باسمه تعالی

یادگیری ماشین

تمرین دوم

آبان ۹۹

لطفا در پاسخ‌گویی به سوالات این تمرین، نکات زیر را مدنظر قرار دهید.

- از اطناب پرهیز کنید و پاسخ‌ها را تا حد ممکن خلاصه و شفاف ذکر کنید. نکات نگارشی را رعایت و پاسخ هر سوال را در صفحه‌ی جداگانه تایپ کنید.

- سوالات تحلیلی را **حتما** به‌صورت تایپ‌شده و فقط در قالب فایل pdf تحویل دهید. برای نگارش روابط ریاضی می‌توانید از روابط عکس گرفته و در فایل pdf قرار دهید.

- کدهای مربوط به سوالات کامپیوتری را حتما ضمیمه کنید. گزارش و تحلیل این سوالات بدون انضمام کد نمره‌ای نخواهد داشت.

- کدها فقط می‌توانند در زبان‌های **پایتون** باشند. تنها فرمت مورد قبول فرمت py. می‌باشد. لطفا از آپلود سایر فرمت‌ها (مانند ipynb. و ...) خودداری کنید. هم‌چنین برای هر سوال (یا بخش)، فایل مربوط به آن سوال را جداگانه و با نام خود سوال ضمیمه کنید.

- برای هر کد که در فایل نهایی ضمیمه می‌کنید، گزارش بنویسید. کدهای ضمیمه‌شده بدون گزارش مربوطه نمره‌ای نخواهند داشت. (این گزارش‌ها تنها معیار تفکیک کد شما و کدهای موجود در منابع مختلف مانند اینترنت خواهند بود.)

- عکس‌ها را به‌صورت واضح و همراه با زیرنویس در گزارش خود بیاورید.

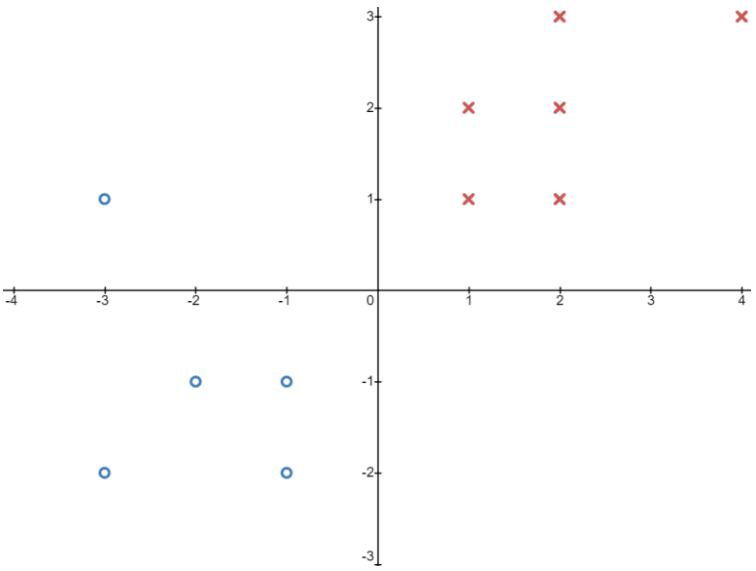
- فایل نهایی خود را در یک فایل زیپ شامل یک فایل pdf گزارشات و فایل کدهای خود آپلود کنید. نام فایل زیپ **حتما** الگوی [ml-hw2-SID] داشته باشد.

- هرگونه مشابهت غیر منطقی در گزارش و کد دانش‌جویان، تقلب محسوب شده و نمره‌ی ۰ برای طرفین منظور خواهد شد. هم‌چنین دقت کنید که نمره‌ی کل این تمرین از ۱۲۰ است که ۲۰ نمره‌ی آن امتیازی می‌باشد.

- در صورت داشتن هرگونه سوال راجع‌به این تمرین، با ایمیل rahimiazghan@gmail.com در ارتباط باشید.

۱ سوال اول (۱۰ نمره)

داده‌های زیر را که به دو کلاس تقسیم شده‌اند را در نظر بگیرید. با فرض توزیع گاوسی برای هر کلاس، معادله‌ی مرز تصمیم مابین این دو کلاس را محاسبه کنید. تمامی مقادیر لازم (میانگین، احتمال پیشین و ..) را از مشاهدات زیر به‌دست آورید. (بعد از محاسبه‌ی مرز، می‌توانید برای تحقیق درستی جواب خود، نمودار آن را در سایت *desmos.com* رسم کنید.)



۲ سوال دوم (۱۰ نمره)

مجموعه داده‌ای ۳بعدی با توزیع گاوسی را فرض کنید که در آن $\underline{\mu} = \begin{bmatrix} ۱ \\ ۲ \\ ۱ \end{bmatrix}$ و $\underline{\underline{\Sigma}} = \begin{bmatrix} ۱ & ۰ & ۰ \\ ۰ & ۸ & ۲ \\ ۰ & ۲ & ۵ \end{bmatrix}$ می‌باشند.

الف) احتمال تعلق نقطه‌ی $\underline{x} = [۲ \quad ۰.۵ \quad ۳]^T$ به این مجموعه داده را محاسبه کنید.

ب) با استفاده از روابط موجود در اسلایدها، ماتریس تبدیل سفیدسازی[1] (A_w) را محاسبه کنید.

ج) با اعمال تبدیل مناسب بر روی این مجموعه داده، توزیع آن را به‌صورت $p(\underline{x}) \sim N(\underline{0}, \underline{\underline{I}})$ در آورده و نقطه‌ی $\underline{x}$ را تحت آن تبدیل به نقطه‌ی $\underline{y}$ برده و مختصات $\underline{y}$ را ذکر کنید.

د) نشان دهید که فاصله‌ی ماهالانوبیس نقطه‌ی $\underline{x}$ از $\underline{\mu}$ برابر فاصله‌ی اقلیدسی $\underline{y}$ از مبدا است.

ه) به‌صورت کلی اثبات کنید که تحت تبدیل سفیدسازی، فاصله‌ی اقلیدسی در فضای ثانویه برابر فاصله‌ی ماهالانوبیس در فضای اولیه است.

[1] Whitening Transformation

۳ سوال سوم (۱۵ نمره)

در برخی از الگوریتم‌های دسته‌بندی، علاوه بر امکان دسته‌بندی یک نقطه در c کلاس موجود، می‌توان بسته به شرایط، یک نقطه را در هیچ‌کدام از کلاس‌های موجود دسته‌بندی نکرد و آن را رد نمود. به‌عبارتی دیگر، اگر هزینه‌ی رد کردن یک نقطه، معقول بوده و کم‌تر از هزینه‌ی اطلاق آن به هرکدام از کلاس‌های موجود باشد، نقطه را رد می‌کنیم.

فرض کنید داریم

$$\lambda(\alpha_i, \omega_j) = \begin{cases} 0 & i = j \qquad 1 \le i,j \le c \\ \lambda_r & i = c+1 \\ \lambda_s & \text{otherwise} \end{cases}$$

در رابطه‌ی بالا، $\lambda(\alpha_i, \omega_j)$ هزینه‌ی عمل α_i است هنگامی که در کلاس ω_j قرار گرفته‌ایم که در آن α_1 تا α_c عمل دسته‌بندی نقطه در کلاس ۱ تا c و α_{c+1} عمل رد کردن نقطه است.

الف) اثبات کنید که کم‌ترین ریسک در انتخاب کلاس ω_i زمانی رخ می‌دهد که دو شرط زیر را برای همه‌ی jها به‌صورت همزمان داشته باشیم: $P(\omega_i|x) \ge P(\omega_j|x)$ و $P(\omega_i|x) \ge 1 - \frac{\lambda_r}{\lambda_s}$

ب) اثبات کنید که تابع زیر، یک تابع افتراق‌ساز[2] بهینه برای همچین مسایلی است.

$$g_i(x) = \begin{cases} p(\underline{x}|\omega_i)P(\omega_i) & i = 1, \dots, c \\ \dfrac{\lambda_s - \lambda_r}{\lambda_s} \displaystyle\sum_{j=1}^{c} p(x|\omega_j)P(\omega_j) & i = c+1 \end{cases}$$

ج) با استفاده از نتایج بخش‌های قبل، تحلیل کنید که اگر نسبت λ_r به λ_s از ۰ تا ۱ تغییر کند به مرور چه تغییراتی در نحوه‌ی دسته‌بندی رخ خواهد داد؟

[2] Discriminant Function

۴ سوال چهارم (۱۵ نمره)

با استفاده از روش ضرایب لاگرانژ، فاصله‌ی نقطه‌ای مانند $\underline{x_0}$ را از ابر صفحه‌ی $\underline{w}^T\underline{x} + b = 0$ به دست آورید.

مجددا با استفاده از ضرایب لاگرانژ و ایده‌ای که از بخش قبل گرفتید، فاصله‌ی نقطه‌ی $\underline{x_0}$ را از بیضی‌گون $\underline{x}^T\underline{\underline{A}}\,\underline{x} = 1$ بیابید. اگر به نظرتان فرم بسته‌ای به‌عنوان راه‌حل موجود نیست، سعی کنید روشی را برای حل عددی این سوال ارائه بدهید. (به کاهش گرادیان فکر کنید) هرچند بایستی تا جای ممکن روابط موجود را ساده کرده و سپس این روش را ارائه دهید.

برای شرایط زیر، مختصات نزدیک‌ترین نقطه روی بیضی‌گون به نقطه‌ی داده شده را حساب کنید. در این‌جا از روش تحلیلی (و نه از روش عددی که در بخش قبل به‌دست آوردید) استفاده کنید. (برای محاسبات سنگین می‌توانید از ابزارهای حل معادله استفاده کنید و برای تحقیق پاسخ خود می‌توانید نمودارهای مربوطه را در سایت *desmos.com* رسم کنید.)

- $\underline{x_0} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \ \underline{\underline{A}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
- $\underline{x_0} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \ \underline{\underline{A}} = \begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix}$

نکته: در این سوال تنها مجاز به استفاده از روش‌های بهینه‌سازی مبتنی بر ضرایب لاگرانژ هستید. استفاده از سایر روش‌ها نمره‌ای نخواهد داشت.

۵ سوال پنجم (۵ نمره)

به سوالات زیر پاسخ کوتاه دهید.

الف) با رسم یک نمودار دلخواه، تاثیر تغییر نسبت هزینه‌های zero-one در حالت باینری ($\frac{\lambda_{12}}{\lambda_{21}}$) را مشاهده کرده و شهودی را که از تغییر این نسبت می‌گیرید توضیح دهید.

ب) مفهوم خطای قابل کاهش[3] را با رسم یک شکل برای دسته‌بند دودویی توضیح دهید.

ج) مفهوم discriminability را توضیح داده و یک معیار برای آن معرفی کرده و شهود خود از آن معیار را توضیح دهید.

د) مفهوم منحنی ROC را توضیح داده و بررسی کنید که آیا این منحنی همواره اکیدا صعودی است یا خیر.

ه) تفاوت رویکردهای discriminative با generative را در مسائل طبقه‌بندی توضیح دهید.

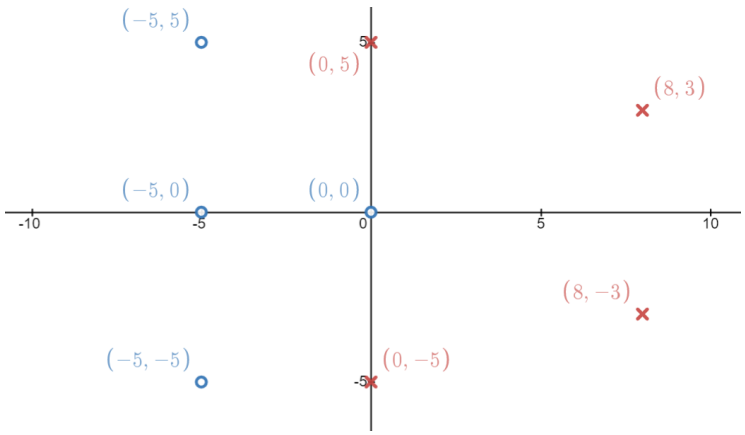
[3] Reducible error

۶ سوال ششم (۲۰ نمره)

سوالات زیر را به‌صورت تحلیلی و با ذکر محاسبات خود پاسخ دهید.

الف) در مساله‌ی دسته‌بندی شکل زیر به روش نزدیک‌ترین همسایه، با معیارهای فاصله‌ی زیر، نتیجه‌ی دسته‌بندی نقطه‌ی $x = [۵ \quad ۰]^T$ کدام کلاس (آبی یا قرمز) خواهد بود؟ (در روش نزدیک‌ترین همسایه، کلاس تعلق‌یافته به یک نقطه‌ی تحت آزمون، کلاس نزدیک‌ترین نقطه در مجموعه‌ی آموزش به این نقطه است.)

- $d_۱(x,y) = \max_i |x_i - y_i|$
- $d_۲(x,y) = \sum_{i=۱}^{d} |x_i - y_i|$
- $d_۳(x,y) = \sum_{i=۱}^{d} (x_i - y_i)^2$



ب) در یک مساله‌ی دسته‌بند دودویی که دسته‌ها دارای توزیع گاوسی هستند، پارامتر هر دسته داده شده است. فاصله‌ی ماهالونوبیس نقطه‌ی $x = [۱.۵ \quad ۱.۵]^T$ را از میانگین هر دسته محاسبه کنید.

- $\mu_۱ = \begin{bmatrix} ۰ \\ ۰ \end{bmatrix}, \Sigma_۱ = \begin{bmatrix} ۳ & -۳ \\ -۳ & ۳.۵ \end{bmatrix}$
- $\mu_۲ = \begin{bmatrix} ۰ \\ ۰ \end{bmatrix}, \Sigma_۲ = \begin{bmatrix} ۳ & ۳ \\ ۳ & ۳.۵ \end{bmatrix}$

ج) بازای چه مقادیری از α، ماتریس $\Sigma = \begin{bmatrix} ۵ & \alpha \\ \alpha & ۴ \end{bmatrix}$ یک ماتریس کوواریانس معتبر است؟

د) در یک مساله‌ی دسته‌بندی باینری، که ریسک شرطی را کمینه می‌کند، مفروضات زیر را داریم. ناحیه‌ی کلاس اول را به‌صورت یک بازه نشان دهید. (λ_{ij} هزینه‌ی تصمیم اشتباه تعلق به ω_i است وقتی کلاس واقعی ω_j باشد.)

- $x \in [۰,۱]$
- $p(x|\omega_۱) = x + \frac{۱}{۲}$, $p(x|\omega_۲) = \frac{۳x^۲}{۴} + \frac{۳}{۴}$, $P(\omega_۲) = \frac{۱}{۴}$
- $\lambda = \begin{bmatrix} ۱ & ۲ \\ ۳ & ۱ \end{bmatrix}$

۷ سوال هفتم (پیاده‌سازی، ۱۵ نمره)

به‌طور مختصر راجع‌به متد Generalized Linear Regression و ارتباط آن با الگوریتم‌های Linear Regression و Logistic Regression توضیح دهید.

برای این سوال مجموعه داده‌ی nyc_cyclist_counts ضمیمه شده است که نشان‌گر تعداد دوچرخه‌سوارهای عبوری از پل بروکلین در تاریخ‌های مختلف و برحسب بالاترین و پایین‌ترین دمای آن روز و درصد رطوبت آن روز است.

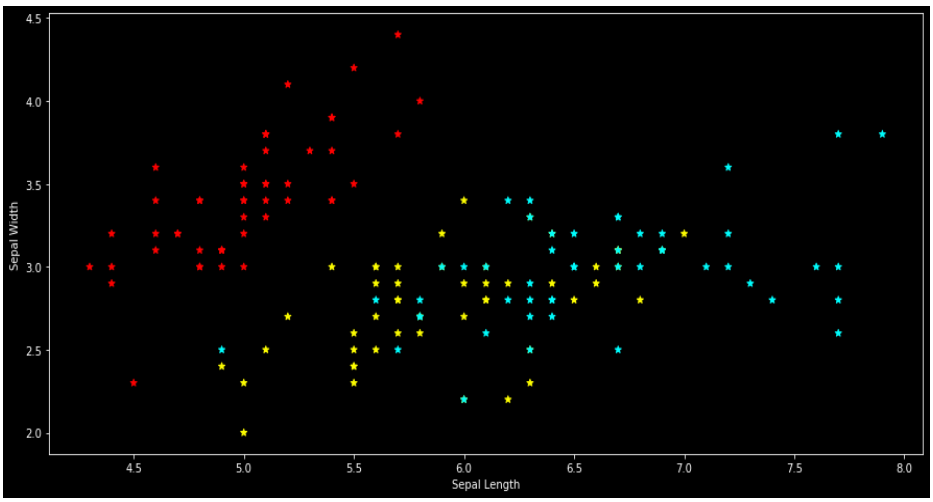
الگوریتم GLM را بر روی مجموعه داده‌ی nyc_cyclist_counts.csv پیاده‌سازی کنید. (مجاز به استفاده از کتاب‌خانه‌های آماده هستید.) هدف الگوریتم شما بایستی این باشد که مقدار BB_COUNT را با استفاده از سایر ستون‌ها تخمین بزند. هم‌چنین فرض کنید که مقادیر ستون BB_COUNT از توزیع پواسون استخراج شده‌اند و سطرها از هم مستقلند. در انتها نمودار مقادیر تخمین‌زده‌شده را کنار مقادیر واقعی بازای هر تاریخ رسم کنید.

دقت کنید که در این سوال گزارشی که می نویسید اهمیت زیادی نسبت به کد دارد. سعی کنید گزارش را کامل و با جزئیات بنویسید.

۸ سوال هشتم (پیاده‌سازی، ۲۰ نمره)

در این سوال بر روی دیتاست iris (که ضمیمه شده است) کار خواهید کرد. برای مطالعه‌ی مختصری راجع‌به این دیتاست، می‌توانید به این‌جا مراجعه کنید. دقت کنید که برای بخش‌های الف تا ج این سوال، امکان استفاده از پکیج‌های یادگیری ماشین را ندارید.

الف) این دیتاست شامل ۴ ویژگی و ۳ کلاس مختلف است. برای درک تصویری بیش‌تر این دیتاست، نمودار پراکندگی آن را برحسب هر دوتایی از ویژگی‌ها رسم کنید. به‌طور مثال، من برای دو ویژگی sepal length و sepal width این نمودار را در شکل ۱ رسم کرده‌ام.



شکل ۱ ــ کلاس‌بندی داده‌های iris برحسب دو ویژگی اول آن‌ها

از روی نمودارهای حاصل، بحث کنید که طبقه‌بندی خطی بر اساس کدام دو ویژگی، دقت بالاتری را به دست خواهد داد؟

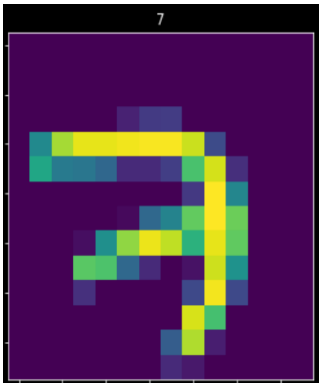
ب) داده‌ها را به‌صورت تصادفی و با نسبت مشخص به داده‌های آموزش و آزمون تفکیک کنید. یک طبقه‌بند نزدیک‌ترین همسایه را پیاده‌سازی کرده و داده‌های آزمون را توسط آن کلاس‌بندی کنید. این‌کار را یک‌بار بدون نرمالایز و یک‌بار با نرمالایز داده‌ها انجام داده و تفاوت در دقت را گزارش کنید.

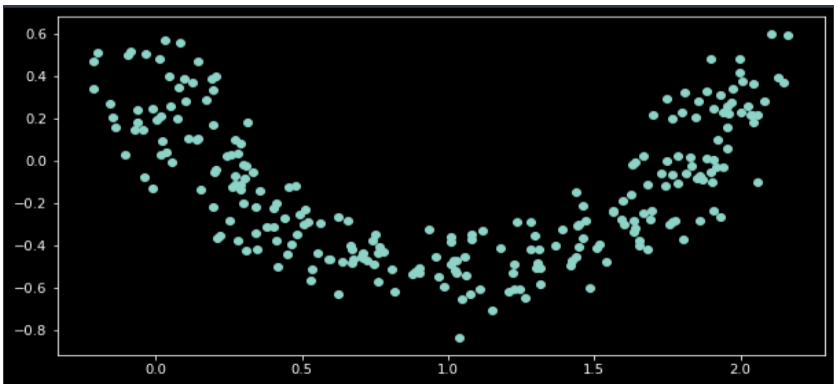
ج) بعد از نرمالایز داده‌ها و طبقه‌بندی، دقت طبقه‌بند، ماتریس آشفتگی (بازای هر کلاس) و مقدار f1-score را گزارش کنید.

د) تک تک گام‌های قبل (جداسازی داده، پیاده‌سازی طبقه‌بند و ...) را توسط پکیج‌های آماده‌ی یادگیری ماشین انجام دهید و با استفاده از آن‌ها نتایج ذکرشده را دریافت و گزارش کنید. هم‌چنین با استفاده از این پکیج‌ها نمودار ROC را (برای هرکلاس در یک نمودار) رسم کرده و مساحت سطح زیر آن را گزارش کنید.

۹ سوال نهم (پیاده‌سازی، ۲۰ نمره)

در قسمت الف تا ج این سوال مجاز به استفاده از هیچ پکیج آماده‌ی یادگیری ماشین نیستید. هم‌چنین توصیه می‌شود برای آشنایی بیش‌تر با مجموعه‌داده‌های استفاده شده آن‌ها را نمایش دهید و با ویژگی‌ها و ساختارهای آن‌ها آشنا شوید. به‌عنوان مثال من در شکل زیر اولین بردار مجموعه داده‌ی tiny mnist و هم‌چنین یک کلاس از noisy moons را در صفحه‌ی ویژگی‌هایش ضمیمه کرده‌ام.





الف) ابتدا توضیح مختصری راجع‌به طبقه‌بندهای naïve bayes و optimal (non naïve) bayes داده و آن‌ها را با فرض گاوسی بودن داده‌های ورودی پیاده‌سازی کنید. دقت کنید که الگوریتم شما نباید هیچ پیش‌فرضی از داده‌های ورودی (اندازه‌ی آن‌ها و ...) داشته باشد و باید برای هر داده‌ای کار کند.

ب) الگوریتم‌هایی که پیاده‌سازی کرده‌اید را برای مجموعه داده‌ی tiny mnist (که نسخه‌ی فشرده‌ای از مجموعه داده‌ی mnist است) تست کنید. نتیجه‌ی به‌دست آمده (شامل دقت طبقه‌بند و f1 score برای هر کلاس) ذکر و راجع‌به تفاوت نتیجه‌ی دو روش بحث کنید. ممکن است برای گرفتن نتیجه‌ی مطلوب بر روی این مجموعه داده، اندکی پیش‌پردازش (نرمال‌سازی، تبدیل whitening و ...) روی آن انجام دهید. در صورت انجام، تمامی این کارها را در گزارش ذکر کنید.

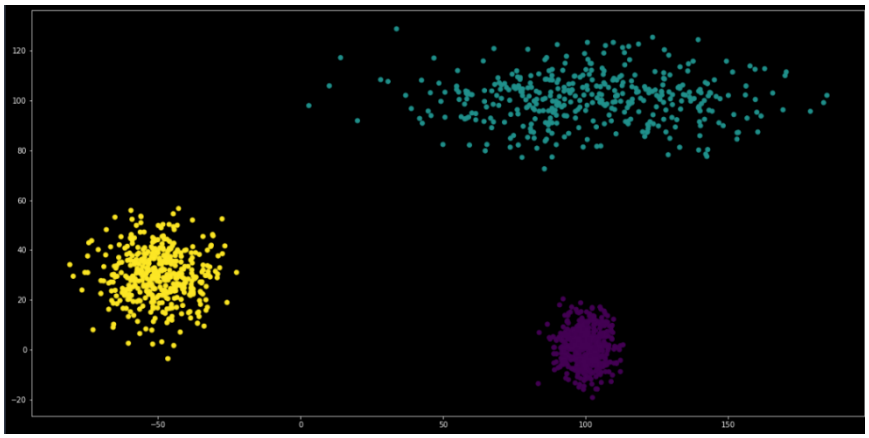
ج) الگوریتم‌های بالا را برای مجموعه داده‌ی noisy moons نیز تست کرده و نتیجه را ذکر کنید. (در فایل noisy moos ستون آخر نشان‌دهنده‌ی برچسب هر کلاس است) توجه کنید که قبل این کار بایستی این داده‌ها را به مجموعه‌های آموزش و آزمون تقسیم کنید.

د) از پکیج‌های آماده‌ی پایتون استفاده کرده و دو مجموعه داده‌ی بالا را توسط الگوریتم naïve bayes در این پکیج طبقه‌بندی کرده و نتیجه را گزارش دهید.

۱۰ سوال دهم (شبیه‌سازی، ۱۵ نمره)

در این سوال حق استفاده از پکیج‌های آماده‌ی یادگیری ماشین را ندارید.

یکی از روش‌های اتخاذشده برای جداسازی داده‌هایی که شامل بیش از یک کلاس هستند، روش one vs rest است. توضیح مختصری راجع‌به این روش بدهید. سپس در الگوریتم logistic regression از آن برای جداسازی مجموعه‌داده‌ای که در فایل random_dataset.csv آورده شده است استفاده کنید. دقت کنید که در این فایل، ستون اول ویژگی اول، ستون دوم ویژگی دوم و ستون سوم برچسب داده‌هاست. این مجموعه داده پراکندگی‌ای به شکل زیر دارد.



نوع خطای استفاده شده را در گزارش ذکر و بازای جداسازی هر کلاس از سایر کلاس‌ها، نمودار نزولی خطا بر حسب تعداد دفعات اجرای الگوریتم آموزش را رسم کنید و در انتها، خطوطی که کلاس‌ها را از هم جدا می‌کنند را توسط پارامترهای آموزش‌داده‌شده نمایش دهید.

۱۱ سوال یازدهم (شبیه‌سازی، ۱۵ نمره)

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

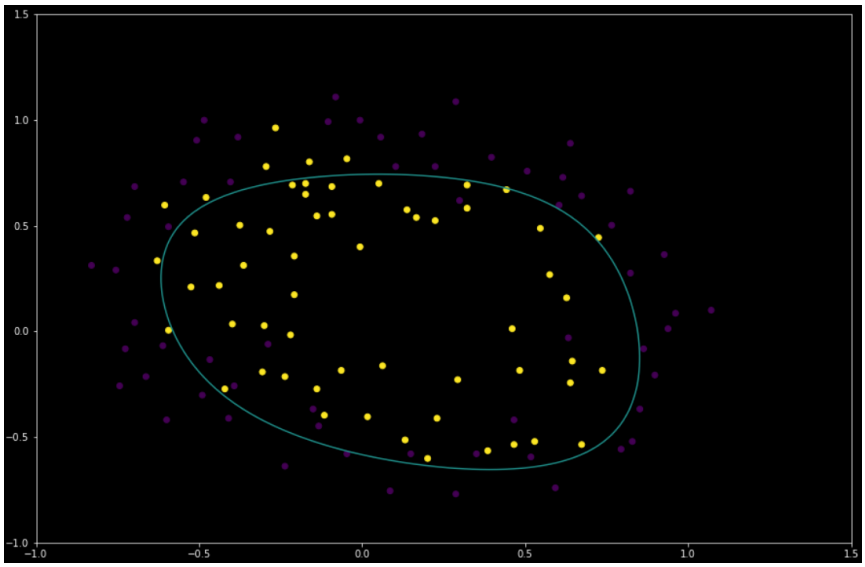
در این سوال بر روی مجموعه داده‌ی quality_test.csv کار خواهید کرد که در آن دو ستون اول نتایج تست یک چیپ و ستون سوم نشان‌دهنده‌ی قبول یا رد کیفیت آن چیپ است.

با استفاده از الگوریتم logistic regression و استفاده از l2 regularization دو کلاس این مجموعه داده را جدا خواهید کرد. همان‌طور که در شکل ۲ معلوم است، این مجموعه داده به‌صورت خطی جداپذیر نیست. بنابراین بایستی ابتدا فضای ویژگی‌ها را به مرتبه‌ی بالاتر برد. تابعی که برای این کار پیاده‌سازی خواهید کرد، عملیات زیر را انجام خواهد داد.

$$X = [x_1 \ x_2]^T \ , \qquad f(X) = [x_1 \ \ x_2 \ \ x_1^2 \ \ x_1x_2 \ \ x_2^2 \ \ x_1^3 \ \ x_1^2x_2 \ \ x_1x_2^2 \ \ x_2^3 \ \ \dots \ \ x_1x_2^5 \ \ x_2^6]$$

$$f(X) \ : \mathbb{R}^2 \to \mathbb{R}^{27}$$

در انتها دقت طبقه‌بند خود را بر روی همین داده‌ها گزارش کرده و مرز تصمیم‌گیری به‌دست آمده توسط الگوریتم خود را رسم کنید. نمودار شما بایستی چیزی شبیه به شکل ۲ باشد.



شکل ۲ ـ طبقه‌بند لاجیستیک با ضریب reg برابر ۱