دانشگاه صنعتی اصفهان

دانشکده مهندسی برق و کامپیوتر

یادگیری ماشین

تمرین سری 2

نام و نام خانوادگی	آرین تشکر
شماره دانشجویی	40023494
تاریخ ارسال گزارش	1401/03/08

فهرست گزارش سوالات

سوال 1 – محاسبه Gini Index

الف)

$$G(D) = 1 - \sum_{i=1}^{2} p_i^2 = 1 - (\left(\frac{10}{20}\right)^2 + \left(\frac{10}{20}\right)^2) = 1 - 0.5 = 0.5$$

نکته ی حائز اهمیت در این بخش این است که چون با case ای طرف هستیم که باید منجر به بیشینه ی Gini Index بشود (از هر کلاس به تعداد برابر در گره موجود است)، بیشینه ی مقدار Gini Index یعنی $1-\frac{1}{k}$ را خواهیم داشت. از آنجایی که مسئله دو کلاسه است، بنابراین $k = 2$ و $\mathrm{G_{max}} = 1 - \frac{1}{2} = 0.5$، همان مقداری ما بدست آوردیم.

ب) در حالت Multiway، در ازای هر حالت نامی برای CustomerID، یک گره در درخت خواهیم داشت و از آنجایی که به ازای هر CustomerID دقیقاً یک کلاس وجود دارد، آنتروپی هر گره به صورت مجزا برابر صفر خواهد بود. در حقیقت داریم:

$$G_{CustomerID}(D) = \frac{1}{20}(1-1^2) + \frac{1}{20}(1-1^2) + \cdots + \frac{1}{20}(1-1^2) = 0$$
$$\gamma(CustomerID, D) = 0.5 - 0 = 0.5$$

پ) در حالت Multiway، به ازای هر حالت نامی Gender یک گره در درخت خواهیم داشت. از آنجایی که ویژگی Gender دو حالته است بنابراین دو گره خواهیم داشت و در نتیجه داریم:

$$G_{Gender}(D) = \frac{10}{20}\left(1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2\right) + \frac{10}{20}\left(1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2\right) = 0.48$$
$$\gamma(Gender, D) = 0.5 - 0.48 = 0.02$$

ت) مانند قسمت های قبل، ویژگی Car Type دارای 3 حالت نامی است و داریم:

$$G_{CarType}(D) = \frac{4}{20}\left(1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2\right) + \frac{8}{20}\left(1 - \left(\frac{8}{8}\right)^2 - \left(\frac{0}{8}\right)^2\right) + \frac{8}{20}\left(1 - \left(\frac{1}{8}\right)^2 - \left(\frac{7}{8}\right)^2\right) = 0.1625$$
$$\gamma(CarType, D) = 0.5 - 0.1625 = 0.3375$$

ث) مانند قسمت های قبل، ویژگی ShirtSize دارای 4 حالت نامی است و داریم:

$$G_{ShirtSize}(D) = \frac{5}{20}\left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right) + \frac{7}{20}\left(1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2\right) + \frac{4}{20}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right) + \frac{4}{20}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right) \approx 0.49$$

$\gamma(ShirtSize, D) = 0.5 - 0.49 = 0.01$

ج) با توجه به مقادیر γ محاسبه شده، ویژگی Car Type بهترین ویژگی برای اولین Split است.

چ) با وجود این که با افراز روی ویژگی شناسه ی مشتری می توانستیم به مقدار γ برابر 0.5 نیز دستیابی پیدا کنیم (کمترین مقدار Gini Score ممکن و بیشتر مقدار Gini Index of Diversity)، از آنجایی که این ویژگی به طور ذاتی هیچ اطلاعاتی در مورد سوژه های مورد نظر به همراه ندارد و صرفاً یک شماره شناسه ی یکتا برای هر سوژه است، نباید از آن به عنوان یک ویژگی مورد تست در درخت تصمیم استفاده شود.

سوال ۲ – اثبات رابطه حساسیت Softmax

اگر رابطه softmax را به صورت Vectorized بنویسیم خواهیم داشت:

$$softmax(z) = \begin{bmatrix} e^{z_1} \\ e^{z_2} \\ \vdots \\ e^{z_J} \end{bmatrix} = \frac{e^{z}}{\sum_{1 \le j \le J} e^{z_j}}$$

اکنون داریم:

$$softmax(z+c) = \begin{bmatrix} e^{z_1+c} \\ e^{z_2+c} \\ \vdots \\ e^{z_J+c} \end{bmatrix} = \frac{e^{z+c}}{\sum_j e^{z_j+c}} = \frac{e^{z}e^{c}}{\sum_j e^{z_j}e^{c}} = \frac{e^{z}e^{c}}{e^{c}\sum_j e^{z_j}} = \frac{e^{z}}{\sum_j e^{z_j}} = softmax(z)$$

سوال 3 – تاثیر موارد مختلف بر روی Bias و Variance یک مدل Logistic Regression

مدل رگرسیون مورد نظر به وضوح از مشکل Overfitting (بیش برازش) رنج می برد چرا که پس از گذشت زمان اندک، میزان خطای Train در حال کاهش بوده در حالی که خطای Validation رو به افزایش است. این بدان معناست که مدل در حال یادگیری ویژگی های تمیز دهنده ی بین نمونه های مختلف موجود در دیتاست Train است و بنابراین در حال از دست دادن قدرت تعمیم پذیری خود می باشد.

الف) افزودن ویژگی های جدید در صورتی که تعداد نمونه های آموزشی نیز به تبع آن زیاد نشوند، وضعیت بیش برازش کنونی را بد تر می کند. با افزودن ویژگی های جدید بدون تغییر تعداد نمونه های آموزشی، معیار هایی بیشتری برای تمیز دادن نمونه های درون یک کلاس موجود خواهد بود، بنابراین مدلی که بیش از حد برای مسئله ی پیش رو پیچیده است، می تواند راحت تر روی نمونه های آموزشی بیش برازش شود. با افزودن ویژگی های جدید انتظار می رود که bias مدل کاهش و variance آن افزایش پیدا کند.

ب) بزرگتر کردن مجموعه ی آموزشی یکی از بهترین روش های مقابله با Overfitting است. در صورتی که مجموعه ی آموزشی بزرگتری در اختیار داشته باشیم، قدرت تعمیم پذیری مدل به تبع آن افزایش خواهد یافت چراکه مدل روی بخش بزرگ تری از فضای کل مسئله آموزش دیده است. انتظار می رود که با بزرگتر کردن مجموعه ی آموزشی bias مدل افزایش و variance آن کاهش پیدا کند.

ج) بزرگتر کردن پارامتر منتظم سازی نیز می تواند به این وضعیت کمک کند. پیش تر توضیح دادیم که در وضعیت کنونی مدل بیش از حد برای مسئله ی پیش رو پیچیده است. بنابراین اگر وزن پارامتر های اضافی مدل توسط یک پارامتر منتظم سازی بزرگ، متعادل شوند، می توان انتظار داشت که در نهایت با یک مدل ساده تر مواجه باشیم که بتواند با قدرت تعمیم پذیری بهتری مسئله را حل کند. انتظار می رود که با بزرگتر کردن پارامتر منتظم سازی، bias مدل افزایش و variance آن کاهش پیدا کند.

سوال 4 – دسته بند SVM

الف) در این مسئله داده های ورودی به شرح زیر هستند:

$$x^{(1)} = \begin{bmatrix}2\\3\end{bmatrix}, y^{(1)} = -1$$
$$x^{(2)} = \begin{bmatrix}1\\4\end{bmatrix}, y^{(2)} = -1$$
$$x^{(3)} = \begin{bmatrix}4\\5\end{bmatrix}, y^{(3)} = 1$$

از آنجایی که مقدار C برای این مسئله مشخص نشده است، فرض می کنیم که با مسئله ی Hard Margin SVM مواجه هستیم. برای حل تحلیلی این مسئله بهتر است که مستقیماً از صورت Dual مسئله ی بهینه سازی SVM استفاده کنیم. یعنی:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)}, x^{(j)}\rangle$$
$$subject\ to: \begin{cases}\alpha_i \geq 0, \quad 1 \leq i \leq n\\ \displaystyle\sum_{i=1}^{n} \alpha_i y^{(i)} = 0\end{cases}$$

از آنجایی که مسئله در فضای دو بعدی فعلی قابل حل است، نیازی به استفاده کرنل غیر خطی نیست و بنابراین قرار می دهیم $\langle x^{(i)}, x^{(j)}\rangle = x^{(i)}.x^{(j)}$. پس از جایگذاری تمام نقاط خواهیم داشت:

$$W(\alpha) = \alpha_1 + \alpha_2 + \alpha_3 - \frac{1}{2}(13\alpha_1^2 + 14\alpha_1\alpha_2 - 23\alpha_1\alpha_3 + 17\alpha_2^2 + 14\alpha_1\alpha_2 - 24\alpha_2\alpha_3 + 41\alpha_3^2 - 23\alpha_1\alpha_3 - 24\alpha_2\alpha_3)\ (I)$$

از طرفی طبق قید دوم مسئله، داریم:

$$-\alpha_1 - \alpha_2 + \alpha_3 = 0 \rightarrow \alpha_3 = \alpha_1 + \alpha_2\ (II)$$

با جایگزینی (II) در (I) داریم:

$$W(\alpha) = -4\alpha_1^2 - 8\alpha_1\alpha_2 + 2\alpha_1 - 5\alpha_2^2 + 2\alpha_2$$

برای بیشینه کردن این تابع، لازم است که مشتق آن را نسبت به α برابر صفر قرار دهیم. یعنی:

$$\nabla W(\alpha) = \begin{bmatrix}\dfrac{\partial W(\alpha)}{\partial \alpha_1}\\ \dfrac{\partial W(\alpha)}{\partial \alpha_2}\end{bmatrix} = \begin{bmatrix}-8\alpha_1 - 8\alpha_2 + 2\\ -8\alpha_1 - 10\alpha_2 + 2\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$$

بنابراین باید دستگاه زیر را حل کنیم:

$$\begin{cases} -8\alpha_1 - 8\alpha_2 + 2 = 0 \\ -8\alpha_1 - 10\alpha_2 + 2 = 0 \end{cases}$$

با کم کردن معادله ی دوم از معادله ی اول داریم:

$$2\alpha_2 = 0 \to \alpha_2 = 0\ (III)$$

با جایگذاری (III) در یکی از معادلات دستگاه $\alpha_1 = \frac{1}{4}\ (IV)$ بدست خواهد آمد. همچنین با جایگذاری (III)و (IV) در (II)، مقدار α_3 نیز برابر $\frac{1}{4}$ بدست خواهد آمد. اکنون برای بدست آوردن ضرایب معادله ی SVM و با توجه به این که ضرایب لاگرانژ بدست آمده برای مسئله ی dual در مسئله ی primal نیز صدق می کنند، می توانیم از یکی از شرایط KKT مسئله ی primal به شرح زیر استفاده کنیم:

$$w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)} = \frac{1}{4}(-1)\left(\begin{bmatrix}2\\3\end{bmatrix}\right) + 0 + \frac{1}{4}(1)\left(\begin{bmatrix}4\\5\end{bmatrix}\right) = \begin{bmatrix}0.5\\0.5\end{bmatrix}$$

اکنون برای محاسبه ی تنها پارامتر باقی مانده یعنی b کافیست از یکی از complementary slackness condition های KKT استفاده کنیم اما از آنجایی که پیش از حل مسئله نمی دانیم که کدام یک از نقاط support vector هستند، مجبوریم b را به ازای تمام complementary slackness condition ها محاسبه کنیم و بررسی کنیم که کدام b پاسخ مطلوب را به ما می دهد.

بنابراین داریم:

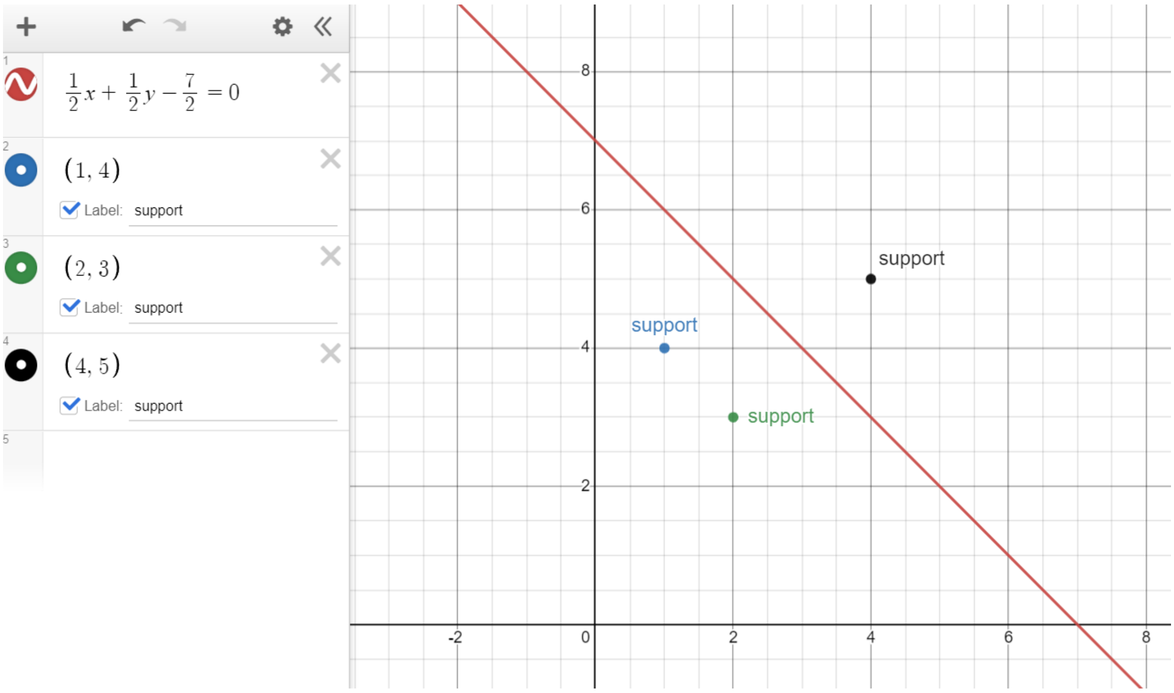
$$\begin{cases} b_1: y^{(1)}\left(w.x^{(1)} + b_1\right) - 1 = 0 \to (-1)\left(\begin{bmatrix}0.5\\0.5\end{bmatrix}.\begin{bmatrix}2\\3\end{bmatrix} + b_1\right) = 0 \to b_1 = -\frac{7}{2} \\ b_2: y^{(2)}\left(w.x^{(2)} + b_2\right) - 1 = 0 \to (-1)\left(\begin{bmatrix}0.5\\0.5\end{bmatrix}.\begin{bmatrix}1\\4\end{bmatrix} + b_2\right) = 0 \to b_2 = -\frac{7}{2} \\ b_3: y^{(3)}\left(w.x^{(3)} + b_3\right) - 1 = 0 \to (1)\left(\begin{bmatrix}0.5\\0.5\end{bmatrix}.\begin{bmatrix}4\\5\end{bmatrix} + b_3\right) = 0 \to b_3 = -\frac{7}{2} \end{cases}$$

بنابراین هر سه نقطه ی داده شده بردار پشتیبان هستند و $b = -\frac{7}{2}$.

در نهایت برای محاسبه ی margin می توانیم از رابطه ی زیر استفاده کنیم:

$$m = \frac{2}{||w||} = \frac{2}{\sqrt{0.5}} = \frac{2}{0.5\sqrt{2}} = 2\sqrt{2}$$

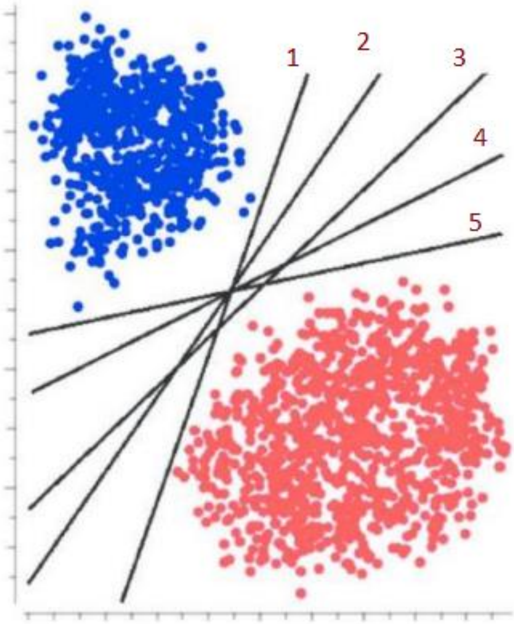
ب) هر سه نقطه بردار پشتیبان هستند.



شکل 1: بردار های پشتیبان و مرز تصمیم **SVM**

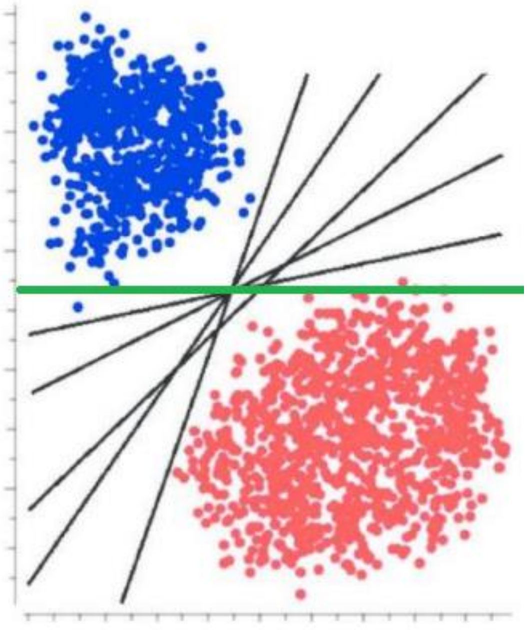
سوال 5 – Logistic Regression با منظم سازی

اگر فرض کنیم که $h_\theta(x) = e^{z(x;\theta)}, z(x;\theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_0$ آنگاه می توانیم تاثیر منتظم سازی روی هر کدام از پارامتر ها را بررسی کنیم (در ضمن از شکل زیر به عنوان مرجع در قسمت های بعدی پاسخ استفاده خواهیم کرد):



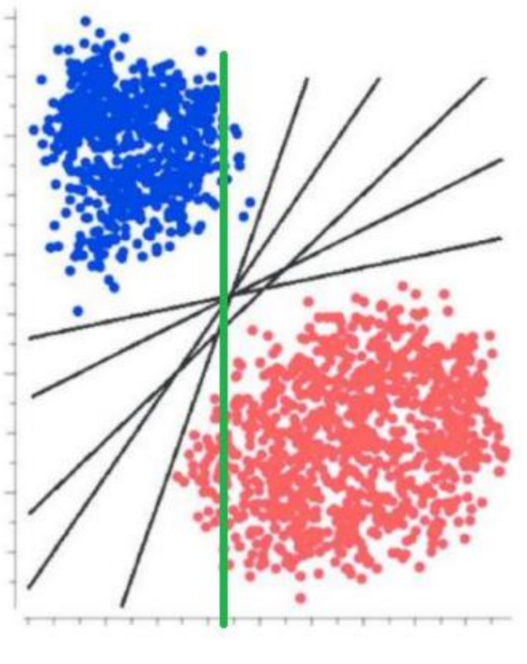
شکل 2: مرز تصمیم **Logistic Regression** به ازای منتظم سازی پارامتر های متفاوت

1. منتظم سازی θ_0: از آنجایی که θ_0 جمله ی بایاس را مرز تصمیم مدل تشکیل می دهد، با زیاد کردن مقدار λ برای θ_0، مرز تصمیم به سمت عبور از مبدا مختصات می رود (عرض از مبدا یا بایاس صفر). در شکل (2)، انتظار می رود که مرز تصمیم به حوالی مرز (2) همگرا شود. از آنجایی که چنین مرز تصمیمی می تواند تمام داده های آموزشی را به طور بی نقص دسته بندی کند، انتظار می رود که خطای نهایی نزدیک به حالت بهینه باشد و تغییر چندانی در مقدار loss function مشاهده نشود.
2. منتظم سازی θ_1: این پارامتر مربوط به ویژگی x_1 (محور افقی در شکل (2)) می باشد. در صورت زیاد کردن مقدار λ برای این پارامتر انتظار می رود که مرز تصمیم به حوالی مرز شماره 5 در شکل (2) همگرا شود (مقدار x_1 در آن تقریباً بی تاثیر شود). اگر مقدار λ خیلی بزرگ باشد مرز تصمیم به یک خط افقی تبدیل خواهد شد و بهترین خط افقی ممکن بر روی داده های موجود همچنان دارای چندین نقطه ی اشتباه دسته بندی شده می باشد. بنابراین با افزایش اندک مقدار loss function نسبت به حالت منتظم نشده، مواجه خواهیم بود.



شکل 3: مرز تصمیم احتمالی پس از منتظم سازی روی $\boldsymbol{\theta_1}$

3. منتظم سازی θ_2: حالت دوگان θ_1، با منتظم سازی روی این پارامتر انتظار داریم که مرز تصمیم به حوالی مرز 1 در شکل (2) همگرا شود و به ازای مقادیر بسیار بزرگ λ، مرز تصمیم یک خط عمودی خواهد بود. همانطور که از تصویر به وضوح مشخص است، بهترین خط افقی دارای تعداد زیادی نقطه ی به اشتباه دسته بندی شده خواهد بود و بنابراین با این منتظم سازی مقدار loss function به میزان قابل توجهی افزایش خواهد یافت.



شکل 4:مرز تصمیم احتمالی پس از منتظم سازی روی $\boldsymbol{\theta_1}$