گزارش تمرین دوم یادگیری ماشین

Fall 2020 - Machine Learning

چکیده

در این گزارش به بررسی سوالات تمرین دوم خواهیم پرداخت.

۱

سوال ۱

در این سوال فرض بر این است که با توضیح گاوسی مواجه هستیم. ابتدا ممان‌های مربوط به هر دسته را تخمین می‌زنیم. برای اینکار جمع برداری تمام نمونه‌های مربوط به یک کلاس را تقسیم بر تعداد می‌نماییم.

$$\hat{\mu}_{cross} = \frac{[1,1]^T + [1,2]^T + [2,1]^T + [2,2]^T + [2,3]^T + [4,3]^T}{6} = [2,2]^T$$
$$\hat{\mu}_{circle} = \frac{[-1,-1]^T + [-1,-2]^T + [-2,-1]^T + [-3,1]^T + [-3,-2]^T}{5} = [-2,-1]^T$$

حال برای تخمین ماتریس کوواریانس هر کلاس از رابطه‌ی زیر استفاده می‌نماییم:

$$\hat{\Sigma} = \frac{1}{N-1} X^T X$$

که در رابطه فوق مقصود از N همان تعداد داده‌های هر کلاس است و ماتریس X ماتریس متشکل از داده‌های شیفت داده شده حول میانگین می‌باشد، به عبارت دیگر سطر i ام ماتریس X به فرم زیر می‌باشد:

$$X_i = x_i - \hat{\mu}$$

که x_i بیانگر دیتای کلاس است. همچنین توجه بفرمایید که استفاده از $N-1$ به جای N برای حذف بایاس در تخمین کوواریانس [1] می‌باشد. با استفاده از روابط مذکور ماتریس‌های کواریانس دو کلاس به شرح زیر

[1] Unbiased Estimation

خواهد بود:

$$\hat{\Sigma}_{Cross} = \begin{pmatrix} 1.2 & 0.6 \\ 0.6 & 0.8 \end{pmatrix}$$
$$\hat{\Sigma}_{Circle} = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1.5 \end{pmatrix}$$

حال می‌دانیم معادله تابع افتراق‌ساز [۲] به صورت زیر است:

$$\begin{cases} g_i(x) = x^T W_i + w_i^T x + w_{i0}; \\ W_i = \frac{-1}{2}\Sigma_i^{-1} \\ w_i = \Sigma_i^{-1}\mu_i \\ w_{i0} = -0.5\mu_i^T\Sigma_i^{-1}\mu_i - 0.5\ln|\Sigma_i| + \ln P(\omega_i) \end{cases}$$

بنابراین برای دو کلاس مذکور داریم:

$$\begin{cases} g_{cross}(x) = -0.67x_1^2 - x_2^2 + x_1x_2 + 0.67x_1 + 2x_2 - 1.866; \\ g_{circle}(x) = -0.6x_1^2 - 0.4x_2^2 - 0.4x_1x_2 - 2.8x_1 - 1.6x_2 - 3.257 \end{cases}$$

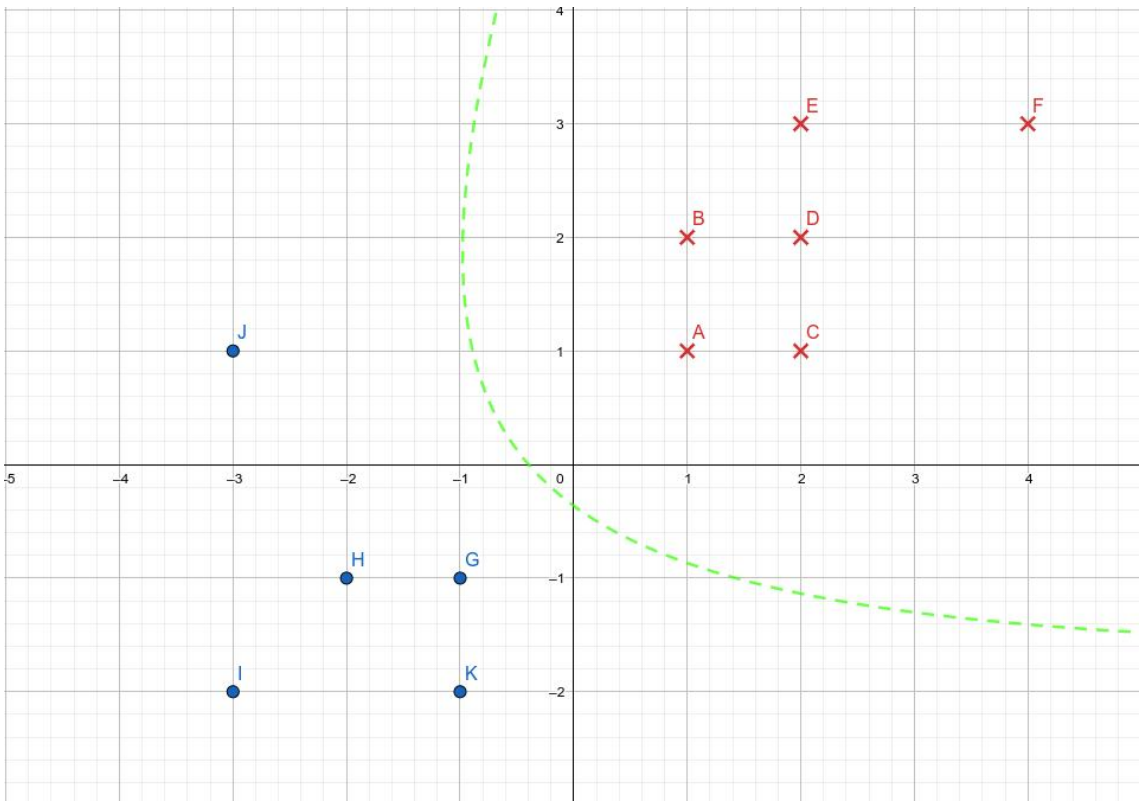
حال مرز بین دو کلاس به صورت زیر خواهد بود:

$$\text{Boundary of Decision} = \{x | f(x) = g_{cross}(x) - g_{circle}(x) = 0\}$$

$$\rightarrow \text{Boundary of Decision} = \{x | f(x) : -0.07x_1^2 - 0.6x_2^2 + 1.4x_1x_2 + 3.47x_1 + 3.6x_2 + 1.391 = 0\}$$

بنابراین مرز تصمیم‌گیری توسط منحنی $f(x) = 0$ تعیین می‌شود و به ازای x هایی که $f(x) > 0$ است، داده متعلق به کلاس ضربدر است و در غیر اینصورت داده متعلق به کلاس دایره می‌باشد. منحنی $f(x) = 0$ را می‌توانید در شکل زیر ملاحظه بفرمایید:

[2] Discriminant Function



شکل ۱.۱: طبقه‌بند بیز(منحنی سبزرنگ) و داده‌های کلاس‌های ضربدر و دایره

۲

سوال ۲

۱.۲ الف

می‌دانیم که احتمال یک نقطه در یک توزیع پیوسته برابر با صفر می‌باشد. به هر صورت به نظر می‌رسد در این سوال مقصود محاسبه چگالی احتمال در نقطه‌ی مذکور است. بدین منظور با استفاده از رابطه‌ی کلی تابع چگالی گوسی داریم:

$$P([2,-0.5,3]^T) = \frac{1}{\sqrt{(2\pi)^3}\sqrt{36}} exp^{(-0.5\times[1,-1.5,2]\begin{pmatrix}1&0&0\\0&8&2\\0&2&5\end{pmatrix}^{-1}[1,-1.5,2]^T)} = 0.003$$

۲.۲ ب

ابتدا بردارهای ویژه‌ی ماتریس Σ را محاسبه کرده و هر بردار ویژه را در ستون این ماتریس قرار می‌دهیم:

$$\Phi = \begin{pmatrix}\text{eig۱} & \text{eig۲} & \text{eig۳}\end{pmatrix} = \begin{pmatrix}0 & 0 & 1\\0.447 & -0.894 & 0\\-0.894 & -0.447 & 0\end{pmatrix}$$

همچنین ماتریس قطری Λ را می‌سازیم، به طوریکه درایه‌ی iام قطر اصلی، جذر مقدار ویژه متناسب با بردار ویژه i ام ماتریس Σ (به عبارت دیگر، ستون i ام ماتریس Φ) می‌باشد:

$$\Lambda^{\frac{1}{2}} = \begin{pmatrix}2&0&0\\0&3&0\\0&0&1\end{pmatrix}$$

حال ماتریس معادل تبدیل سفید ساز A_W [1] ، به صورت زیر محاسبه می‌گردد:

$$A_W = \Phi\Lambda^{\frac{-1}{2}} = \begin{pmatrix} 0 & 0 & 1 \\ 0.223607 & -0.298 & 0 \\ -0.447 & -0.149 & 0 \end{pmatrix}$$

۳.۲ ج

برای تبدیل توزیع فعلی به توزیع مطلوب از تبدیل زیر استفاده می‌نماییم:

$$y = T(x) = A_W^T x - A_W^T \mu = \begin{pmatrix} 0 & 0 & 1 \\ 0.223607 & -0.298 & 0 \\ -0.447 & -0.149 & 0 \end{pmatrix} ([2, 0.5, 3]^T - [1, 2, 1]^T)$$
$$\rightarrow y = [-1.23, 0.15, 1]^T$$

۴.۲ د

ابتدا فاصله‌ی ماهالانوبیس [2] x از μ را به صورت زیر محاسبه می‌نماییم:

$$\begin{cases} d^2_{\text{mahal}} = (x-\mu)^T \Sigma^{-1} (x-\mu) = 2.534 \\ \\ d^2_{\text{euqlidean}} = ||y||^2 = 2.534 \end{cases}$$

۵.۲ ه

نقطه‌ی دلخواه x را از فضای اول در نظر می‌گیریم. مطابق قسمت (ج) می‌توانیم برای نقطه‌ی y (نقطه‌ی تبدیل شده تخت تبدیل سفید سازی بنویسیم):

$$||y||^2 = ||A_W^T(x-\mu)||^2 = (A_W^T(x-\mu))^T(A_W^T(x-\mu)) = (x-\mu)^T A_W A_W^T (x-\mu) \quad (۱.۲)$$

[1] Whitening Transformation

[2] Mahalanobis

از طرفی داریم:

$$A_W^T \Sigma A_W = I \rightarrow \Sigma = (A_W^T)^{-1} A_W^{-1} \rightarrow \Sigma^{-1} = A_W A_W^T \tag{۲.۲}$$

حال با ترکیب دو رابطه ۱.۲ و ۲.۲ داریم:

$$||y||^2 = (x-\mu)^T \Sigma^{-1} (x-\mu)$$

۳

سوال ۳

۱.۳ الف

برای آنکه کمترین ریسک را در انتخاب کلاس ω_i داشته باشیم، باید ریسک انتخاب این کلاس از انتخاب سایر کلاس‌ها کمتر باشد

$$\begin{aligned} &\forall j \in \{1, ..., c\}; j \neq i : R(\alpha_i|X) \leq R(\alpha_j|X) \\ &\rightarrow \Sigma_{m=1}^{c} \lambda_{m,i} P(\omega_m|X) \leq \Sigma_{m=1}^{c} \lambda_{m,j} P(\omega_m|X) \\ &\rightarrow \Sigma_{m=1,m\neq i}^{c} \lambda_s P(\omega_m|X) \leq \Sigma_{m=1,m\neq j}^{c} \lambda_s P(\omega_m|X) \\ &\rightarrow \lambda_s P(\omega_j|X) \leq \lambda_s P(\omega_i|X) \rightarrow P(\omega_i|X) \geq P(\omega_j|X) \end{aligned} \tag{۱.۳}$$

همچنین ریسک انتخاب این کلاس باید از هزینه رد کردن کمتر باشد:

$$\begin{aligned} &R(\alpha_i|X) < \lambda_r : R(\alpha_i|X) = \Sigma_{m=1,m\neq i}^{c} \lambda_s P(\omega_m|X) = \\ &\lambda_s \Sigma_{m=1,m\neq i}^{c} P(\omega_m|X) = \lambda_s (1 - P(\omega_i|X)) \leq \lambda_r \\ &\rightarrow P(\omega_i|X) \geq 1 - \frac{\lambda_r}{\lambda_s} \end{aligned} \tag{۲.۳}$$

۲.۳ ب

برای نشان دادن حکم مذکور، باید ثابت کنیم که انتخاب بر اساس بیشینه تابع افتراق ساز هم ارز با حکم (الف) می‌باشد. در واقع فرض می‌کنیم کلاس w_i برآورد کننده مسئله $\arg\max_k g_k(x)$ باشد. در این صورت داریم:

$$i = \arg\max_j g_j(X) \rightarrow \forall j \in \{1, ..., c+1\} : g_i(X) \geq g_j(X)$$

$$if j \in \{1, ..., c\} : ۰۱$$

$$
\begin{aligned}
& P(X|\omega_i)P(\omega_i) \geq P(X|\omega_j)P(\omega_j) \\
\xleftrightarrow{P(X|\omega)P(\omega)=P(\omega|X)P(X)} \quad & P(\omega_i|X)P(X) \geq P(\omega_j|X)P(X) \qquad (۳.۳) \\
\equiv P(\omega_i|X) \geq P(\omega_j|X); \qquad & \forall j \in \{1, ..., c\}
\end{aligned}
$$

$$if j = c+1 : \quad .۲$$

$$
\begin{aligned}
& P(X|\omega_i)P(\omega_i) \geq \frac{\lambda_s - \lambda_r}{\lambda_s} \Sigma_{j=1}^{c} P(X|\omega_j)P(\omega_j) \\
\xleftrightarrow{P(X|\omega)P(\omega)=P(\omega|X)P(X)} \quad & P(\omega_i|X)P(X) \geq \frac{\lambda_s - \lambda_r}{\lambda_s} \Sigma_{j=1}^{c} P(\omega_j|X)P(X) \qquad (۴.۳) \\
\equiv P(\omega_i|X) \geq \frac{\lambda_s - \lambda_r}{\lambda_s} \Sigma_{j=1}^{c} P(\omega_j|X) = & \frac{\lambda_s - \lambda_r}{\lambda_s} \times 1 \\
\equiv P(\omega_i|X) \geq 1 - & \frac{\lambda_r}{\lambda_s}
\end{aligned}
$$

مشاهده می‌نماییم که عبارات بدست آمده در معادلات ۳.۳ و ۴.۳ همان معادلات ۱.۳ و ۲.۳ می‌باشند و از آنجا که تمامی مراحل استدلال در معادلات ۳.۳ و ۴.۳ دو طرفه می‌باشد، متوجه هم‌ارزی مسئله تعریف شده در بخش (ب) و (الف) می‌شویم.

۳.۳ ج

در نظر بگیرید: $\gamma = \frac{\lambda_r}{\lambda_s}$ و همچنین از بخش (الف) می‌دانیم یکی از شروط انتخاب کلاس ω_i شرط زیر است:

$$P(\omega_i|X) \geq 1 - \gamma. \qquad (۵.۳)$$

حال اگر ضریب γ به سمت 0 میل کند، طبق معادله‌ی ۵.۳ داریم $P(\omega_i|X) \geq 1 - \gamma = 1$، که با توجه به مفهوم احتمال($P(\omega_i|X) \leq 1$) مشاهده می‌کنیم که این ناتساوی تقریبا هیچ‌گاه برقرار نمی‌باشد(مگر اینکه مسئله تک کلاسه باشد که در این صورت طبقه بندی اندکی بی‌معناست) و در واقع همواره ترجیح می‌دهیم نمونه را رد کنیم. این نتیجه منتظره است، چرا که لازمه‌ی صفر شدن γ آن است که هزینه‌ی رد کردن برابر صفر باشد ($\lambda_s = 0$). بنابراین وقتی رد کردن یک نمونه هیچ هزینه‌ای برای ما ندارد، دقیقا معادل آن است

که کلاس درست را انتخاب نماییم($\lambda_{i,i} = 0$) و ضمنا به دلیل آنکه رد کردن نمونه هیچ احتمال وشبهه‌ای ندارد(عمل رد کردن قطعی است)، بنابراین همواره رد کردن بهترین استراتژی است.

از طرفی، اگر ضریب γ به سمت 1 میل کند، طبق معادله‌ی ۵.۳ داریم $P(\omega_i|X) \geq 1 - \gamma = 0$، و از آنجا که احتمال همیشه غیرمنفی است، مشاهده می‌کنیم تساوی فوق همواره برقرار است و در واقع همواره ترجیح می‌دهیم از رد کردن نمونه بپردازیم و حتی اگر شده، خطای زیاد را تحمل کرده، اما تصمیمی را اتخاذ نماییم. این نتیجه منتظره است، زیرا لازمه‌ی یک شدن ضریب گاما آن است که، ریسک رد کردن برابر با مجموع تمامی اقدامات باشد، بنابراین همواره ریسک رد کردن از ریسک انتخاب یک عمل بیشتر است و انتخاب یک عمل به رد کردن نمونه ارجح است.

به طور کلی در صورت که ضریب گاما به سمت صفر میل کند، بدین معناست که در طبقه بندی رویکرد محافظه کارانه‌ای داریم و در صورتی که چندان از صحت طبقه بندی نمونه مطمئن نیستیم، ترجیح می‌دهیم آنرا رد کنیم(برای مثال این موضع در تشخیص بیماری قابل درک است)، اما در صورتی که ضریب گاما به سمت ۱ میل کند، بدین معناست که ریسک اشتباه کردن چندان بالا نیست و ما ترجیح می‌دهیم حتما یک انتخاب داشته باشیم و تا حد امکان از رد کردن نمونه بپرهیزیم.

۴

سوال ۴

۱.۴ فاصله نقطه از خط

ابتدا مسئله را در قالب بهینه سازی بازنویسی می‌نماییم:

$$\begin{cases} d^2 = \min_x f(x) := ||x - x_0||^2 \\ h(x) := w^T x + b = 0 \end{cases} \tag{۱.۴}$$

حال با استفاده از روش ضرایب لاگرانژ [۱] مسئله را به فرم زیر بازنویسی می‌نماییم:

$$g(x, \lambda) = ||x - x_0||^2 + \lambda(h(x)) \tag{۲.۴}$$

حال به محاسبه نقاط بحرانی تابع $g(.)$ می‌پردازیم:

$$\begin{cases} \frac{\partial g}{\partial x} = 0 \to (x - x_0) + \lambda w = 0 \to x = -\lambda w + x_0 \\ \frac{\partial g}{\partial \lambda} = 0 \to w^T x + b = 0 \xrightarrow{x=-\lambda w+x_0} \lambda = \frac{w^T x_0 + b}{w^T w} \end{cases} \tag{۳.۴}$$

$$\to x = -\frac{w^T x_0 + b}{w^T w} w + x_0$$

[1] Lagrange multiplier

طبق قضیه‌ی لاگرانژ، می‌دانیم برای آنکه x محاسبه شده در معادله ۳.۴ مقدار تابع هزینه را کمینه نماید، اولا تابع در نقطه محاسبه شده باید کمینه باشد، ثانیا شرط کافی مرتبه دو(مثبت معین بودن مشتق دوم) برقرار باشد. بدین منظور داریم:

$$\begin{cases} \nabla h(x) = w^T \\ \frac{\partial^2 g}{\partial x^2} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} > 0 \end{cases} \tag{۴.۴}$$

بنابراین با توجه به معادله ۴.۴ کافی است $w \neq 0$ تا به ازای x تابع $f(.)$ در مسئله بهینه سازی کمینه باشد. بنابراین در صورت برقراری فرض فاصله نقطه از خط برابر است با:

$$d = f(x^*) = || - \frac{w^T x_0 + b}{w^T w} w + x_0 - x_0|| = ||\frac{w^T x_0 + b}{w^T w} w||$$

۲.۴ فاصله نقطه از بیضی گون

در نظر بگیرید که:

- با توجه به بیضی‌گون بودن نتیجه می‌گیریم که A مثبت معین است.
- هر ماتریس مثبت معین با بعد محدود، پایه یکامتعامد از بردارهای ویژه دارد

بنابراین مطابق موارد فوق می‌توان ادعا نمود که همواره می‌توان مسئله را به فضای مختصات جدید برد که در این فضا اندازه‌ها ثابت است و متر حفظ می‌شود و ماتریس A در این فضا قطری است (زیرا همانطور که گفتیم پایه یکا متعامد از بردارهای ویژه موجود است که با استفاده از این پایه ماتریس را می‌توان به فضای قطری برد و از طرفی چون پایه یکامتعامد است پس معکوس ماتریس انتقال عملا خود ترنسپوز ماتریس می‌شود و خلاصه آنکه این تبدیل نرم را حفظ می‌نماید). بنابراین بدون از دست رفتن کلیت مسئله را برای A قطری با درایه‌های قطر اصلی γ_i برای عنصر i ام قطر بررسی می‌نماییم (زیرا در هر صورت می‌توان A را قطری کرد). حال مجددا مانند قسمت قبل مسئله را در چهارچوب بهینه سازی حل می‌نماییم:

$$\begin{cases} d^2 = \min_x f(x) := ||x - x_0||^2 \\ h(x) := x^T A x - 1 = 0 \end{cases} \tag{۵.۴}$$

حال با استفاده از روش ضرایب لاگرانژ مسئله را به فرم زیر بازنویسی می‌نماییم:

$$g(x,\lambda) = ||x - x_0||^2 + \lambda(h(x)) \tag{۶.۴}$$

حال به محاسبه نقاط بحرانی تابع $g(.)$ می‌پردازیم:

$$\begin{cases} \frac{\partial g}{\partial x} = 0 \rightarrow x - x_0 + \lambda A x = 0 \rightarrow x = (\lambda A + I)^{-1} x_0 \\ \frac{\partial g}{\partial \lambda} = 0 \rightarrow x^T A x - 1 = 0 \end{cases} \tag{۷.۴}$$

از طرفی با توجه به قطری بودن A (به توضیحات مقدمه همین بخش مراجعه شود) داریم:

$$\begin{cases} x = 0 \rightarrow x = (\lambda A + I)^{-1} x_0 \rightarrow x^i = \frac{x_0^i}{1+\lambda\gamma_i} \\ x^T A x = \Sigma_i \gamma_i (x^i)^2 \end{cases} \tag{۸.۴}$$

که در معادله فوق مقصود از بالاگذار i همان درایه‌ی i ام بردار است. حال با ترکیب دو رابطه‌ی ۷.۴ و ۸.۴ داریم:

$$\Sigma_i \gamma_i \frac{\gamma_i (x_0^i)^2}{(1+\lambda\gamma_i)^2} = 1 \tag{۹.۴}$$

معادله فوق را می‌توان با استفاده از روش نیوتن به صورت عددی حل نمود و مقدار λ را محاسبه نمود(برای روش نیوتن و رافسون رجوع شود به گزارش تمرین اول اینجانب).

طبق قضیه‌ی لاگرانژ، می‌دانیم برای آنکه x محاسبه شده در معادله ۹.۴ مقدار تابع هزینه را کمینه نماید، اولا تابع در نقطه محاسبه شده باید کمینه باشد، ثانیا شرط کافی مرتبه دو(مثبت معین بودن مشتق دوم) برقرار باشد. بدین منظور داریم:

$$\begin{cases} \nabla h(x) = Ax \neq 0 \\ \frac{\partial^2 g}{\partial x^2} = \lambda A + I > 0 \end{cases} \tag{۱۰.۴}$$

همانطور که ملاحظه می‌فرمایید هرچند که با توجه به مثبت معین بودن ماتریس A هموار شرط اول معادله ۱۰.۴ برقرار است اما بسته به A و همینطور λ محاسبه شده در معادلات ۹.۴ و ۸.۴ عملا ممکن است شرط

دوم برقرار نباشد و عملا نتوان از قضیه استفاده کرد. بنابراین، **به نظر حل عددی به صورت کلی نمی‌تواند قابل اعتماد باشد** و باید نتایج بدست آمده به دقت بررسی گردند و سپس در صورت اقناع شرایط قضیه مورد استفاده قرار بگیرند. در نهایت فاصله بیضی‌گون، فاصله x محاسبه شده تا نقطه‌ی x_0 می‌باشد.

۳.۴ حل مثال‌ها

- مثال اول:

در این مثال بیضی‌گون و نقطه مورد نظر به صورت زیر انتخاب شده‌اند:

$$x_0 = [1,1]^T; \qquad A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

با توجه به قطری بودن ماتریس بیضی‌گون مستقیم از معادلات ۸.۴ ۹.۴ استفاده می‌نماییم:

$$\frac{2}{(2+\lambda)^2} = 1 \rightarrow \lambda = -1+\sqrt{2}, -1-\sqrt{2}$$

مطابق ۱۰.۴ تنها مقدار $\lambda = -1+\sqrt{2}$ داریم:

$$x = \frac{-x_0}{2} = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}] \rightarrow d = ||x - x_0|| = 0.614$$

- مثال دوم:

این مثال بیضی‌گون نیست(ماتریس مثبت معین نیست)، پس تمام محاسبات لاگرانژ را از ابتدا انجام می‌دهیم $A = \begin{pmatrix} 2 & 3 \\ 1 & 1 \end{pmatrix}$; $x_0 = [1,1]^T$ از آن‌جا که ماتریس A متقارن نیست معادله ۷.۴ را با کمی اصلاح بکار می‌گیریم. داریم:

$$\begin{cases} x = (\lambda(A+A^T)+I)^{-1}x_0 = [\frac{-2\lambda+1}{-8\lambda^2+6\lambda+1}, \ \frac{1}{-8\lambda^2+6\lambda+1}]^T \\ \xrightarrow{x^TAx=1} 2(\frac{-2\lambda+1}{-8\lambda^2+6\lambda+1})^2 + 4\frac{-2\lambda+1}{-8\lambda^2+6\lambda+1} \times \frac{1}{-8\lambda^2+6\lambda+1} + (\frac{1}{-8\lambda^2+6\lambda+1})^2 = 1 \end{cases}$$

از آن‌جا که معادله‌ی بالا یک معادله درجه ۴ است و از آنجا که ماتریس هسین برابر با $I+\lambda(A+A^T)$ می‌باشد، تنها جواب $\lambda = 0.2$ از بین ۴ جواب دیگر قابل قبول هستند. در این حالت داریم:

$$x^T = [x_1, x_2]^T = [0.28, 53] \rightarrow d = 0.85$$

همچنین شکل‌ها باید اضافه شوند و حل دوم باید اصلاح گردد

۵

سوال ۵

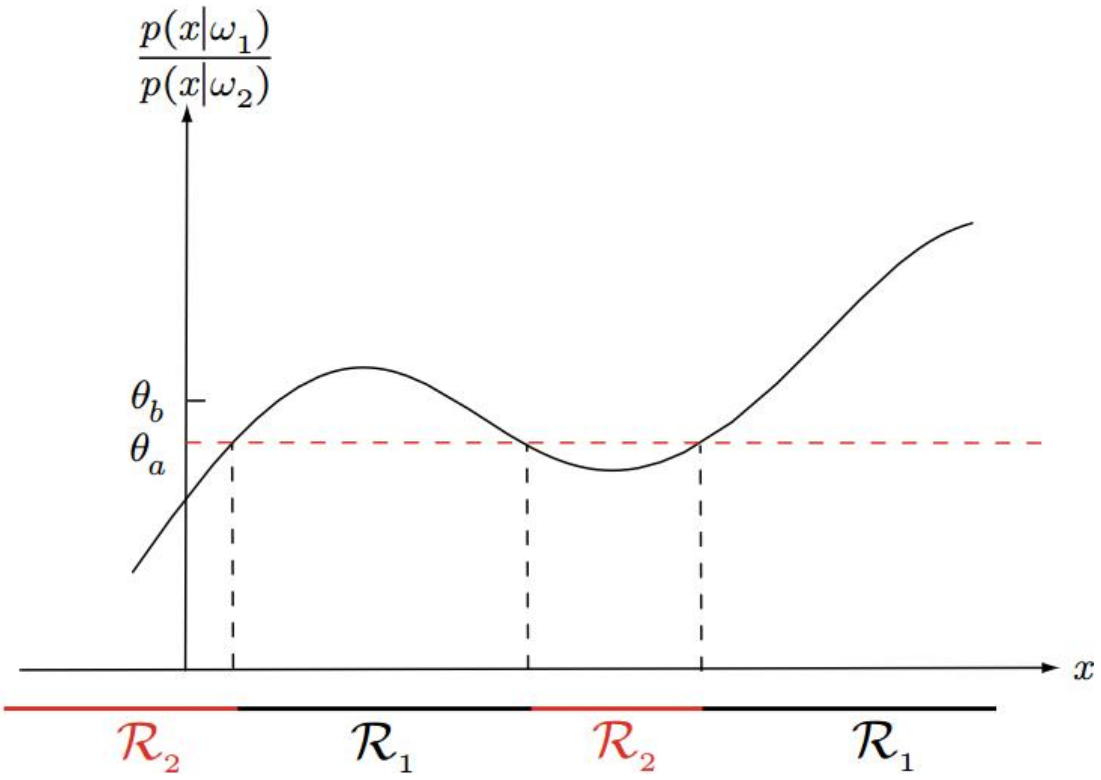
۱.۵ الف

می‌دانیم در صورت وجود هزینه‌های فوق و با در نظر گرفتن $\lambda_{11} = \lambda_{22} = 0$ تصمیم گیری به صورت زیر صورت خواهد پذیرفت:

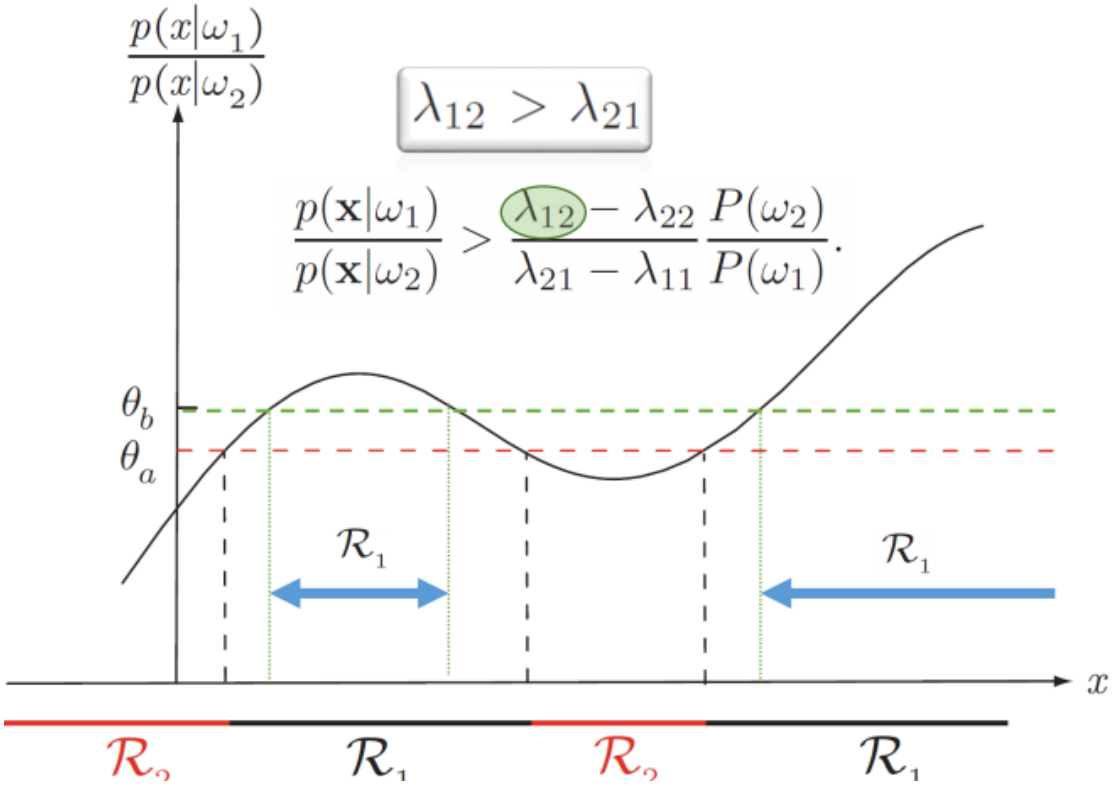
$$\begin{cases} \omega_1 & if \frac{P(X|\omega_1)}{P(X|\omega_2)} \geq \frac{\lambda_{12}}{\lambda_{21}} \frac{P(\omega_2)}{P(\omega_1)} \\ \omega_2 & \text{Othw} \end{cases} \tag{۱.۵}$$

بنابراین عبارت $\zeta = \frac{\lambda_{12}}{\lambda_{21}}$ ماتتد یک پیچ تنظیم عمل می‌نماید و سعی می‌کند تصمیم گیری را بر مبنای هزینه‌ها تنظیم کند. بدین صورت که هرچه مقدار ζ بیشتر باشد، ما در انتخاب تصمیم ω_1 محتاط‌تر می‌شویم و در مواقع کمتری این تصمیم را اتخاذ می‌نماییم و بالعکس. این موضوع بدیهی است، در واقع زمانی که ریسک تصمیم یک بیشتر از تصمیم دوم باشد، منطقی است که در قبال تصمیم اول محتاطانه‌تر عمل نماییم.

برای مثال به شکل ۱.۵ توجه بفرمایید. در ابتدا حد تصمیم‌گیری برابر با θ_a می‌باشد. اما با افزایش نسبت ζ همانطور که در شکل ۲.۵ مشاهده می‌شود که این حد به θ_b تغییر می‌یابد، که باعث کوچک تر شدن نواحی R_1 می‌شود.



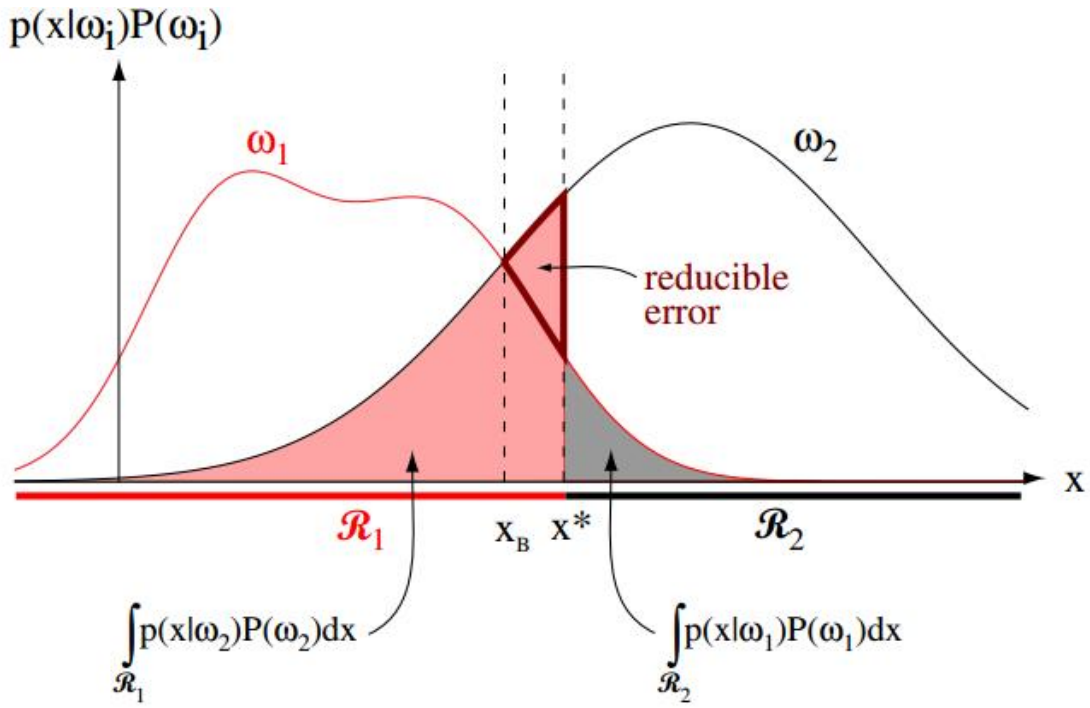
شکل ۱.۵: طبقه بند بر اساس ریسک اقدام



شکل ۲.۵: مقایسه طبقه بندها با ریسک‌های متفاوت

۲.۵ ب

خطای قابل کاهش [۱] به خطایی گفته می‌شود که در اثر اتخاذ استراتژی(تعیین حد تصمیم) غیربهینه بوجود می‌آید و خطای ذاتی طبقه بند نیست. این خطا با اتخاذ تصمیم بهینه قابل جبران است. همانطور که در شکل ۳.۵ مشاهده می‌شود، ناحیه مثلثی خطای قابل کاهش است که در اثر تصمیم غیربهینه(x^*) بوجود آمده است. در اثر این تصمیم، بخش زیادی از نمونه‌های موجود در کلاس ω_2 به عنوان نمونه‌ی کلاس ω_1 تشخیص داده شده‌اند، درحالیکه در این ناحیه کلاس ω_2 شانس بیشتری داشته است. با اتخاذ تصمیم بهینه X_B این خطا جبران می‌گردد.



شکل ۳.۵: خطای قابل کاهش

[1] Reducible Error

۳.۵ ج

تمایزپذیری [۲] یک معیار ذاتی برای داده‌های کلاس‌های مختلف است و بیانگر میزان جداپذیری و قابل تمایز بودن داده‌هاست. هرچقدر این متر و معیار بیشتر شود، به این معناست که طبقه بند با تمونه‌های متفاوتی مواجه است و برای طبقه‌بندی آن‌ها کار راحت‌تری پیش رو دارد. یک تعریف مناسب برای این معیار می‌تواند بر مبنای فاصله‌ی میانگین دسته‌ها و میزان پراکندگی آن‌ها باشد(در مسائل بیز). در واقع به طور شهودی می‌توان گفت که نمونه‌ها (در مسائل با توزیع گاوسی) حول میانگین دسته و به اندازه واریانس پراکنده می‌شوند. در صورتی که میانگین این دسته‌ها از یکدیگر دور باشند و از طرفی واریانس نیز کم باشد، مشاهده می‌گردد که نمونه‌های کلاس‌های متمایز در فواصل دور از هم و به صورت متمرکز حول میانگین خود قرار گرفته‌اند. رابطه‌ی ریاضی این معیار به شرح زیر می‌باشد:

$$d' = \frac{||\mu_2 - \mu_1||}{\sigma}$$

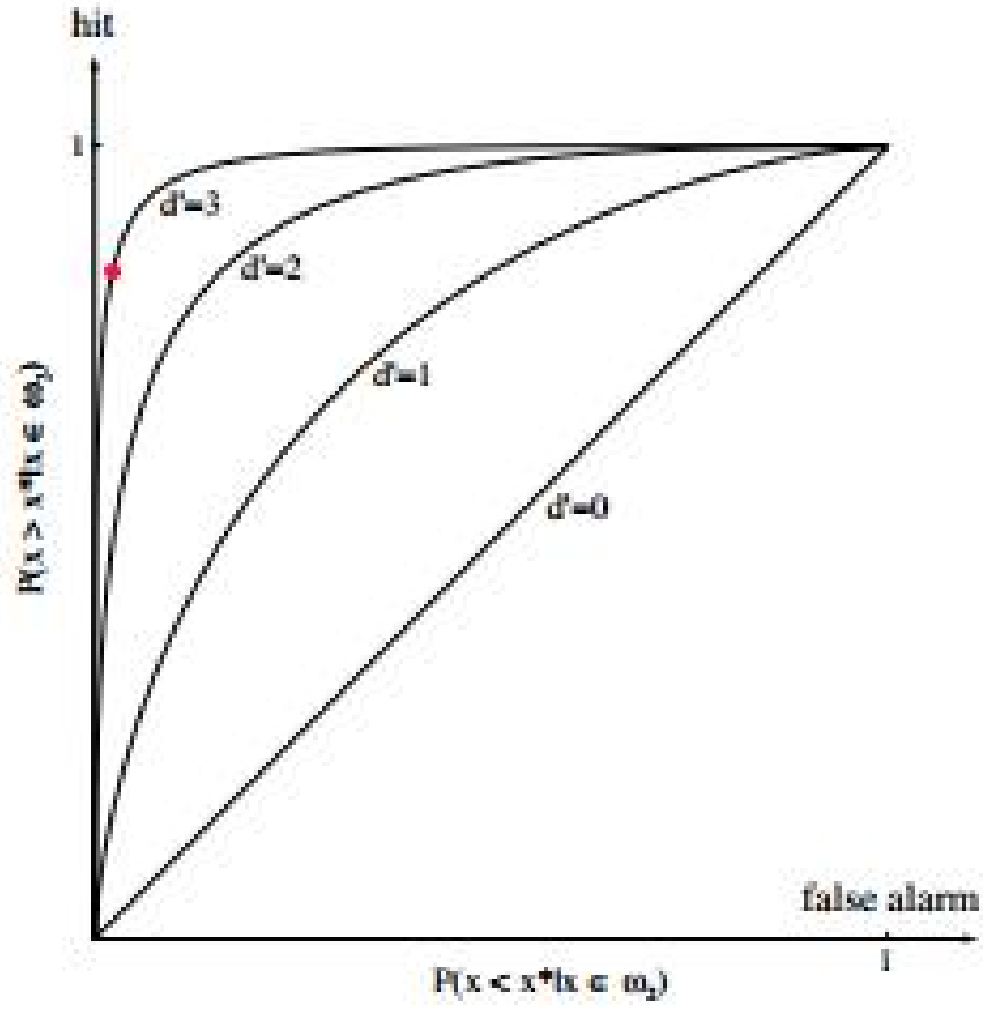
۴.۵ د

در مواقعی که پارامترهای توزیع دادها را نداریم عملا نمی‌توانیم از رابطه‌ی بخش گذشته استفاده نماییم. در این صورت، برای تعیین یک معیار مناسب برای تمایزپذیری از منحنی ROC استفاده می‌نماییم. در واقع در رسم این منحنی، محورها را به صورت hit و false alarm انتخاب می‌نماییم. سپس به ازای مرز تصمیم‌های متفاوت (X^*) به اندازه‌گیری میزان hit و false alarm می‌پردازیم، سپس به ازای هر X^* یک نقطه متناسب با این دو معیار در صفحه رسم می‌نماییم. وقتی این کار را به ازای تموم X^* ها انجام دهیم، در نهایت به نمودار ROC می‌رسیم. سطح بین این نمودار و ربع اول و سوم می‌تواند معیار مناسبی برای جداپذیری باشد، هرچه این سطح بیشتر باشد، درنهایت تمایز پذیری داده‌ها بیشتر خواهد بود.

این نمودار هرچند که همواره صعودی است(مانند cdf یک توزیع)، اما می‌تواند اکید نباشد. برای مثال در دو توزیع گاوسی که کاملا از هم متمایزند(تمایز پذیری بالایی دارند) در ابتدا به صورت بسیار سریع hit زیاد

[2] Discriminability

شده در حالیکه false alarm تقریبا ثابت است(نمونه کاملا صعودی)، اما سپس false alarm زیاد شده اما hit ثابت می‌ماند.



شکل ۴.۵: نمودار ROC بعضا صرفا صعودی است و نه اکیدا صعودی

۵.۵ ه

- Generative در این رویکرد سعی می‌شود تمام نمونه‌ها یه یک باره در نظر گرفته شود و سپس از روی تمامی نمونه‌ها توزیع واقعی هر کلاس مشخص شود. در واقع در این حالت سعی می‌گردد که احتمال توام $P(x, y)$ برآورد شود.
- Discriminative ذر این رویکرد هدف ساختن مرز تصمیم‌های کلاس است. در واقع سعی می‌شود که

توزیع هر کلاس در تعداد نمونه‌های مرتبط با خود بررسی شود. می‌توان ادعا نمود که در این حالت تلاش بر تعیین توزیع شرطی : $P(x|y)$ است.

- مقایسه و نتیجه‌گیری هرچند که در رویکرد اول به نظر عمل محاسبه احتمال کار ساده‌ای به نظر می‌رسد، اما چالش اصلی در جمع آوری نمونه‌ها می‌باشد. در واقع در این روش بسیار مهم است که نمونه برداری به صورت کاملا عادلانی و به دور از بایاس باشد، که این کار بسیار دشوار و در مواردی غیر ممکن است. اما روش دوم با آنکه محاسبه و استدلال آن شاید کمی قابل توجه باشد، اما در عوض در مقابل بایاس و نمونه برداری مقاوم است و سعی می‌کند در هنگام بررسی هر کلاس صرفا تمرکز خود را بر روی داده‌ها مربوط به آن کلاس قرار دهد.

۶

سوال ۶

۱.۶ مقدمه

در تمامی بخش‌های این سوال، بنابر اعلام دستیار آموزشی محترم $k = 1$ لحاظ خواهد شد.

۲.۶ الف

- d_1:

طبق این معیار داده‌های $[8, 3]^T, [8, -3]^T$ نزدیکترین داده‌ها به نقطه‌ی $[5, 0]^T$ می‌باشند که در این حالت فاصله آن‌ها از این نقطه برابر است با:

$$d_1(x, y) = \max(8 - 5, 3 - 0) = 3$$

همچنین از آنجا که این نقاط قرمز هستند پس **نقطه‌ی** $[5, 0]$ **متعلق به کلاس قرمز می‌باشد.**

- d_2:

طبق این معیار داده‌ی $[0, 0]$ نزدیکترین داده به نقطه‌ی $[5, 0]^T$ می‌باشند که در این حالت فاصله آن‌ها از این نقطه برابر است با:

$$d_1(x, y) = (5 - 0) + (0 - 0) = 5$$

همچنین از آنجا که این نقاط قرمز هستند پس **نقطه‌ی** $[5, 0]$ **متعلق به کلاس آبی می‌باشد.**

- d_1:

طبق این معیار، مجددا داده‌های $[8,3]^T, [8,-3]^T$ نزدیکترین داده‌ها به نقطه‌ی $[5,0]^T$ می‌باشند که در این حالت فاصله آن‌ها از این نقطه برابر است با:

$$d_1(x,y) = \sqrt{(8-5)^2 + (3-0)^2} = \sqrt{18}$$

همچنین از آنجا که این نقاط قرمز هستند پس **نقطه‌ی** $[5,0]$ **متعلق به کلاس قرمز می‌باشد.**

۳.۶ ب

با استفاده از رابطه‌ی $d^2_{\text{mahal}} = (x-\mu)^T\Sigma^{-1}(x-\mu)$ به محاسبه فواصل می‌پردازیم:

- $\mu_1 = [0,0]^T; \qquad A = \begin{pmatrix} 3 & -3 \\ -3 & 3.5 \end{pmatrix}$:

$$d^2_{\text{mahal}} = [1.5, 1.5] \begin{pmatrix} \frac{3.5}{1.5} & \frac{3}{1.5} \\ \frac{3}{1.5} & \frac{3}{1.5} \end{pmatrix} [1.5, 1.5]^T = 18.75$$

- $\mu_1 = [0,0]^T; \qquad A = \begin{pmatrix} 3 & 3 \\ 3 & 3.5 \end{pmatrix}$:

$$d^2_{\text{mahal}} = [1.5, 1.5] \begin{pmatrix} \frac{3.5}{1.5} & \frac{-3}{1.5} \\ \frac{-3}{1.5} & \frac{3}{1.5} \end{pmatrix} [1.5, 1.5]^T = 0.75$$

۴.۶ ج

می‌دانیم ماتریس کواریانس یک ماتریس مثبت نیمه معین [1] است. از طرفی به طور معادل می‌توان گفت تمامی مقادیر ویژه این ماتریس باید نامنفی باشند. از طرفی از آنجا که ماتریس مورد بحث یک ماتریس با ابعاد 2×2 است، تنها شامل دو بردار ویژه است. از طرفی از آنجا که اثر [2] ماتریس مورد بحث مثبت است بنابراین جمع دو مقدار ویژه نیز مثبت است ($\lambda_1 + \lambda_2 = \text{Trace} = 4 + 5 > 0$) بنابراین اگر ضرب دو مقدار ویژه نیز مثبت باشد، آنگاه به طور هم‌ارز می‌توان گفت که ماتریس مثبت معین است. از طرفی ضرب مقادیر ویژه برابر با دترمینان ماتریس است. بنابراین به طور هم ارز، برای حل مسئله تنها نا منفی بودن دترمینان را بررسی

[1] Positive Semi Definite

[2] Trace

می‌نماییم.

$$det(\Sigma) = 20 - \alpha^2 \geq 0 \equiv \alpha \in (-\sqrt{20}, +\sqrt{20})$$

۵.۶ د

برای حل این سوال از رابطه‌ی ۱.۵ استفاده می‌نماییم. مطابق این رابطه:

$$\begin{cases} \omega_1 & if \frac{x+\frac{1}{2}}{\frac{3x^2}{4}+\frac{3}{4}} \geq \frac{2-1}{3-1}\frac{\frac{1}{4}}{1-\frac{1}{4}} \\ \omega_2 & \text{Othw} \end{cases}$$

حال به طور خاص در مورد بازه بندی داریم:

$$\begin{aligned} \frac{x+\frac{1}{2}}{\frac{3x^2}{4}+\frac{3}{4}} \geq \frac{1}{2}\frac{\frac{1}{4}}{1-\frac{1}{4}} &\equiv x^2 - 8x - 3 \leq 0 \\ &\equiv x \in (4-\sqrt{19}, 4+\sqrt{19}) \end{aligned}$$

با توجه به محاسبات فوق و بازه‌ی داده شده در صورت سوال داریم:

$$\begin{cases} \omega_1 & if : x \in [0,1] \\ \omega_2 & \end{cases}$$

بنابراین مشاهده می‌شود بازه‌ی،

$x \in [0,1]$ متعلق به کلاس اول می باشد.

۷

سوال ۷

۱.۷ مقدمه

روش GLM یک فریمورک کلی برای حل مسائل رگرسیون و طبقه‌بندی، به کمک پارامترهایی است که روابط خطی دارند(هر چند که تابع نهایی در فریمورک ممکن است شکل غیرخطی داشته باشد). این چهارچوب به صورت زیر تعریف می‌گردد:

$$P(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

$$\begin{cases} T(y) : \text{Statistic Sufficient} \\ a(\eta) = \text{function partition Log} \end{cases}$$

در رابطه‌ی بالا، با انتخاب T, a, b عملا می‌توان به خانواده‌ای از روش‌های بر اساس پارامتر η دست یافت. برای مثال می‌توان دو زیر خانواده زیر را تعریف نمود:

۱. **Regression: Logistic** فرض کنید انتخاب‌های زیر برای پارامترهای فریمورک را داشته باشیم:

$$\begin{cases} T(y) = y \\ a(\eta) = -\log(1-\phi) \\ b(y) = 1 \end{cases}$$

در این صورت داریم:

$$p(y,\phi) = \phi^y (1-\phi)^{1-y}$$

که یک توزیع برنولی با پارامتر ϕ می‌باشد که با پیش‌برد طبقه‌بندی از این روش عملا به روش Logistic Regression می‌رسیم.

۲. **Regression Linear** فرض کنید اجزای فریمورک را به صورت زیر انتخاب نماییم:

$$\begin{cases} \eta = \mu T(y) = y \\ a(\eta) = \frac{\mu^2}{2} \\ b(y) = (\frac{1}{2\pi}) \exp(-\frac{y^2}{2}) \end{cases}$$

در این صورت مشاهده می‌شود که:

$$p(y, \mu) = \mathcal{N}(\mu, 1)$$

که این همان روش ML خطی است، که می‌دانیم معادل با رگرسیون خطی است.

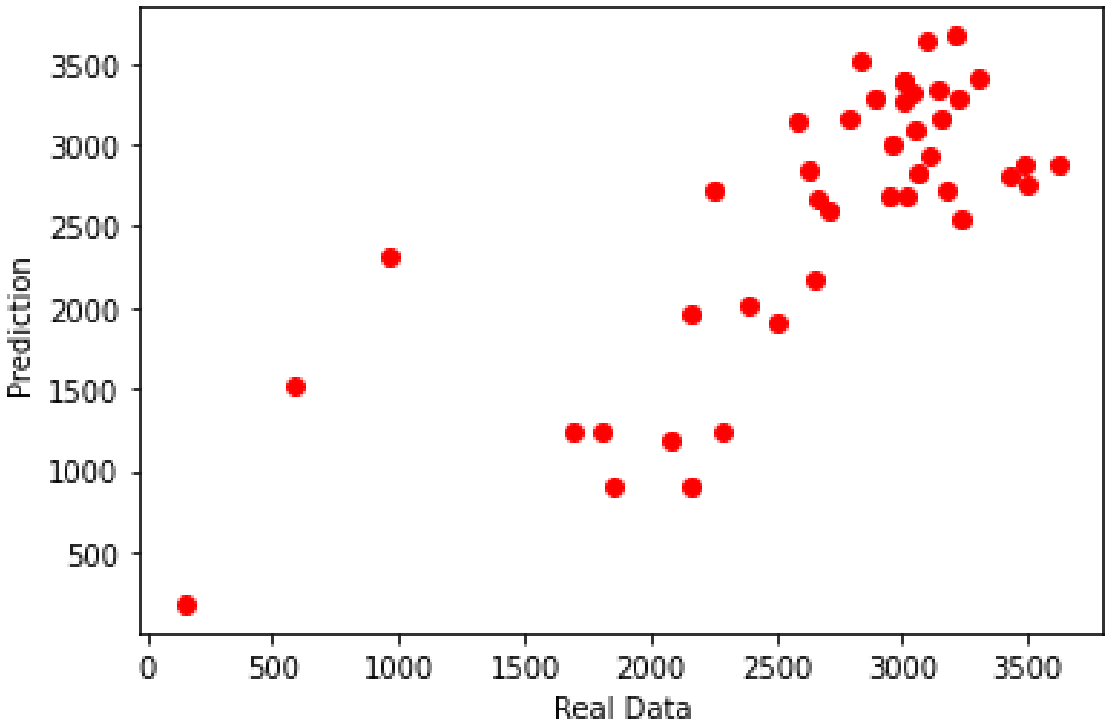
بنابراین مطابق آنچه که گفته شد به صورت کلی GLM سعی در گسترش رگرسیون خطی به صورت استفاده در دل توابع دیگر است: $\text{Estimation} = h(X, \theta)$ که پارامتر θ در این تابع به صورت $\theta^T X$ ظاهر می‌شود. همچنین این روش کماکان معیار نزدیکترین فاصله مجموع را سعی می‌نماید برآورده کند. در واقع با در نظر گرفتن عبارات رگولایز شده می‌توان قیدهای زیر را در نظر رگفت:

$$min \Sigma d(y, \hat{y}) + \lambda F(\theta)$$

که مقصود از $F(\theta)$ هر نوع نرم برای ماتریس وزن است و همچنین $\hat{y}$ نیز مقادیر تخمین زده شده می‌باشند.

۲.۷ شبیه سازی و نتایج

پس از اجرای الگوریتم مشاهده می‌شود که خروجی‌های پیش‌بینی شده به شرح شکل ۱.۷ می‌باشند. همانطور که در این شکل مشاهده می‌گردد، داده‌ها تقریبا بر روی نیمساز ربع اول و سوم هستند که این نشان دهنده نزدیکی تخمین می‌باشد.



شکل ۱.۷: نمودار ویژگی‌های مختلف دیتاست iris

۸

سوال ۸

۱.۸ الف

در شکل ۱.۸ نمودار تمام انتخاب‌های دوتایی از ویژگی‌ها را رسم می‌نماییم. بر مبنای این نمودارها می‌توان دریافت که در صورت استفاده از ویژگی‌های Petal Width و Petal Length می‌توان این سه کلاس را به صورت خطی از هم جدا نمود.

۲.۸ پیاده سازی کد و الگوریتم

کد این سوال در فایل P8.py ذخیره شده است.

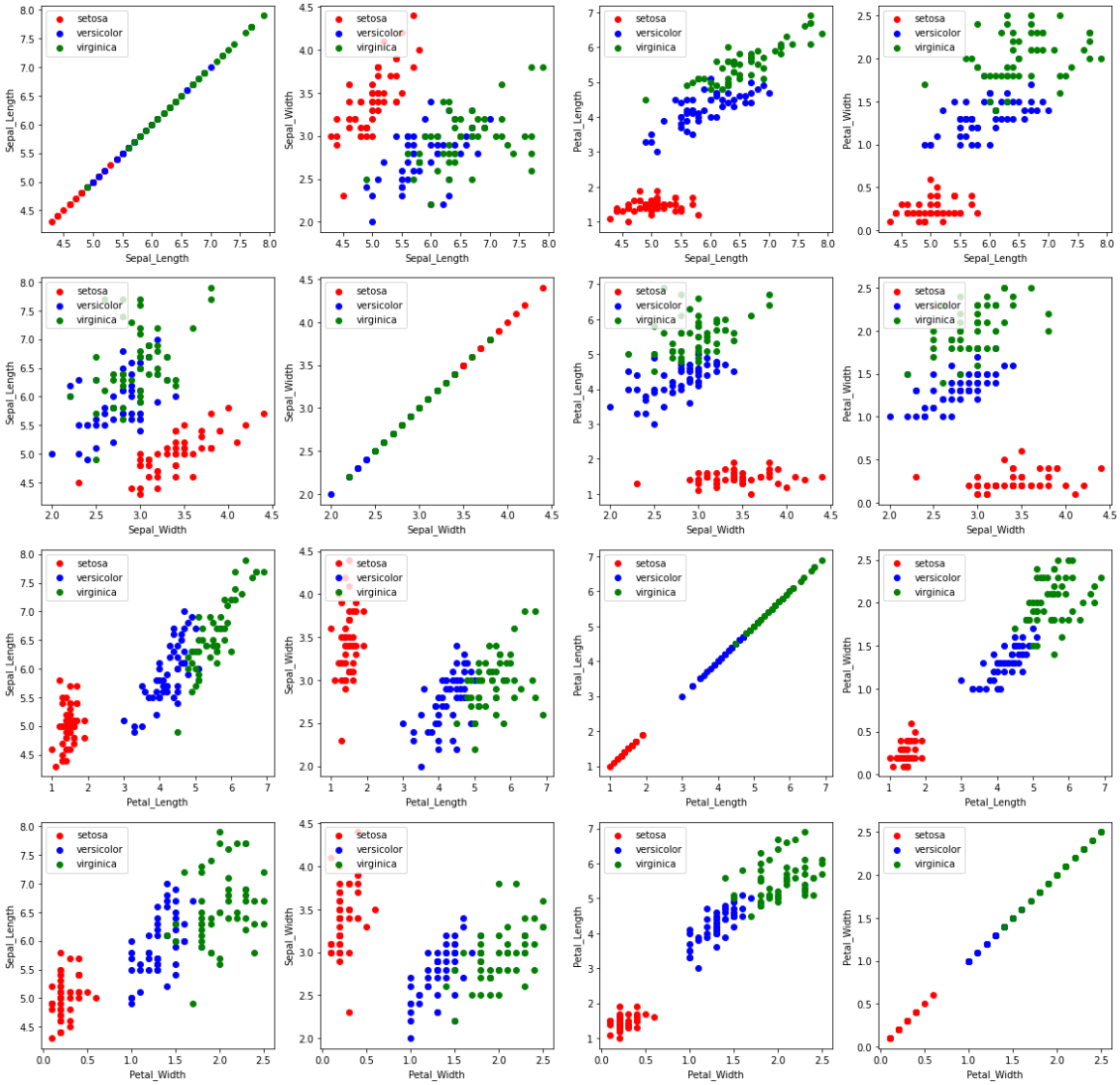
در این کد، توابعی به منظور جدا سازی رندوم دیتا و تقسیم آن‌ها به دادهای تست و آموزش طراحی شده است. همچنین توابع نرمالیزه کننده داده نیز پیاده سازی شده‌اند. در ادامه برای هر دیتای تست، نزدیکترین همسایه را بررسی کرده و بر اساس برچسب آن به تصمیم‌گیری می‌پردازیم.

۳.۸ ب

ذر این قسمت عملکرد طبقه بند را به ازای دو حالت بررسی می‌نماییم:

- **طبقه بندی بدون نرمالیزه** در این حالت مشاهده می‌شود که دقت مدل برابر با 87% می‌باشد.
- **طبقه بندی همراه با نرمالیزه کردن** در این حالت مشاهده می‌شود که دقت افت کرده و برابر با 84% می‌گردد.

- **تحلیل و نتیجه‌گیری** مشاهده می‌شود که این طبقه بند دارای دقت 87% می‌باشد که با نرمالیزه کردن این دقت کاهش یافته است. دلیل این امر آنست که نرمالیزه سعی در یکسان کردن ارزش ویژگی‌های مختلف دارد. این امر در مواردی که یک ویژگی در تمایز دادن دیتاها مهم‌تر از باقی است می‌تواند موجب ضرر شود. برای مثال در تشخیص گل‌ها شاید رنگ یک گل به مراتب مهم‌تر از پهنای درونی گل‌برگ باشد. بنابراین نرمالیزه کردن از آن‌جا که تمام ویژگی‌ها را هم ارزش می‌کند ممکن است موجب شود بعضا یک ویژگی کم‌اهمیت‌تر را بر ویژگی مهم مقدم بدانیم و دسته بندی را بر اساس آن انجام دهیم.



شکل ۱.۸: نمودار ویژگی‌های مختلف دیتاست iris

۴.۸ ج

مشاهده می‌شود ماتریس درهم‌ریختگی را می‌توانید در شکل ۱.۸ مشاهده بفرمایید. همچنین سایر مشخصات طبقه‌بند به شرح زیر است:

۱. **دقت طبقه بند:** منظور از دقت طبقه بند تعداد پیش‌بینی‌های درست به کل پیش‌بینی‌ها می‌باشد:

$$Acc = \frac{TP}{TP + TN + FP + FN} = \frac{26}{31} = 83\%$$

۲. **f1 score:** می‌دانیم به صورت کلی این معیار بر اساس رابطه‌ی زیر برای هر کلاس محاسبه می‌گردد:

$$f_1s = \frac{TP}{TP + 0.5(FP + FN)}$$

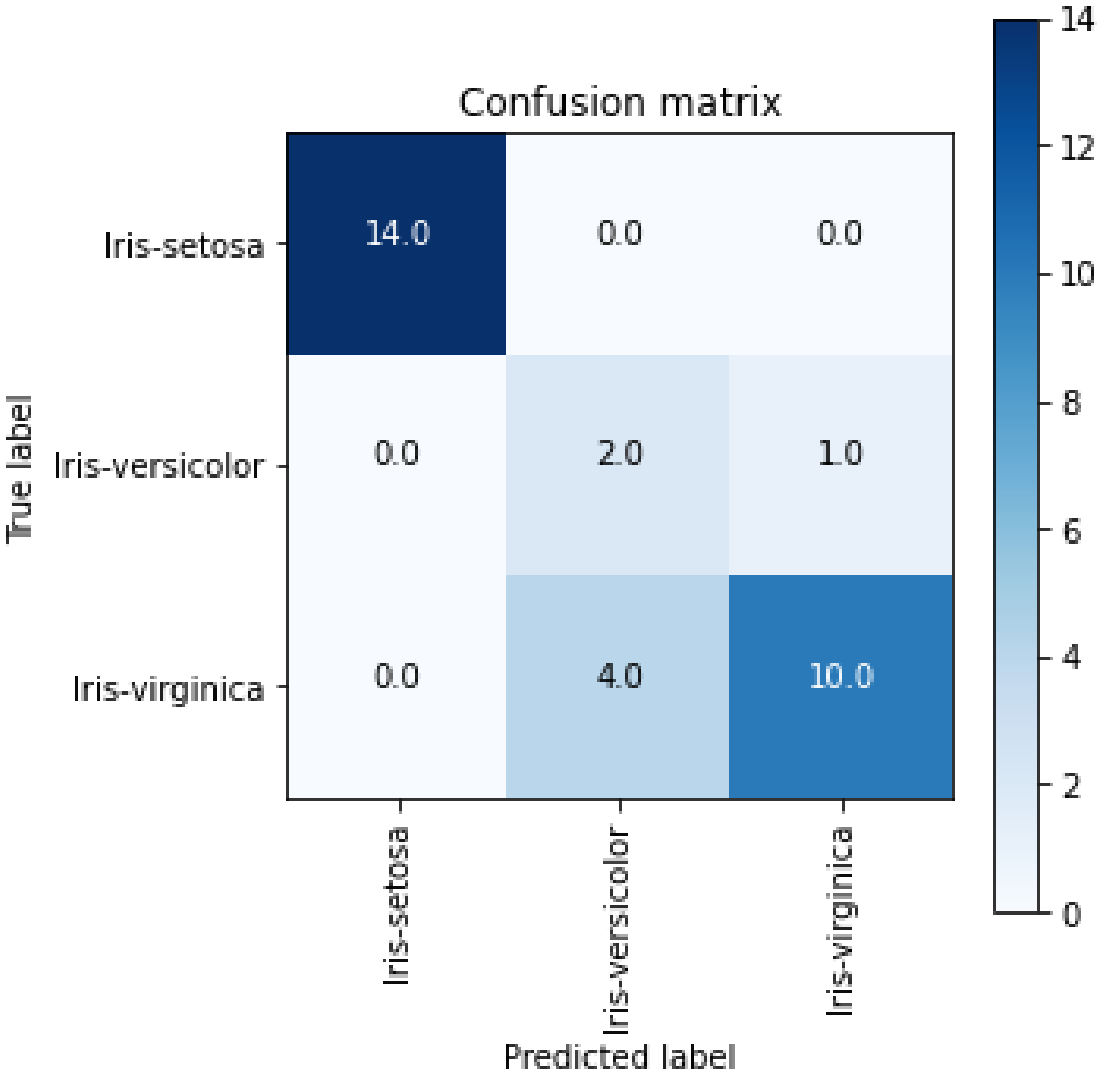
که در این مسئله برای سه کلاس مذکور داریم:

آ) **Iris Setosa:** $f_1s = 1$

ب) **Iris Versicolor:** $f_1s = 0.44$

ج) **Iris Virginica:** $f_1s = 0.8$

۳. **تحلیل و نتیجه‌گیری**: در این بخش مشاهده شد که در کلاس دوم کمی معیار اندازه کمی دارد، که این عمدتا به دلیل کم بودن دیتاست و تا حدی تحت تاثیر رندوم بودن نمونه برداری. در بخش آینده خواهیم دید که نتایج پکیج‌ها اندکی با پکیج‌های فعلی تفاوت دارد که بخشی از این مسئله مرتبط با تصادفی بودن این نمونه برداری است. راهکار مناسب استفاده از روش‌های مناسب ارزیابی مانند Cross Validation می‌باشد.



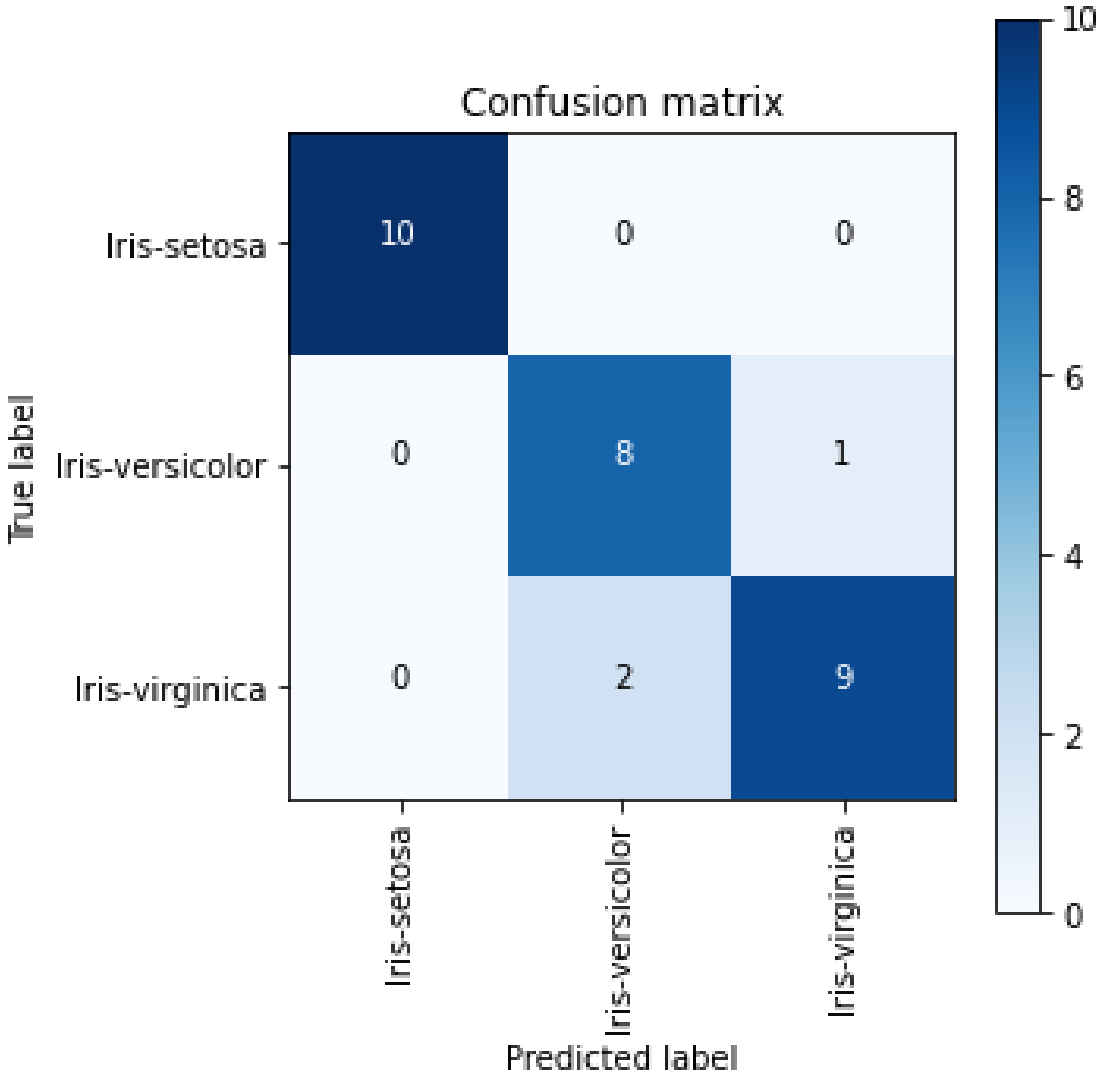
شکل ۲.۸: ماتریس درهم ریختگی

۵.۸ د

حال تمام بخش‌های قبل را با استفاده از پکیج‌ها انجام می‌دهیم. مشاهده می‌گردد نتایج به شرح زیر است:

- **طبقه بندی بدون نرمالیزه** در این حالت دقت طبقه بندی برابر با 90% می‌باشد. حدس زده می‌شود که علت عمده این امر در متفاوت بودن تقسیم داده‌ها(تصادفی بودن) در این حالت با بخش "ب" است.
- **طبقه بندی بدون نرمالیزه** در این حالت دقت طبقه بندی برابر با 90% که این خلاف روند طبقه‌بند دستی می‌باشد، که در بخش "ب" با کاهش دقت همراه بوده است.بنابراین الگوریتم پکیج نسبت به این موضوع مقاوم است.

- **ماتریس درهم ریختگی**: در این حالت مشاهده می‌شود که ماتریس به صورت شکل ۳.۸



شکل ۳.۸: ماتریس درهم ریختگی

همانطور که ذکر شد علت اصلی تفاوت ماتریس درهم ریختگی دوحالت به صرف تصادفی بودن نمونه برداری است.

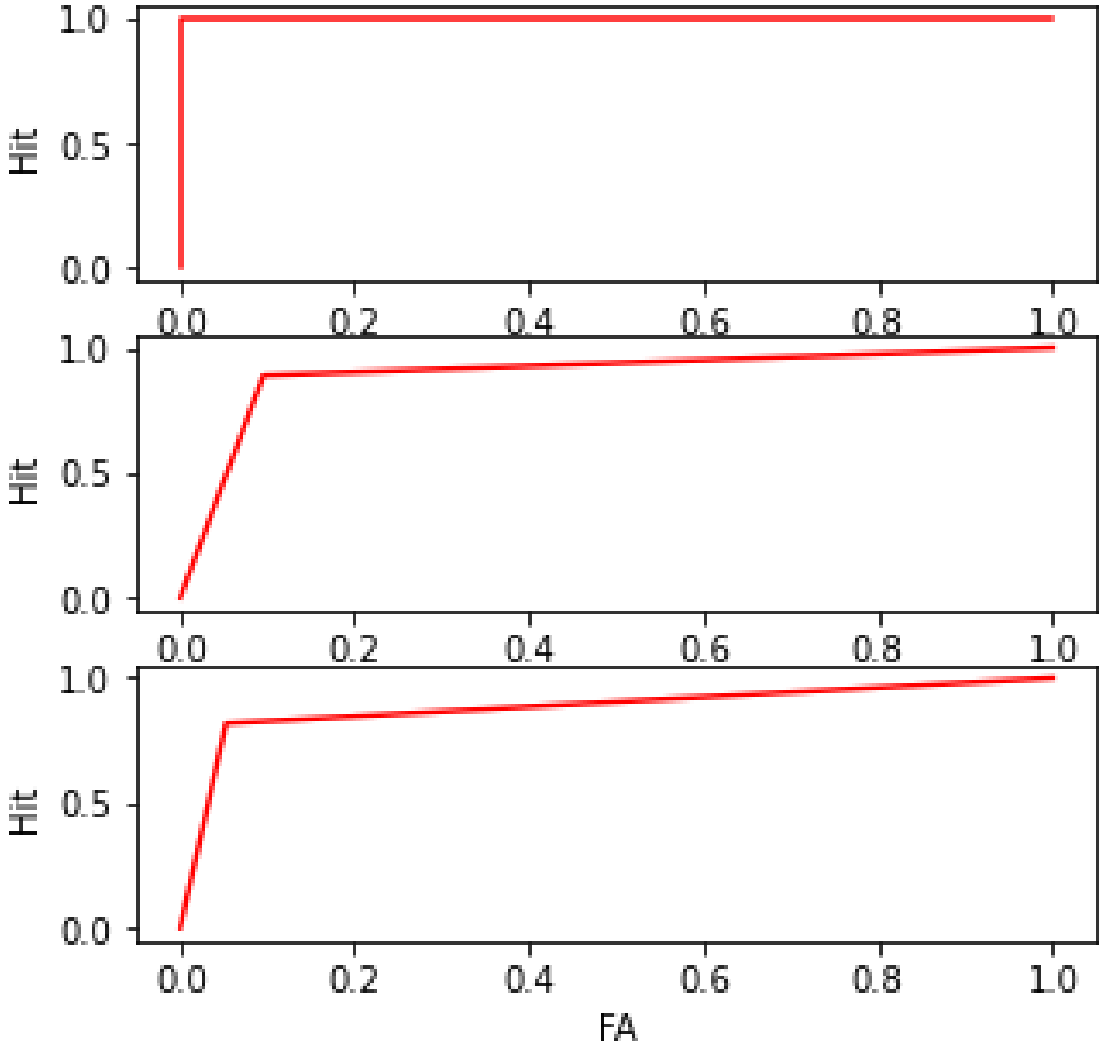
- **f1 score:** مشاهده می‌گرد که این معیار برای سه کلاس این مسئله به شرح زیر می‌باشد:

۱. **Iris Setosa:** $f_1s = 1$

۲. **Iris Versicolor:** $f_1s = 0.84$

۳. **Iris Virginica:** $f_1s = 0.86$

- **نمودارهای ROC**: همچنین نمودارهای ROC در شکل ۴.۸ آورده شده‌اند. مساحت هریک از نمودارها



شکل ۴.۸: ماتریس درهم ریختگی

نیز به ترتیب به شر زیر است:

۱. **کلاس ۱**: $Auc = 1$

۲. **کلاس ۲**: $Auc = 0.897$

۳. **کلاس ۳**: $Auc = 0.828$

۹
سوال ۹

:

۱.۹ الف

به صورت کلی می‌دانیم تعداد نمونه‌های لازم برای تخمین یک توزیع با بعد بردار ویژگی، رابطه‌ی نمایی دارد. برای مثال اگر در تخمین یک توزیع یک بعدی به تعداد N نمونه احتیاج داشته باشیم، آنگاه برای تخمین توزیع l بعدی به N^l بعد احتیاج داریم. این موضوع با توجه به هزینه بر بودن جمع آوری داده(هرینه زمانی، مکانی و ...) برای ما ایجاد مشکل می‌نماید. در بستر Naive Bayes برای حل این مشکل فرض می‌شود که ویژگی‌های مختلف نسبت به هم مستقل هستند، در این صورت می‌توان یک توزیع N بعدی را به صورت N توزیع یک بعدی در نظر گرفت آنگا برای تخمین این توزیع از رنج Nl داده احتیاج است که پیشرفت محسوسی نسبت به رابطه‌ی نمایی محسوب می‌شود. البته صحت این روش در منطقی بودن فرض استقلال ویژگی‌ها می‌باشد. به صورت کلی تابع چکالی به فرم زیر خواهد بود:

$$P(x|w_i) = \Pi_{j=1}^{l} p(x_j|w_i)$$

بنابراین برای پیشبرد الگوریتم Naive Bayes ، باید آماره‌های هر ویژگی را به شرط کلاس محاسبه نمود و سپس در رابطه‌ی کلی بیز از آن استفاده نمود،

در طبقه‌بندی به روش بیز بیهنه، میزان ارتباط [۱] ویژگی‌ها با یکدیگر در نظر گرفته می‌شوند. بنابراین باید

[1] Correlation

تمامی داده‌ها و ماتریس کواریانس آن‌ها را در نظر داشت که موجب افزایش بعد مسئله به صورت نمایی می‌شود. همچنین محاسبه‌ی ماتریسی بعد بالا در این روش از علل دیگر کاهش سرعت می‌باشد اما عملا می‌توان با پرداخت هزینه‌ی سرعت و پیچیدگی الگوریتم، دقت بهتری را کسب نمود. باید آماره‌های تمامی ویژگی‌ها به شرط کلاس‌های مختلف را محاسبه نمود و سپس در رابطه‌ی کلی بیز از آن استفاده کرد. بر مبنای توضیحات داده شده، عملا فرق دو الگوریتم در استقلال و عدم استقلال ویژگی‌هاست و در سایر موارد مشابه یکدیگر عمل می‌نمایند. پس از پیاده سازی کد، به بررسی عملکرد هرکدام می‌پردازیم.

۲.۹ ب

۲

۱.۲.۹ پیاده سازی کد الگوریتم

در این قسمت بدون پیش پردازش داده، الگوریتم را پیاده می‌نماییم و تنها با توجه به داده‌ها یک ضریب مناسب برای نرمی واریانس ویژگی‌ها تعیین کرده و در ابتدای امر به عنوان ورودی به متد طبقه بند وارد می‌نماییم. دقت نمایید در الگوریتم Optimal Bayes این ضریب نرمی، به صورت یک ضریب در ماتریس همانی تبدیل می‌شود.

- **Bayes Naive** کد این سوال در فایل P9Bayes.py ذخیره شده است. در پیاده سازی این کد داده‌های مختلف امکان ورود دارند(داده‌های Noisy Moon کامنت شده اند) و همچنین توابعی به منظور جدا سازی داده‌ها آموزش و تست طراحی گردیده. سپس سعی شده است که میانگین هر ویژگی به شرح یک کلاس محاسبه گردد و ماتریسی متناسب با کلاس و ویژگی برای واریانس‌ها و میانگین‌ها طراجی گردد. سپس در فرآیند بهسازی با استفاده از این موارد به طبقه بندی پرداخته می‌شود/
- **Bayes Optimal** این بخش نیز در فایل P9Optimal.py ذخیره شده است. این کد نیز مشابه بخش قبل است با این تفاوت که ماتریس کواریانس تمام ویژگی‌ها ب شرط کلاس و میانگین تمامی ویزگی‌ها

[2] Smoothness

به شرط کلاس محاسبه شده و در فرآیند محاسبه استفاده می‌گردد.

۲.۲.۹ نتایج و تحلیل آن

مطابق روند خواسته شده، نتایج به شرح زیر می‌باشند:

جدول ۱.۹: معیارهای گزارش شده برای دو طبقه بند

Classifier	f score Class 10	f score Class 9	f score Class 8	f score Class 7	f score Class 6	f score Class 5	f score Class 4	f score Class 3	f score Class 2	f score Class 1	Acc
Naive Bayes	0.6688	0.592	0.765	0.779	0.554	0.633	0.743	0.710	0.812	0.864	0.717
Optimal Bayes	0.873	0.639	0.883	0.878	0.686	0.859	0.800	0.846	0.865	0.855	0.82

۳.۹ ب

حال بخش قبل را برای دیتاست Noisy Moons تکرار می‌کنیم:

جدول ۲.۹: معیارهای گزارش شده برای دو طبقه بند

Classifier	f score Class 2	f score Class 1	Acc
Naive Bayes	0.884	0.879	0.881
Optimal Bayes	0.934	0.924	0.93

حال با استفاده از پکیج پایتون طبقه بندی را انجام می‌دهیم و نتیجه به شرح زیر می‌باشد:

جدول ۳.۹: معیارهای گزارش شده برای دو طبقه بند

Acc	Data Set
0.678	Tiny Mnist
0.882	Nosiy Moon

۴.۹ تحلیل و نتیجه‌گیری

به طور کلی مشاهده می‌شود که دقت تصمیم بهینه، به شرط حضور داده به اندازه کافی از تصمیم‌گیر Naive بیشتر است که دلیل آن‌هم دقیق بودن الگوریتم است. اما علاوه بر عدم دسترسی به دیتای به اندازه کافی، می‌توان مشاهده کرد که حجم محاسبه‌زیاد نیز از دیگر اشکالات Optimal است که عمده‌ی این محاسبات از معکوس ماتریس نشئت می‌گیرد.

۱۰
سوال۱۰

۱.۱۰ مقدمه

در روش one vs rest می‌توان مسئله را به یک مسئله‌ی طبقه‌بندی باینری تبدیل کرد. بدین صورت که در هر مرحله یک کلاس خاص در نظر گرفته شده و باقی کلاس‌ها به عنوان یک کلاس بزرگ غیر، تلقی می‌گردد. سپس با استفاده از الگوریتم مناسب بدنبال جدا سازی(طبقه بندی) کلاس مورد نظر و کلاس بزرگ غیر می‌باشیم. این کار را برای تمامی کلاس‌ها تکرار می‌نماییم و در نهایت مجموعه‌ی تمامی طبقه بندها را در نظر می‌گیریم.

همچنین روش Logistic Regression یک روش هوشمندانه برای تبدیل مسئله‌ی طبقه‌بندی به مسئله رگرسیون است. بدین ترتیب که از تابع واصل سیگموئید به عنوان تابعی استفاده می‌شود که نتایج رگرسیون را در بازه‌ی احتمال ($[0, 1]$) تصویر می‌نماید. سپس می‌توان با اتخاذ روش One Vs Rest مسئله را یک مسئله طبفه بندی دو کلاسه در نظر گرفت و سپس به بهینه کردن تابع Likelihood با استفاده از روش های عددی پرداخت.

در این سوال از تابع Log Likelihood به عنوان تابع هزینه و روش Gradient Ascent برای بیشینه کردن آن استفاده می‌شود. این روش به صورت کلی در دسته‌ی روش‌های Generative قرار می‌گیرد و به شرط داشتن تعداد داده‌های آموزش به اندازه کافی عملکرد مناسبی خواهد داشت.

۲.۱۰ شرح الگوریتم و کد

کد مربوط به این سوال در فایل P10.py ذخیره گردیده است.

در این روش از تابع هزینه‌ی Log Likelihood استفاده می‌گردد. این تابع هزینه برای نمونه مشاهده شده i

ام به شرح زیر می‌باشد:

$$l(\theta) = \Sigma_{i=1}^{n} y^i \log h(x^i) + (1 - y^i) \log(1 - h(x^i))$$

همچنین برای بیشنیه کردن این تابع هزینه از روش گرادیان افزایشی [۱] استفاده می‌نماییم. در واقع داریم:

$$\begin{cases} \theta := \theta + \alpha \nabla_\theta l(\theta) \\ \nabla_\theta l(\theta) = (y - h_\theta(x))x \end{cases}$$

که در آن مقصود از α همان طور پله می‌باشد. در این مسئله قصد استفاده از الگوریتم گرادیان تصادفی [۲] را داریم به همین دلیل فرآیند به روز رسانی را برای هر مشاهده i انجام می‌دهیم. در نهایت قانون به روز رسانی به فرم زیر می‌باشد:

$$\theta := \theta + \alpha(y^i - h_\theta(x^i))x^i$$

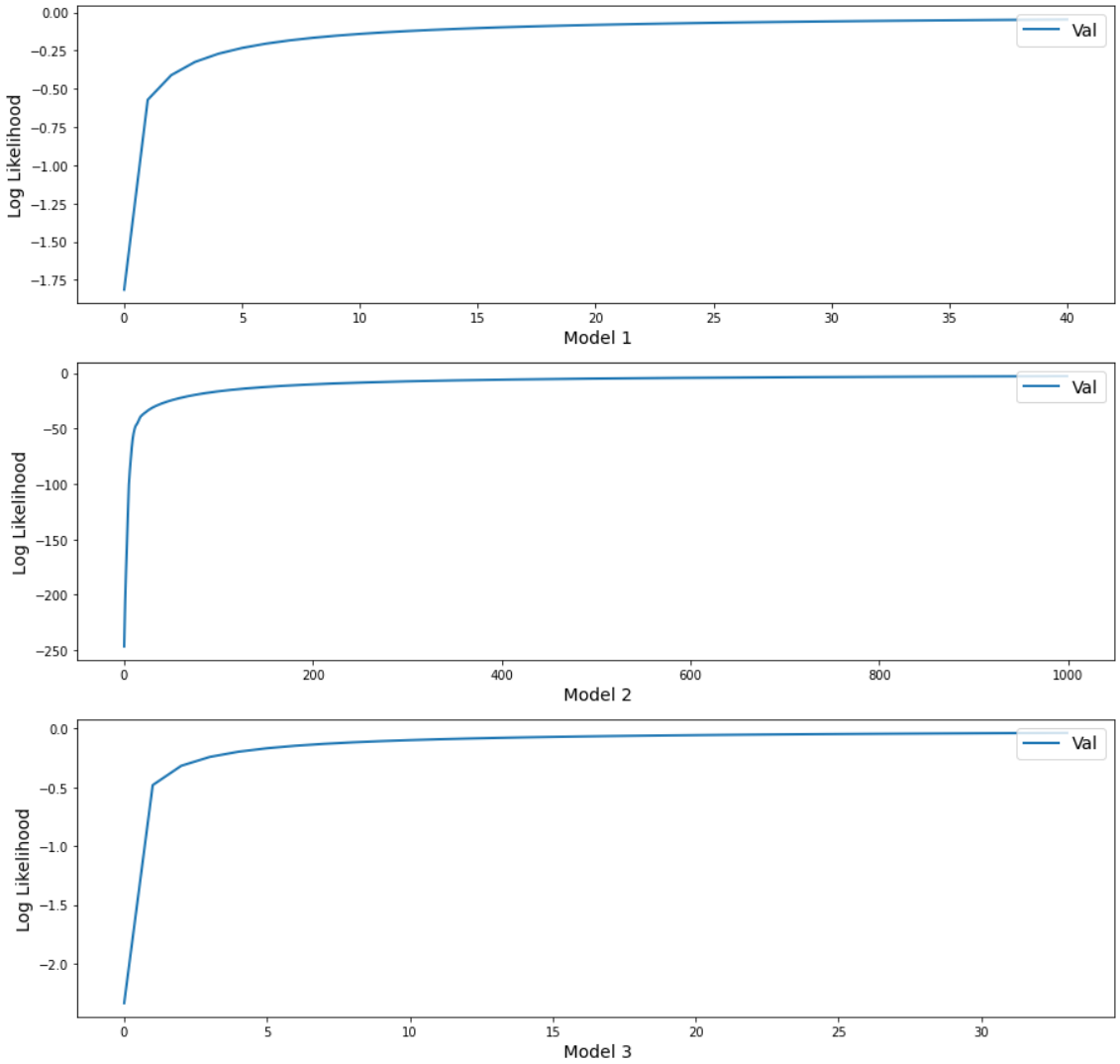
که با استفاده از طول پله مناسب و نقطه اولیه می‌توان الگوریتم را احرا نمود. در این مسئله نقطه اولیه به صورت : $[x, y, bias]^T = [-0.01, -0.01, 0.0]$ و طول پله به صورت استاتیک برابر با: $\alpha = 0.001$ انتخاب شده و شرط پایان هم به صورت ترکیب حداکثر تعداد تکرار مجاز 1000 بار و تلورانس کمتر از 0.001 برای تابع هزینه در نظر گرفته شده است.

۳.۱۰ نتایج اجرای الگوریتم

نمودار تابع هزینه الگوریتم به صورت زیر می‌باشد:

[1] Gradient Ascent

[2] Stochastic Gradient Descent



شکل ۱.۱۰: تابع هزینه در مراحل مختلف اجرای الگوریتم

همچنین پس از اجرای الگوریتم مشاهده می‌شود خطوط جدا ساز به صورت زیر عمل می‌نمایند. همانطور که مشاهده می‌شود، خط مربوط به کلاس توانسته است آن کلاس خاص را به خوبی از دو کلاس دیگر جدا کند. معادلات خطوط به شرح زیر است:

- **کلاس ۱:**

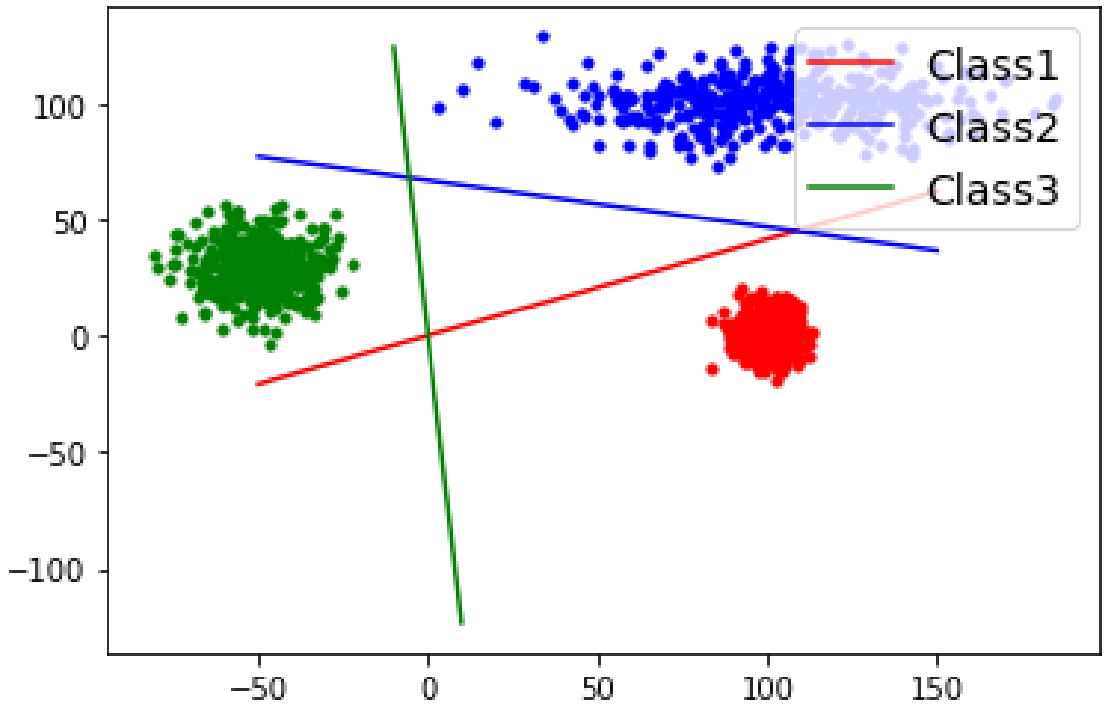
$$0.127x - 0.307y = 0.00136$$

- **کلاس ۲:**

$$0.0284 + 0.14y = 9.38$$

- **کلاس ۳:**

$$-0.256x - 0.0207y = -0.0033$$



شکل ۲.۱۰: تابع هزینه در مراحل مختلف اجرای الگوریتم

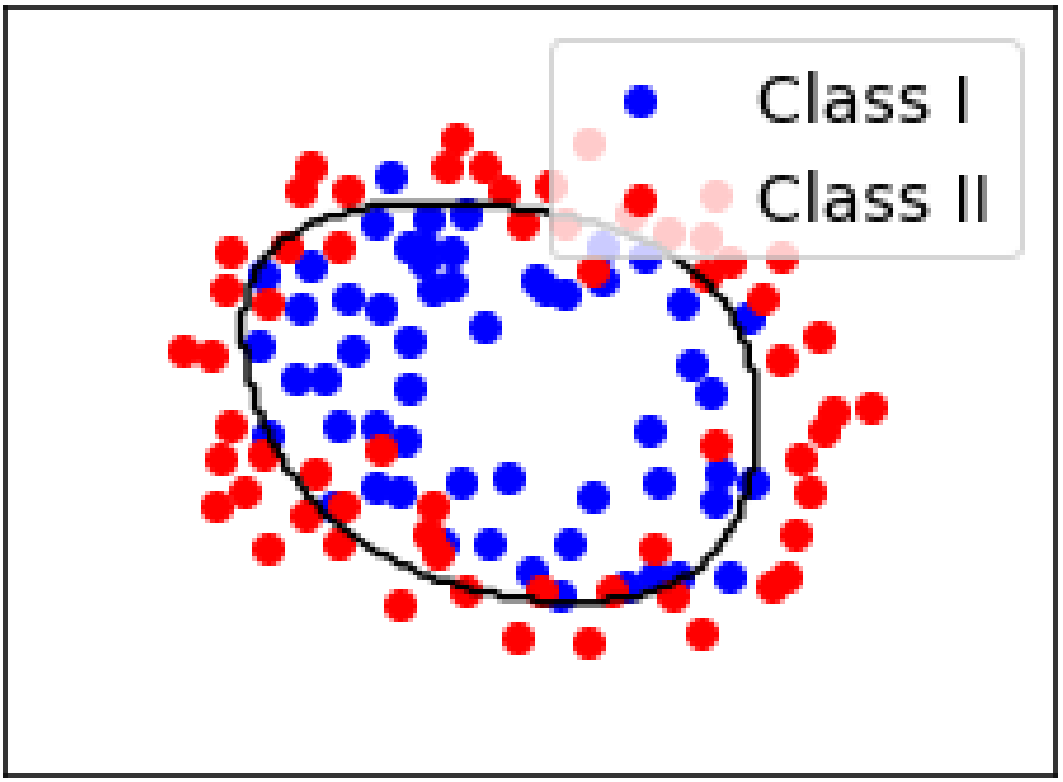
۱۱
سوال ۱۱

با توجه به آنکه داده‌ها به صورت خطی جداپذیر نیستند با استفاده از یک نگاشت مناسب آنرا به خمینه [۱] با بعد بالاتر می‌بریم. در خمینه مذکور داده‌ها به صورت خطی جدا پذیرند. بنابراین آنها را با صفحه مورد نظر حدا می‌کنیم. از انجا که این خمینه دارای یک نگاشت ایزمور می‌باشد، می‌توان صفحه را در بعد بالا تبدیل به بعد پایین نمود. در شکل پایین این نواحی را برای داده‌های تست ملاحظه می‌فرمایید. همانطور که مشاهده می‌شود، داده‌ها با مرزی بسته جدا شده‌اند.

۱.۱۱ شرح کد

در این قسمت، از پکیج‌های آماده استفاده شده است. در ابتدا مش بندی را حول تمامی داده‌ها انجام داده، سپس داده‌ها را به بعد بالا برده طبقه بندی را انجام می‌دهیم و در نهایت بر اساس برچسب‌های پیش‌بینی شده نواحی در فضای اولیه دو بعدی را تعیین می‌نماییم. برای طبقه بندی در بعد بالا نیز از Logistic Regression استفاده می‌نماییم.

[1] Manifold



شکل ۱.۱۱: طبقه‌بندی داده‌ها به صورت غیرخطی