تمرین سری پنجم

درس: یادگیری ماشین- پاییز ۱۴۰۳

استاد درس: دکتر فاطمه زمانی

دستیاران آموزشی: حسین آقاگل زاده، ابوالفضل حسینی فر

دانشکده مهندسی برق و کامپیوتر دانشگاه صنعتی نوشیروانی بابل

- این تمرین دارای سه بخش **تئوری**، **پیاده سازی** و **مطالعاتی** است.
- پاسخنامه بخش تئوری می تواند به صورت دست نویس یا تایپی باشد. در هر صورت اما می بایست در قالب **یک فایل pdf** تحویل گردد.
- پاسخ بخش پیاده سازی می بایست حاوی **کد پیاده سازی** شده **به همراه گزارش آن در قالب فایل pdf** باشد. از آوردن کد در گزارش اجتناب کنید (مگر آوردن بخشی از آن از نظر شما ضرورت داشته باشد).
- پاسخ بخش مطالعاتی می بایست به صورت تایپ شده در قالب یک فایل pdf باشد.
- گزارش بخش پیاده سازی می بایست میزان تلاش شما رو با ٤ اولویت مهم **۱- شفافیت ۲- درستی ۳-زیبایی ٤- کوتاهی**، نشان بده
- بخش تئوری بدون استفاده از کد انجام بشه مگر دلیل موجهی وجود داشته باشه.
- کل پاسخ (کد و فایل های pdf) در **یک فایل zip** با فرمت زیر ارسال شود.

HW5_شماره دانشجویی.zip

تئوری

مسئله ۱

اگر داشته باشیم $p(x|c_i) = N(\mu_i, \sigma^2)$، برای مسئله طبقه بندی دو کلاسه با ورودی تک بعدی اثبات کنید که تابع خطا از رابطه زیر بدست می آید(احتمال prior دو کلاس را یکسان فرض کنید):

$$P_e = \frac{1}{\sqrt{۲\pi}} \int_{\frac{|\mu_۲ - \mu_۱|}{۲\sigma}}^{\infty} e^{-\frac{u^۲}{۲}} du$$

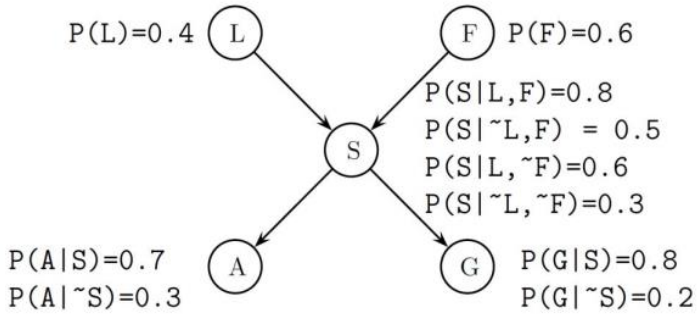
مسئله ۲

برای سناریو های استقلال شرطی ذکر شده موارد زیر را اثبات کنید.

- Head to Tail $P(X,Y,Z) = P(X)P(Y|X)P(Z|Y)$
- Tail to Tail $P(X,Y,Z) = P(X)P(Y|X)P(Z|X)$
- Head to Head $P(X,Y,Z) = P(X)P(Y)P(Z|X,Y)$

مسئله ۳

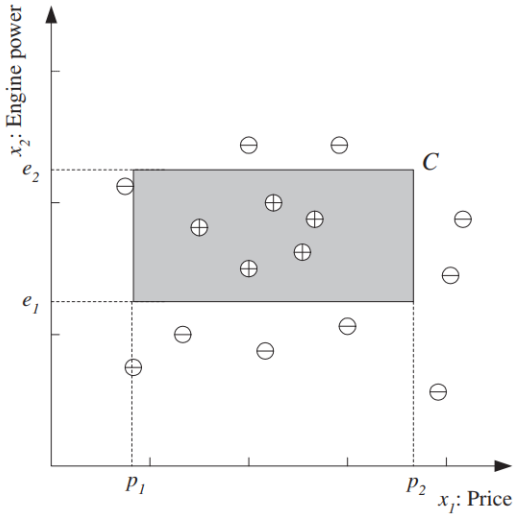
با توجه به مدل گرافی زیر، مطلوب است p(S) و p(S|A).



- تمام متغیر های تصادفی را باینری فرض کنید.
- برای سادگی S=True به صورت "S" و S=False به صورت"~S" نمایش داده شده است (و به همین صورت برای باقی متغیر ها).

- **مسئله ۴**

در مسئله یادگیری مطرح شده در کتاب، مدل یک مستطیل فرض شده است که شکل زیر نماینده این مسئله است.



الف) رابطه بین دو معیار false positive و false negative با مساحتی که این مستطیل در بر میگیرد چیست؟ (بیان کنید که با کم زیاد شدن مساحت مستطیل این دو معیار کم می شوند یا زیاد- مرکز مستطیل را ثابت در نظر بگیرید)

ب) اگر مدل فیت شده را یک دایره در نظر بگیریم، در فرآیند یادگیری چه پارامتری/پارامتر هایی باید پیدا شود؟

پ) اگر به جای دایره یک بیضی در نظر بگیریم چه پارامتر هایی باید یادگرفته شوند؟ چرا استفاده از بیضی به جای دایره منطقی تر است؟

ت) حال فرض کنید به جای یک مدلی که شامل یک مستطیل است مدلی شامل بیش از یک مستطیل داریم. این مدل چه مزیتی نسبت به حالت یک مستطیلی دارد؟

ث) در حالت ت فرض کنید تعداد مستطیل ها برابر با تعداد نمونه های یکی از کلاس ها باشد و عرض و طول مستطیل ها بسیار کوچک باشد. در این حالت رخداد چه پدیده ای را پیش بینی می کنید؟

ج) در حالت تک مستطیلی، فرض کنید فردی حضور دارد که به ازای تمام مقادیر ورودی، برچسب مربوطه (مثبت یا منفی) را می داند. اگر شما اجازه داشته باشید فقط یک بار کلاس ورودی دلخواه را از او سوال کنید کدام ورودی است؟ ورودی دو بعدی است و جواب یکتا نیست و به صورت یک ناحیه است (یعنی هر نقطه از این ناحیه می تواند جواب این مسئله باشد)، آن ناحیه را بر روی شکل نشان دهید.

- **مسئله ۵**

برای یک مسئله دو کلاسه اثبات کنید خروجی y1 و y2 برای دو حالت زیر یکسان است (یعنی نشان دهید y1=y1 و y2=y2)

- y1=softmax(o1), y2=softmax(o2)
- y1=sigmoid(o1-o2), y2=1-y1

- **مسئله ۶**

روابط بروزرسانی وزن ها در الگوریتم گرادیان کاهشی را برای یک مدل Logistic regression با تابع خطای MSE بنویسید. مسئله را دو کلاسه و ورودی را دو بعدی فرض کنید. وزن ها شامل w1، w2 و b هستند.

پیاده سازی

مسئله ۱

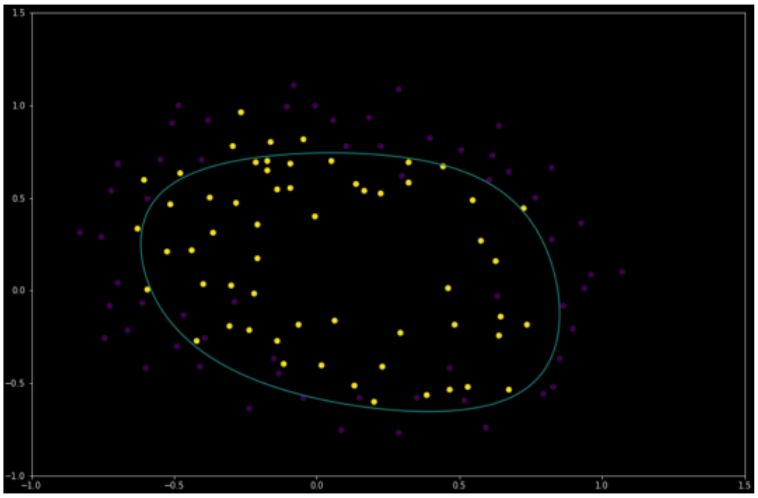
در این مسئله بر روی مجموعه داده ی quality_test.csv کار خواهید کرد که در آن دو ستون اول نتایج تست یک چیپ و ستون سوم نشان دهنده قبول یا رد کیفیت آن چیپ است.

با استفاده از الگوریتم logistic regression دو کلاس این مجموعه داده را جدا خواهید کرد. همانطور که در شکل زیر معلوم است، این مجموعه داده به صورت خطی جداپذیر نیست. بنابراین بایستی ابتدا فضای ویژگی ها را به مرتبه ی بالاتر برد. تابعی که برای این کار پیاده سازی خواهید کرد، عملیات زیر را انجام خواهد داد.

$$X = [x_1 \ x_2]^T \ , \quad f(X) = [x_1 \ \ x_2 \ \ x_1^2 \ \ x_1x_2 \ \ x_2^2 \ \ x_1^3 \ \ x_1^2x_2 \ \ x_1x_2^2 \ \ x_2^3 \ \ \dots \ \ x_1x_2^5 \ \ x_2^6]$$
$$f(X) \ : \mathbb{R}^2 \rightarrow \mathbb{R}^{27}$$

- فقط برای پیاده سازی الگوریتم logistic regression مجاز به استفاده از دستور آماده sklearn هستید.
- نیاز به تقسیم بندی داده به ترین و تست نیست.

در انتها دقت طبقه بند خود را بر روی همین داده گزارش کرده و مرز تصمیم گیری بدست آمده توسط الگوریتم خود را رسم کنید. نمودار شما بایستی چیزی شبیه به شکل زیر باشد.



مسئله ۲

مجاز به استفاده از دستوراتی که کل الگوریتم را در یک خط پیاده می کند نیستید. تلاش شود تا حداکثر فقط از numpy و matplotlib.pyplot استفاده شود.

الف) تعداد ۱۰۰ نقطه یک بعدی به عنوان ورودی از بازه ۰ تا ۱۰ بسازید و آن را در x قرار دهید. نقاط فواصل یکسان از یکدیگر داشته باشند.

ب) نقاط y را بر اساس رابطه $y=3x+2$ حاصل کنید و نمودار y بر حسب x را رسم کنید.

پ) نویز تصادفی بر آمده از توزیع گوسی با میانگین ۰ و واریانس ۰٫۸ را به y اضافه کنید (مراحل بعدی مسئله با y نویزی پیش خواهد رفت). ۲۰ درصد دادگان را به عنوان داده تست در نظر بگیرید و دادگان آموزش را رسم کنید (y برحسب x).

ت) می خواهیم مدلی **خطی** را برای y تخمین بزنیم که به ازای هر ورودی x ممکن خروجی تخمینی را به ما بدهد. این کار به کدام مسئله از درس اشاره دارد؟

ث) با فرض در نظر گرفتن تابع خطای MSE روابط بروز رسانی وزن ها را بر اساس گرادیان کاهشی بنویسید.

ج) الگوریتم گرادیان کاهشی را بر روی مسئله برای یافتن وزن های مناسب با تکرار کافی پیاده کنید (نرخ یادگیری مناسب را خود بیابید). هم چنین هنگام آموزش خطا را در هر تکرار ذخیره کنید و در نهایت نمودار خطا به ازای تکرار را نمایش دهید.

چ) خطای دادگان تست را در مقایسه با مدل تخمینی بیابید و سپس مدل تخمینی را به همراه نقاط تست بر روی **یک نمودار** نمایش دهید.

مطالعاتی

مقاله ی پیوست شده را مطالعه کنید و برداشت خود از آن را شرح دهید (حداقل ۲ صفحه).

موفق باشید 😊