به نام خدا

دانشکده مهندسی برق و کامپیوتر

تمرین سری دوم یادگیری ماشین

دانشگاه تهران

پردیس دانشکده های فنی

سلام بر دانشجویان عزیز، چند نکته مهم:

1. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، در حد نیاز توضیح دهید.
2. نکته ی مهم در گزارش نویسی روشن بودن پاسخ ها می باشد، اگر فرضی برای حل سوال استفاده می‌کنید حتما آن را ذکر کنید، اگر جواب نهایی عددی است به صورت واضح آن را بیان کنید.
3. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده کنید.
4. فایل نهایی خود را در یک فایل زیپ شامل، pdf گزارش و فایل کدها آپلود کنید. نام فایل زیپ ارسالی الگوی ML_HW2_StudentNumber داشته باشد.
5. از بین سوالات **شبیه سازی** حتما به هر سه مورد پاسخ داده شود.
6. نمره تمرین ۱۰۰ نمره می باشد و حداکثر تا نمره ۱۱۰ (**۱۰ نمره امتیازی**) می توانید کسب کنید.
7. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می باشد و کل تمرین برای طرفین **صفر** خواهد شد.
8. در صورت داشتن سوال، از طریق ایمیل maedehtoosi@gmail.com، سوال خود را مطرح کنید.

سوال ۱: (۲۰ نمره)

بخش های زیر را به صورت تحلیلی و با ذکر محاسبات خود پاسخ دهید:

الف: فرض کنید توزیع احتمال توام P_{XY} را می دانیم و فرم بهینه $f^*: X \to Y$ که MSE را به حداقل می رساند برابر است با:

$$f^* = \arg\min_f E\left[(f(X) - Y)^2\right]$$

نشان دهید $f^*(X) = E[Y|X]$.

ب: یکی از راه‌های گسترش رگرسیون لجستیک به مجموعه های چند کلاسه (K class labels) ، در نظر گرفتن مجموعه‌های (K-1) از بردارهای وزن و تعریف

$$P(Y = y_k|X) \propto exp(\omega_{k0} + \sum_{i=1}^{d} \omega_{ki}X_i), \qquad for\ k = 1, \ldots, K-1$$

می باشد. این تعریف چه مدلی را برای $P(Y=y_k|X)$ نشان می دهد؟

ج: قانون طبقه بندی در مورد ب چه خواهد بود؟

سوال ۲: (۱۵ نمره)

الف: L1 Regularization و L2 Regularization را تعریف کنید و تفاوت های آنها را بیان کنید.

ب: اگر تابع هدف برای رگرسیون خطی L2-regularized برابر با:

$$J(\mathrm{w}) = \| X\mathrm{w} - \mathrm{y} \|_2^2 + \lambda \| \mathrm{w} \|_2^2$$

باشد که در آن ردیف های ماتریس X نقاط داده هستند و مینیمم کننده J برابر با

$$\mathrm{w}^* = (X^T X + \lambda I)^{-1} X^T \mathrm{y}$$

باشد، حال با در نظر گرفتن روش نیوتن برای به حداقل رساندن J ، اگر w_0 یک حدس اولیه دلخواه برای روش نیوتن باشد، نشان دهید که w_1 (مقدار وزن‌های بعد از یک گام نیوتن)، برابر با w^* است.

سوال ۳: (۲۵ نمره)

جدول زیر مربوط به یک مساله رگرسیون خطی ساده می باشد و در آن مقادیر y و x بیانگر نمره و میزان مطالعه می باشد، با توجه به آن به قسمت الف و ب پاسخ دهید.

i	x_i	y_i
1	16	46
2	27	80
3	11	36
4	20	52
5	30	98
6	25	75
7	5	10
8	24	70
9	21	64
10	10	30

الف: در رابطه رگرسیون خطی ساده، عبارت شیب خط را در نظر نگیرید و در مدل زیر β_0 را بیابید.

$$y = \beta_0 + \varepsilon_i$$

ب: در رابطه رگرسیون عبارت عرض از مبدا را در نظر نگیرید و در مدل زیر β_1 را بیابید.

$$Y_i = \beta_1 x_i + \varepsilon_i$$

ج: تفاوت دو معادله زیر را توضیح دهید.

$$\hat{Y} = b_0 + b_1 X$$
$$Y = \beta_0 + \beta_1 X + \epsilon$$

اگر تجزیه و تحلیل رگرسیون مربوط به نمرات آزمون (y) به میزان مطالعه (x) عبارت $\widehat{y} = 25 - 0.5x$ را ایجاد کند :

د : یک نمره آزمون اضافی برای مشاهده جدید در x = 6 به دست آمده است. آیا نمره آزمون برای مشاهده جدید لزوماً 22 خواهد بود؟ دلیل خود را توضیح دهید.

هـ : اگر مجموع مربعات خطا (SSE) برای این مدل 7 باشد و به اندازه n = 16 مشاهده وجود داشته باشد ، بهترین تخمین را برای σ^2 ارائه دهید.

سوال ۴: (شبیه سازی، ۱۵ نمره)

هدف از انجام این سوال بررسی Overfitting و Underfitting برای یک سری داده می باشد.

ابتدا با توجه به کد زیر داده های مربوطه را تولید کنید:

```
x = np.arange(-10 , 10 , 0.2)
y = 2 * cos(x)/-pi + 2 * sin(2 * x)/(2 * pi) + 2 * cos(3 * x)/(-3 * pi)
```

سپس y این داده ها در حالت اول با نویز گاوسی سفید جمع کنید و در حالت دوم با نویز پواسون با $\lambda = 2$ جمع کنید. نویزها را با ضریب تاثیر 0.12 به داده ها اضافه کنید. حال سعی کنید که تابع درجه 1 تا 15 را برای این داده ها برازش کنید.

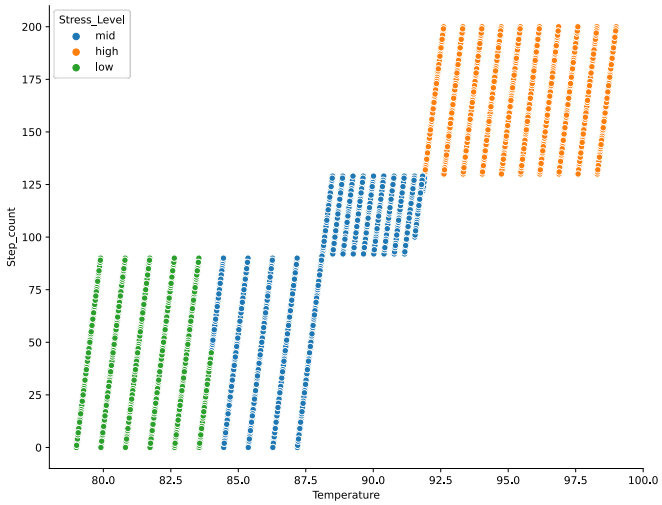
الف: مشخص کنید بهترین درجه و بدترین درجه کدام می باشد.

ب: برای بهترین درجه و درجات 4،1، 7 و 15 نمودار برازش شده را رسم کنید و مقادیر MSE برای هر یک از موارد گزارش دهید.

ج : مشاهدات خود را از نتایج به دست آمده شرح دهید. همچنین مقادیر بایاس و واریانس را ذکر کنید.

سوال ۵: (شبیه سازی، ۱۵ نمره)

در این سوال بر روی دیتاست Stress-Lysis که ضمیمه شده است کار خواهید کرد.



شکل 1، نمایش دیتاست بر اساس دو ویژگی Step count و Temperature

الف: ابتدا برای درک بهتر این دیتاست نمودار نقاط آن را بر حسب هر دو تایی از ویژگی ها رسم کنید. حتما اسامی ویژگی ها را بر روی نمودار مشخص کنید (مطابق شکل 1). حال از بین این نمودار ها مشخص کنید که یک طبقه بند خطی بر حسب کدام ویژگی می تواند با دقت بیشتری کلاس ها را جدا نماید.

ب: داده ها را به صورت تصادفی و با نسبت مشخص به داده های آموزش و آزمون تفکیک کنید (80 درصد برای داده های آموزش و 20 درصد برای داده های آزمون) و یک طبقه بند چند کلاسه با استفاده از **Logistic Regression** و تکنیک one against all پیاده سازی کنید و داده های آزمون را توسط آن کلاس بندی کنید. دقت طبقه بند، confusion matrix، معیار Jaccard و f1-score را گزارش کنید.

ج: تک تک گام های قبل شامل جداسازی داده، پیاده سازی طبقه بند و .. را توسط پکیج های آماده یادگیری ماشین (ترجیحا scikit-learn) انجام دهید و با استفاده از آنها نتایج ذکرشده را گزارش کنید. همچنین با استفاده از این پکیج ها نمودار ROC را برای هرکلاس در یک نمودار رسم کرده و مساحت سطح زیر آن را گزارش کنید.

*** دقت کنید که برای بخش‌های الف و ب این سوال، امکان استفاده از پکیج های یادگیری ماشین را ندارید.

سوال ۶: (شبیه سازی، ۲۰ نمره)

در این سوال برای دو حالت زیر اقدام به تولید داده می کنیم. در هر دو حالت دو دسته نقطه با مختصات (X, Y) داریم.

حالت اول:

دسته اول شامل 200 نقطه درون دایره ای به مرکز (1.5,0) محدود به شعاع های 4 و 9 می باشد.

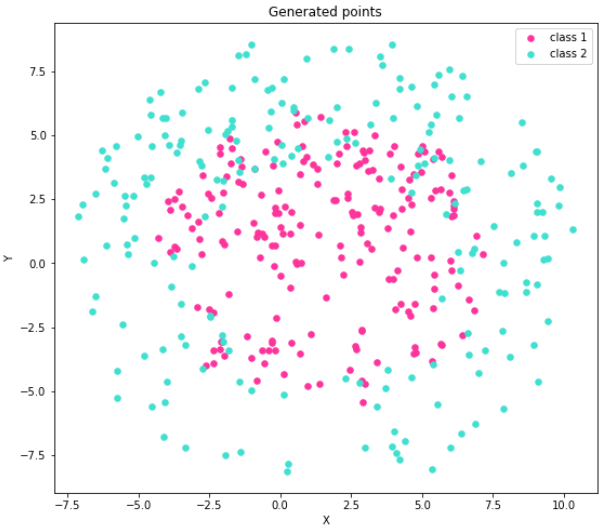
دسته دوم شامل 200 نقطه درون دایره ای به مرکز (1.5,0) محدود به شعاع های 0 و 6 می باشد.

حالت دوم:

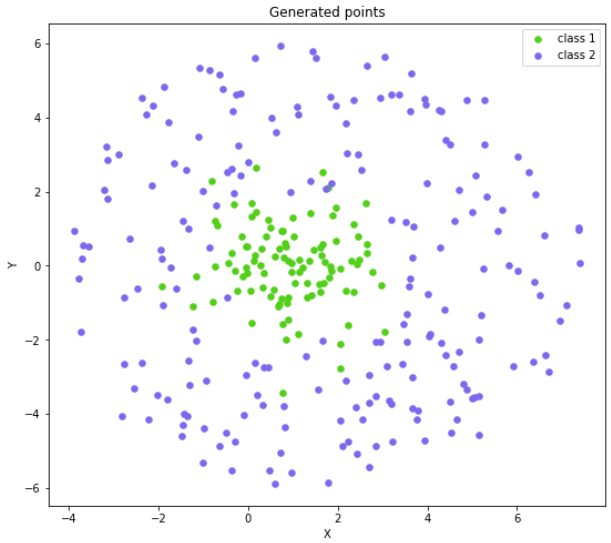
دسته اول شامل 100 داده با میانگین (1,0) برای (X, Y) ، نقاط به صورت تصادفی با فرض انحراف از معیار 1 تولید می شوند.

دسته دوم شامل 200 نقطه درون دایره ای به مرکز (1.5,0) محدود به شعاع های 2 و 6 می باشد.

الف: نمودار ها را برای هر دو حالت رسم کنید. نتیجه چیزی حدودا شبیه به شکل ۲و۳ خواهد بود.



شکل 2 ، نمایش داده های حالت اول



شکل 3 ، نمایش داده های حالت دوم

ب: با استفاده از الگوریتم Logistic Regression و استفاده از L2 Regularization دو کلاس این مجموعه داده را جدا کنید. همانطور که در شکل ها مشخص است، این مجموعه داده ها به صورت خطی جداپذیر نیستند. بنابراین باید ابتدا فضای ویژگی ها را به مرتبه‌ی بالاتر برد. تابعی که برای این کار پیاده سازی خواهید کرد، عملیات زیر را انجام خواهد داد که در مثال زیر ابعاد از 2 به 35 افزایش یافته است.

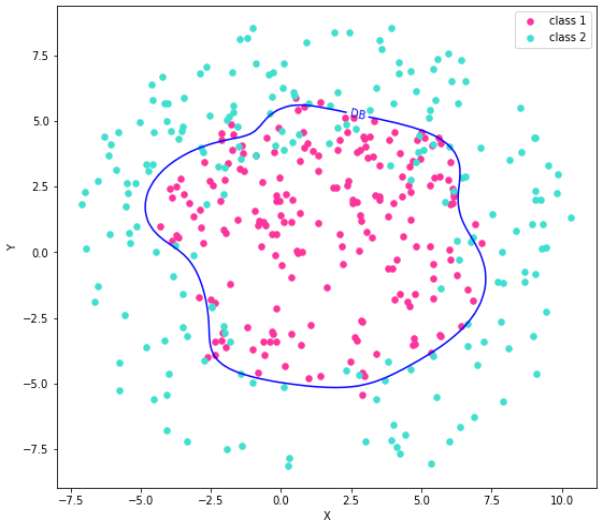
$$X = [x_1, x_2]^{\mathrm{T}}$$

$$f(X) = [x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3, \dots, x_1x_2^6, x_2^7]^T, f: R^2 \to R^{35}$$

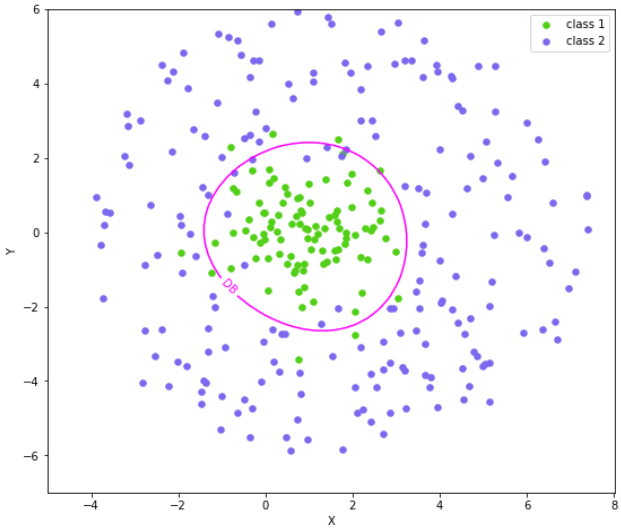
در ادامه دقت طبقه بند خود را بر روی داده های هر دو حالت گزارش کنید و مرز تصمیم گیری به دست آمده توسط الگوریتم خود را رسم کنید. در هر دو حالت بهترین درجه ای که با آن مرز تصمیم رسم شده است را گزارش کنید. شکل حاصل باید حدودا مشابه شکل 4 و 5 باشد.

ج : نتایج به دست آمده در هر دو حالت را تحلیل کنید.

*** درصورت استفاده از پکیج های آماده ی یادگیری ماشین، نصف نمره ی این سوال را خواهید گرفت.



شکل 4 ، نمایش مرز تصمیم گیری برای داده های حالت اول



شکل 5 ، نمایش مرز تصمیم گیری برای داده های حالت دوم