، 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 دوم:

مجموعه داده CSM_dataset 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 TwitterوYouTube وIMDB جمع‌آوری شده‌است. در این سوال می‌خواهیم مقدار متغیر Ratings را پیش‌بینی کنیم.

الف) در برخی از ستون‌های مجموعه داده مقادیر گمشده[7]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.

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 سوم (امتیازی)**

داده‌های مربوط به ویژگی‌های شخصی و هزینه‌ی پزشکی افراد در فولدر dataset3 قرار دارد. هر داده دارای ۶ 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 (OHE) استفاده کنیم.

الف) توضیح دهید که integer encoding و one hot encoding (OHE) چگونه انجام می‌گیرد.

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 : female=0, male=1 - برای سیگاری بودن: no = 0 , yes=1)

ج) به نظر شما چرا برای ویژگی منطقه از OHE استفاده می‌کنیم و از integer encoding استفاده نمی‌کنیم؟

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:

---

[6] iteraions

[7] Missing values

$$\phi_{age}(x_{age}) = x_{age}^{۲}$$

و برای سایر ویژگی‌ها از تابع همانی استفاده کنید:

$$\phi_(x) = x$$

ه) در این قسمت ابتدا تعداد داده‌های آموزش را ۱۰ در نظر بگیرید و با گام‌های ۱۰ تایی، تا ۱۰۰۰ 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 stochastic gradient descent را بر روی این مجموعه داده پیاده سازی کنید و مقدار دقت و خطا را گزارش کنید.

پایدار باشید

نکات مربوط به تحویل تمرین

- کدهای خود را ترجیحا به زبان پایتون و در محیط jupyter پیاده‌سازی کنید.
- نظم و خوانایی در نوشتن گزارش و کدها از اهمیت بالایی برخوردار است. کدهای خود را تا حد امکان کامنت‌گذاری کنید.
- در پیاده‌سازی بخش‌های مختلف امکان استفاده از کتابخانه‌های آماده مربوط به الگوریتم‌های یادگیری ماشین رابه طور کلی ندارید. مگر در مواردی که در صورت سوال ذکر شده باشد.
- برای خواندن داده‌ها می‌توانید از کتابخانه pandas و برای نمایش نمودارها و عملیات ماتریسی می‌توانید از کتابخانه‌های numpy و matplotlib استفاده کنید. برای محاسبه معیارهای ارزیابی مانند دقت، ماتریس درهم‌ریختگی و تقسیم داده‌ها به مجموعه‌های آموزش و آزمون نیز استفاده از کتابخانه آماده مجاز است.
- در صورتی که داده‌ها را به دسته‌های آموزشی، تست (ویا validation) تقسیم می‌کنید، درصد هر کدام را در گزارش ذکر کنید.
- فایل‌های کد و گزارش خود را در قالب یک فایل فشرده با فرمت HW01_StdNumber.zip که StdNumber شماره دانشجویی شماست، در سامانه بارگذاری کنید.
- سوالات ستاره دار(*) دارای نمره اضافی بر تمرین است.
- مطابق قوانین دانشگاه هر گونه کپی‌برداری ممنوع است و در صورت مشاهده، نمره **هر دو طرف صفر** داده می‌شود.
- در صورت وجود هر گونه سوال یا ابهامی با ایمیل درس در تماس باشید:

ml.ce.aut@gmail.com