سلام بر دانشجویان عزیز، چند نکته مهم:

۱. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، در حد نیاز توضیح دهید.

۲. نکته‌ی مهم در گزارش نویسی روشن بودن پاسخ‌ها می‌باشد، اگر فرضی برای حل سوال استفاده می‌کنید حتما آن را ذکر کنید، اگر جواب نهایی عددی است به صورت واضح آن را بیان کنید.

۳. کدهای ارسال شده **بدون گزارش** فاقد نمره می‌باشند.

٤. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده کنید.

٥. 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.

٦. از بین سوالات **شبیه سازی** حتما به هر دو مورد پاسخ داده شود.

۷. نمره تمرین ۱۰۰ نمره می‌باشد و حداکثر تا نمره ۱۱۰ ( **۱۰ نمره امتیازی**) می توانید کسب کنید.

۸. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می‌باشد و کل تمرین برای طرفین **صفر** خواهد شد.

۹. در صورت داشتن سوال، از طریق ایمیل z.habibzadeh213@gmail.com، سوال خود را مطرح کنید.

## سوال ۱: (۱۵ نمره)

در مدل رگرسیون $D$ بعدی زمانی که می‌خواهیم برچسب $x_i$ در داده آموزش را به دست آوریم ، از تابع خطی زیر استفاده می‌کنیم:

$$f(X_i; \theta_0, \theta_1, \theta_2, \dots, \theta_D) = \sum_{d=1}^{D} \theta_d . x_i(d) \; + \; \theta_0$$

در فرمول بالا، منظور از $x_i(d)$ المان $d$ام $x_i$ می‌باشد. در این حالت $x_i$ یک داده‌ی بعدی $D$ است. فرض می‌کنیم که $y_i$ برچسب واقعی $x_i$ باشد و تعداد کل داده‌های آموزش $N$ باشد. پارامترهای مدل $(\theta_0, \theta_1, \dots, \theta_D)$ با یافتن مقدار کمینه تابع هزینه زیر بدست می‌آید:

$$R(\theta) = \frac{1}{2N} ||y - X\theta||_2^2 + \; \theta^T H \theta + \theta^T \theta + \; a^T \theta$$

$a$ یک بردار، $H$ یک ماتریس و $H = H^T$ است. هر دو مقدار $H$ و $a$ را داریم. مقادیر بهینه پارامترهای مدل را محاسبه کنید. همه مراحل را توضیح دهید.

## سوال ۲: (۲۰ نمره)

**الف)** L1 Regularization و L2 Regularization (Ridge Regularization) را با ذکر فرمول هایشان، تعریف کنید و تفاوت‌های آنها را بیان کنید. (به ۴ مورد از تفاوت‌ها اشاره کنید)

**ب)** یک رگرسیون خطی با L2 Regularization به صورت عبارت زیر در نظر بگیرید:

$$\hat{\beta} = arg\underbrace{min}_{\beta}\sum_{i=1}^{n}(Y_i - X_i\beta)^2 + \lambda\,||\beta||_2^2$$

$\lambda \geq 0$ پارامتر L2 Regularization و $X_i = [X_i^{(1)} \dots X_i^{(p)}]$ است. با فرض $Y = [Y_1; \dots; Y_n]$ , $A = [X_1; \dots; X_n]$ **ثابت کنید** جواب فرم بسته $\hat{\beta}$ برابر عبارت زیر است:

$$\hat{\beta} = (A^TA + \lambda I)^{-1}A^TY$$

## سوال ۳: (۱۵ نمره)

یکی از راه‌های گسترش Logistic Regression به مجموعه‌های چند کلاسه مثلاً برچسب‌های کلاس $K$، این است که مجموعه‌های $(K-1)$ از بردارهای وزن را در نظر بگیریم و تعریف کنیم:

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

**الف)** این تعریف چه مدلی را برای $P(Y = y_k|X)$ نشان می‌دهد؟

**ب)** قانون طبقه بندی در مورد قسمت الف چه خواهد بود؟

## سوال ۴: (۱۵ نمره)

**بخش‌های مختلف این سوال را با محاسبات خود پاسخ دهید (برای این سوال، کد نزنید، تمامی محاسبات باید به صورت دستی انجام گیرد).

در جدول زیر، نمره‌های میان ترم دانشجویان درس یک کلاس برحسب میزان ساعت مطالعه آنها آورده شده است، که یک مسئله رگرسیون خطی[1] ساده می‌باشد.

| **میزان ساعت مطالعه ($x_i$)** | ۴ | ۹ | ۱۰ | ۱۴ | ۴ | ۷ | ۱۲ | ۲۲ | ۱ | ۱۷ |
|---|---|---|---|---|---|---|---|---|---|---|
| **نمره ($y_i$)** | ۳۱ | ۵۸ | ۶۵ | ۷۳ | ۳۷ | ۴۴ | ۶۰ | ۹۱ | ۲۱ | ۸۴ |

**الف)** مقدار پارامترهای $\beta_0$،$\sigma^2$ و$\beta_1$ را محاسبه کنید.

**ب)** واریانس مربوط به $\beta_0$ و$\beta_1$ را محاسبه کنید.

**ج)** مقدار کوریلیشن[2] مربوط به دو پارامتر $\beta_0$ و$\beta_1$ را بدست آورید.

[1] Linear Regression
[2] Correlation

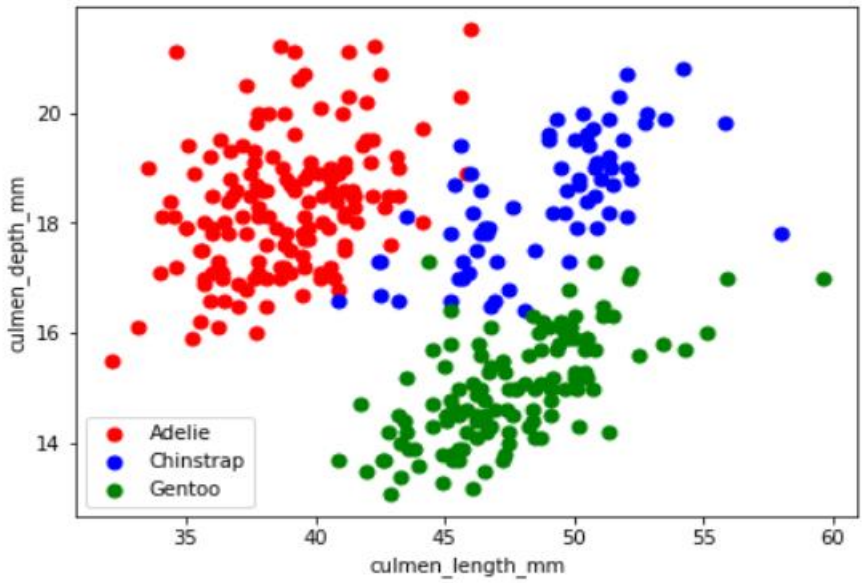
** در سوالات شبیه سازی، در صورت لزوم، پیش پردازش‌های ممکن باید انجام گیرد.

## سوال ۵: (شبیه سازی، ۲۵ نمره)

در این سوال با دیتاست penguins کار خواهید کرد (دیتاست ضمیمه شده است). در این دیتاست داده مربوط به ۴ ویژگی مختلف سه گونه پنگوئن (Chinstrap, Adelie, Gentoo ) فراهم شده است.

** برای بخش ب امکان استفاده از پکیج‌های یادگیری ماشین را ندارید.

**الف) ابتدا** برای درک بهتر این دیتاست، نمودار نقاط[3] آن را بر حسب هر دو تایی از ویژگی‌ها، رسم کنید. **(حتما** اسامی ویژگی‌ها را بر روی نمودار مشخص کنید به عبارت دیگر توجه کنید که محورها، لیبل مناسب داشته باشد. برای داده‌های سه گونه مختلف، سه رنگ مختلف در نظر بگیرید و با legend آن‌ها را مشخص کنید (مطابق شکل ۱)). **حال** از بین این نمودارها مشخص کنید که طبقه بندی خطی بر حسب کدام دو ویژگی، می‌تواند با دقت بیشتری کلاس‌ها را جدا نماید؟ **سپس** در ادامه‌ی این سوال، برای **هر** ویژگی، توزیع کلاس‌های مختلف را با استفاده از هسیتوگرام[4]، رسم کنید.




شکل ۱- کلاس بندی داده‌های penguin بر ا اساس دو ویژگی culmen_depth_mm , culmen_length_mm

[3] Scatter plot
[4] Histogram

**ب)** داده‌ها را به صورت تصادفی و با نسبت مشخص به داده‌های آموزش و آزمون تفکیک کنید و یک طبقه بند چندکلاسه با استفاده از **Logistic Regression** و تکنیک one against all پیاده سازی کنید. دقت طبقه بند، precision، confusion matrix، معیار Jaccard ، f1-score و Recall را گزارش کنید.

**ج)** تک تک گام‌های قبل ( جداسازی داده ، پیاده‌سازی طبقه بند و ...) را توسط پکیج‌های آماده‌ی یادگیری ماشین(scikit-learn) انجام دهید و نتایج را گزارش نمایید. همچنین به کمک این پکیج‌ها نمودار ROC برای هر کلاس در یک نمودار رسم کرده و مساحت سطح زیر آن را نیز گزارش نمایید.

## سوال ۶: (شبیه سازی، ۲۰ نمره)

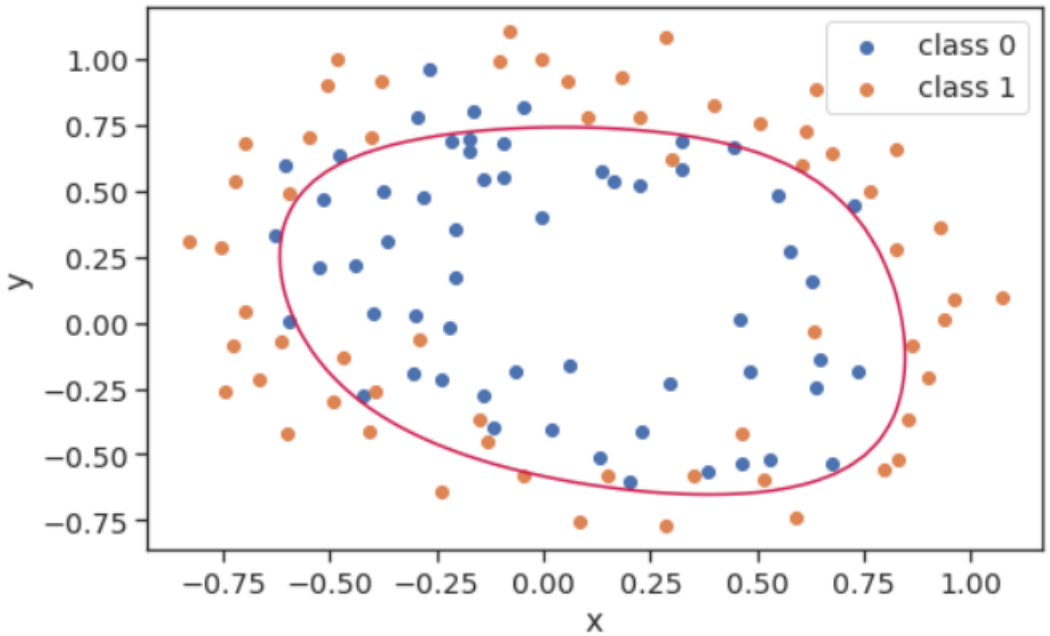
در این سوال، شما با دیتاست Quality.csv (دیتاست ضمیمه شده است) کار خواهید کرد که این مجموعه‌ی داده به بررسی کیفیت چیپ‌های موجود براساس نتایج تست، در یک آزمایشگاه می‌پردازد که دو ستون اول، نتایج تست و ستون سوم نشاندهنده‌ی قبول یا رد کیفیت چیپ است.

با استفاده از الگوریتم Logistic Regression و L2 Regularization دو کلاس این مجموعه داده را جدا کنید. همانطور که در شکل ۲ مشخص است، این مجموعه داده‌ها به صورت خطی جداپذیر نیست. بنابراین بایستی ابتدا فضای ویژگی‌ها را به مرتبه‌ی بالاتر برد. تابعی که برای این کار پیاده سازی خواهید کرد، عملیات زیر را انجام خواهد داد.

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

در انتها دقت طبقه بند خود را بر روی همین داده‌ها گزارش و تحلیل کرده و مرز تصمیم گیری به دست آمده توسط الگوریتم خود را رسم کنید. شکل شما باید حدودا شبیه شکل ۲ شود.

** درصورت استفاده از پکیج آماده‌ی یادگیری ماشین، نصف نمره‌ی این سوال شبیه سازی را خواهید گرفت.




شکل۲- نمایش مرز تصمیم گیری