به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

یادگیری ماشین

استاد درس: دکتر میرزایی

تمرین 1

اسفند ماه ۹۸

سوال اول:

هدف این سوال طراحی یک طبقه بند لاجستیک است .داده های این سوال مربوط به یک بیماری خونی است در این داده دو ویژگی خون(میزان گلوکز و اکسیژن در خون)موجود است. در ستون سوم ، لیبل مربوط به وجود بیماری در هر فرد آمده است که 1 به معنای وجود بیماری و 0 به معنای سالم بودن فرد مورد بررسی است .هدف این سوال طراحی طبقه بند برای پیش بینی وجود بیماری براساس دو ویژگی خون است.

1 .نموداری رسم کنید که محورهای آن دو ویژگی دیتاست باشد .بر روی این نمودار بیماربودن یا نبودن را با رنگ های مختلف مشخص کنید.

2 .میدانیم که رگرسیون لاجستیک مطابق فرمول زیر تعریف میشود:

$$h\theta = sigmoid\,(\theta^T x)$$

یکی از مهمترین قسمتهای هر طبقه بند Loss Function است .در این سوال Loss Function را به صورت زیر تعریف میکنیم:

$$J(\theta) = \frac{1}{m}\sum_{k=0}^{n}\{-y^{(i)}\log[\mathrm{h}_\theta(x^{(i)})] - (1-y^{(i)})\log[1-\mathrm{h}_\theta(x^{(i)})]\}$$

با استفاده از گرادیان نزولی بردار θ را به گونه ای پیدا کنیدکه تابع هزینه کمینه شود .همچنین فرمولهای محاسبه شده براساس گرادیان نزولی را در گزارش خود ذکر کنید.

3 .با استفاده از L2 Norm تابع هزینه را تغییر دهید و regularization را نیز به طبقه بند اضافه کنید و مجددا θ بهینه را بدست آورید.

4 .بهترین دقت حاصله را گزارش کنید.

سوال دوم:

در سوال قصد داریم با رگرسیون خطی ساده، چند متغیره و چند جمله ای آشنا شویم. هدف از رگرسیون یافتن ضرایبی برای تابع پیش بینی کننده است به نحوی که بیشترین مطابقت با دادگان هدف را داشته باشند. به این منظور:

1 .در ابتدا برای دادگان آموزش و هدف زیر یک مدل رگرسیون خطی آموزش دهید. تابع خطا و روش بهینه سازی ضرایب را با جستجو در منابع و مطالعه دقیق تر رگرسیون خطی، به اختیار خود انتخاب کنید.

```
X = [5, 15, 25, 35, 45, 55]
y = [5, 20, 14, 32, 22, 38]
```

2 .پس از یافتن ضرایب تابع، با استفاده از تابع خطایی که در قسمت قبل مشخص کردید، مقدار کلی خطا را برای پیشبینی که انجام داده اید محاسبه کنید.

3 .نقاط آموزش به همراه تابع رگرسیون ، که در قسمت 1 مدل کردید، را در یک `plot` بیاورید.

4 .قسمت 1تا3 را برای دادگان آموزش و هدف زیر تکرار کنید:

```
x = [[0, 1], [5, 1], [15, 2], [25, 5], [35, 11], [45, 15], [55,
34], [60, 35]]
y = [4, 5, 20, 14, 32, 22, 38, 43]
```

5 .قسمت 1تا3 را برای دادگان آموزش و هدف زیر تکرار کنید:

```
x = [5, 15, 25, 35, 45, 55]
y = [15, 11, 2, 8, 25, 32]
```

6 .سعی کنید برای دادگان قسمت 5 با اضاف کردن ترم x^2، به جای رگرسیون خطی، رگرسیون چند جمله ای انجام دهید و همه مراحل 1تا3 را مجددا تکرار کنید. نتایج را تحلیل نمایید.

سوال سوم-

در این سوال قصد داریم با رگرسیون چند متغیره آشنا شویم و به کمک آن کمیتی از یک دیتاست را پیش بینی کنیم.

برای این منظور از این لینک در ابتدا با جزئیات دادگان **Energy efficiency** و ویژگی های هر داده آشنا شوید. قصد ما این است که دو کمیت ستون 9 ام و 10 ام که Heating Load و Cooling Load هستند، پیش بینی شوند. شما باید با استفاده از ویژگی های 8 ستون اول بتوانید ستون 9 ام و 10 را پیش بینی کنید.

1- 8 ستون اول هر داده به عنوان داده آموزش و ستون 9 ام و 10ام به عنوان داده هدف، سعی کنید مدلی طراحی کنید که بتواند ستون 9 و 10 را پیش بینی کند. همچنین از داده 600 به بعد به عنوان دادگان تست برای ارزیابی دقت مدلتان استفاده کنید.

2- برای قسمت 1 با تحقیق کردن در منابع موجود، تابع خطا و روش بهینه سازی مورد نظرتان انتخاب کرده و به صورت کامل در گزارش کار شرح دهید.

3- در الگوریتم های یادگیری ماشین یک روش مرسوم برای ارزیابی مدل ساخته شده، تقسیم کردن کل دادگان موجود به 3 دسته می‌باشد.دو دسته آموزش و تست را که با آن آشنا هستیم، دسته سوم، دسته دادگان اعتبار سنجی هستند. در خصوص دادگان اعتبار سنجی تحقیق کنید و در گزارش کارتان شرح دهید.

4- در این قسمت، بخش اول همین سوال را تکرار کنید با این تفاوت که 10 درصد از کل دادگان آموزشتان را به عنوان دادگان اعتبار سنجی در نظر بگیرید. برای مثال به ازای آموزش از طریق 45 داده آموزش ، اعتبار مدلتان را بر روی 5 داده بعدی به عنوان دادگان اعتبار سنجی بسنجید و به همین روال تا پایان دادگان آموزش ادامه دهید.

5- مقدار تابع خطا برای دادگان آموزش، اعتبار سنجی و تست را در هر مرحله گزارش کنید.

6- در برخی مواقع برای انجام رگرسیون با دادگان بسیار بزرگی سرو کار داریم که انجام محاسبات بر کل ویژگی های دیتا ست هم از نظر زمانی هم هزینه ای برای ما نامناسب می‌باشد. در این مواقع سعی می‌کنیم که از ویژگی های "موثر" دادگان برای رگرسیون استفاده کنیم. با تحقیق در منابع روش هایی برای شناسایی ویژگی های موثر دایتاست ها پیشنهاد کنید.

7- با استفاده از روش پیشنهادی خود در قسمت قبل سعی کنید ویژگی های موثر دیتاست این سوال برای پیش بینی ستون 9 و 10 را بیابید و بار دیگر قسمت اول را تکرار کنید و نتایج را با وقتی که از همه ویژگی ها استفاده کرده بودید بصورت کامل مقایسه کنید.

سوال چهارم:

با استفاده از روش نیوتن و استفاده از ی backtracking line search مقدار کمینه تابع f(x) را بدست آورید .شرط توقف الگوریتم را بیان کرده و علت انتخاب آن را توضیح دهید.

مقدار تابع و طول گام بدست آمده در هر مرحله را رسم کرده و درمورد تغییرات طول گام توضیح دهید.

$$f(x) = -(10x^3 + 60x - 2x^6 - 3x^4 - 12x^2)$$

نکات:

- مهلت تحویل این تمرین 16 اسفند می‌باشد.
- گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است. لطفا تمامی نکات و فرض‌هایی که برای پیاده سازی‌ها و محاسبات خود در نظر می‌گیرید را در گزارش ذکر کنید.
- برای دادگان لاجستیک، با استفاده از توابع مناسب در کتابخانه های پایتون، فایل mat. را بخوانید و لود کنید.
- در صورت مشاهده‌ی تقلب نمرات تمامی افراد شرکت کننده در آن صفر لحاظ می‌شود.
- استفاده از کدهای آماده برای تمرینها مجاز نمی‌باشد. برای تمرین ها فقط برای قسمتهایی از کد برای پیاده‌سازی می توانید از کدهای آماده راهنمایی بگیرید. بنابراین، کپی کردن سختار ها و کدهای آماده و حل شده از اینترنت **تقلب** محسوب می‌شود.
- نحوه‌ی محاسبه‌ی تاخیر به این شکل است : مهلت بدون کسر نمره تا تاریخ 16 اسفند اعلام شده و تاخیر تا یک هفته بعد یعنی 21 اسفند با 30 درصد کسر نمره محاسبه خواهد شد.
- در صورت وجود هرگونه ابهام یا مشکل می‌توانید از طریق رایانامه‌ی زیر با دستیار آموزشی مربوطه در تماس باشید:

هاشم پور(hamidreza.hashemp@ut.ac.ir)

- گزارش کار های خود را به آیدی @cactus_995 در تلگرام ارسال کنید.