دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

یادگیری ماشین

تمرین شماره ۱

نام و نام خانوادگی

امیرحسین چمی

آبان ۱۴۰۴

فهرست

فهرست شکل‌ها

فهرست جدول‌ها

چکیده

در این تمرین روش های مختلف طبقه بندی را مورد بررسی قرار می دهیم. سوالات ۱ الی ۵ بر مفاهیم تئوری تمرکز دارند. سوال ۱ مفهوم تصمیم‌گیری با حداقل ریسک و استفاده گزینه‌ی reject را بررسی می کند و می خواهد دو نامساوی مربوط به قانون تصمیم گیری را اثبات کنیم. سوال ۲ تحلیل تئوری از naive bayes است و باید مستقیما محاسبات انجام انجام دهیم. در واقع در این تمارین naive bayes داریم اثر مستقل فرض کردن ویژگی ها را می بینیم. سوال ۳ درباره‌ی naive bayes چند متغیره است و اثبات می کند که مرز تصمیم گیری در آن خطی است. سوال ۴ درباه‌ی MLE در حالت چندمتغیره است. این سوال صرفا دنبال کردن روش حل مسائل MLE است. سوال ۵ تخمین MAP با توزیع پیشین Dirichlet را مطرح می کند و تاثیر پارامتر های آلفا را روی احتمال بررسی می کند.

سوالات ۶ الی ۸ وارد فاز عملی می شوند. سوال ۶ یک دیتاست ارائه می کند و ابتدا با فرض مشترک بودن کوواریانس و سپس کوواریانس متفاوت برای هر کلاس، مرز تصمیم را بررسی می کند. همچنین در این سوال اثر تغییر احتمال پیشین را بررسی و گزینه ی reject با هزینه‌ی لاندا را تحلیل می کنیم. سوال ۷ شامل پیاده سازی کامل gaussian naive bayes است که تخمین پارامتر ها، طبقه بندی داده ها و ارزیابی عملکرد مدل را شامل می شود. سوال ۸ برای حل یک مسئله هم از MLE و هم از MAP استفاده می کند و می بینیم که MLE اصلا عملکرد خوبی ندارد و علت این عملکرد ضعیف را بررسی می کنیم. به نظر من تمرین یک در مجموع تعادل خوبی از مسائل تئوری و پیاده سازی کد دارد.

پرسش ۱ – Minimum Risk Classification with Rejection Option

در این سوال باید یک نامساوی را در مسئله ی طبقه بندی حداقل ریسک (طبقه بندی بیزی) اثبات کنیم.

۱-۱. اثبات نامساوی

مطابق صورت سوال cost به صورت زیر تعریف شده است.

$$h(\omega_i|\omega_j) = \begin{cases} ۰ & if\ \omega_i = \omega_j \quad (i.e., CorrectClassification) \\ \lambda_r & if\ \omega_i = \omega_. \quad (i.e., Rejection) \\ \lambda_s & Otherwise \quad (i.e., SubstitutionError) \end{cases} \tag{1}$$

مطابق آنچه در کلاس درس دیدیم، Expected loss به صورت زیر تعریف می شود.

$$R(\alpha_i|x) = \sum_{j=1}^{C} \lambda(\alpha_i|\omega_j)\, P(\omega_j|x) \tag{2}$$

در رابطه ی (۲)، α_i اقدام اتخاذ شده است. برای مسئله ی ۱ داریم: $\alpha_i = \omega_i$

در این مسئله‌ی مینیمم ریسک، برای انتخاب کلاس ω_i باید داشته باشیم:

$$R(\alpha_i|x) \le R(\alpha_j|x)\,, \forall j \tag{3}$$

با جایگذاری مقادیر داریم.

$$\lambda_s(۱ - P(\omega_i|x)) \le \lambda_s\left(۱ - P(\omega_j|x)\right)\,, \forall j \tag{4}$$

با تقسیم طرفین نامساوی بر $\lambda_s > ۰$ خواهیم داشت.

$$۱ - P(\omega_i|x) \le ۱ - P(\omega_j|x)\,, \forall j \tag{5}$$

در نهایت

$$P(\omega_i|x) \ge P(\omega_j|x)\,, \forall j \tag{6}$$

نامساوی اول اثبات شد.

در این مسئله‌ی مینینمم ریسک، ریسک رد کردن باید مساوی یا بیشتر از ریسک بقیه ی حالات باشد در غیر این صورت تصمیم گیری ما بهینه نیست. در واقع داریم:

$$R(\omega_.|x) \ge R(\omega_i|x) \tag{۷}$$

حال هر بخش را جداگانه محاسبه می کنم. مقدار جمع ریسک برای طبقه بندی درست به علاوه ی *substitution error* به صورت زیر است.

$$R(\omega_i|x) = \sum_{j\neq i} \lambda(\omega_i|\omega_j)\, P(\omega_j|x) = ۰ . P(\omega_i|x) + \lambda_s \sum_{j\neq i} P(\omega_j|x) \qquad (۸)$$

مقدار ریسک برای رد کردن به صورت زیر است.

$$R(\omega_.|x) = \lambda_r \qquad (۹)$$

از طرفی می دانیم جمع تمامی احتمالات مساوی ۱ می شود. پس:

$$\sum_{j=۰}^{C} P(\omega_j|x) = ۱ \qquad (۱۰)$$

$$۱ - P(\omega_i|x) = \sum_{j\neq i} P(\omega_j|x) \qquad (۱۱)$$

می توانیم رابطه‌ی (۱۱) را در (۸) اعمال کنیم.

$$R(\omega_i|x) = \lambda_s\, P(\omega_j|x) = \lambda_s\, [۱ - P(\omega_i|x)] \qquad (۱۲)$$

حال از (۹) و (۱۲) در (۷) استفاده می کنم.

$$\lambda_r \geq \lambda_s\, [۱ - P(\omega_i|x)] \qquad (۱۳)$$

$$P(\omega_i|x) \geq ۱ - \frac{\lambda_r}{\lambda_s} \qquad (۱٤)$$

نامساوی دوم اثبات شد.

پرسش ۲ – Naive Bayes

در این سوال محاسبات یک طبقه بندی **Naive Bayes** را انجام می دهیم. سپس با توجه به محاسبات، رابطه‌ی بین کلاس ها و ویژگی ها را تحلیل و بررسی می کنیم. ابتدا این مبحث را از جزوه ی درس و [۱] مرور کردم.

۲-۱. بخش a: $P_{Y|X_1,X_2}(y|x_1=0,x_2=0)$

ابتدا مقادیر احتمال هر حالت را با جایگذاری مقادیر در توزیع ها محاسبه می کنم.

ویژگی اول:

$$P_{X_1|Y}(x_1=0|Y=c)=1-\theta_c=\begin{cases}0.5 & if\ c=0\\0.25 & if\ c=1\\0.5 & if\ c=2\end{cases} \quad (15)$$

ویژگی دوم:

$$P_{X_2|Y}(x_2=0|Y=c)=\frac{1}{\sqrt{2\pi\,\sigma_c^2}}e^{-\frac{(0-\mu_c)^2}{2\sigma_c^2}}=\begin{cases}0.241970724 & if\ c=0\\0.398942280 4 & if\ c=1\\0.241970724 5 & if\ c=2\end{cases} \quad (16)$$

همچنین با استفاده از قانون احتمال کل می توانیم احتمال مشترک ویژگی ها را محاسبه کرد.

$$P(X_1,X_2)=\sum_{c=0}^{2}P(x_1,x_2|Y=c)P(Y=c) \quad (17)$$

با توجه به naive بودن مسئله، قانون احتمال کل به صورت زیر خواهد بود.

$$P(X_1,X_2)=\sum_{c=0}^{2}P(x_1|Y=c)P(x_2|Y=c)P(Y=c) \quad (18)$$

با جایگذاری مقادیر در رابطه ی (۱۸) به ازای هر کلاس داریم:

$$\begin{cases}0.5*0.241970724 5*0.5 & if\ c=0\\0.25*0.398942280 4*0.25 & if\ c=1\\0.5*0.241970724 5*0.25 & if\ c=2\end{cases}=\begin{cases}0.060492681 13 & if\ c=0\\0.024933892 53 & if\ c=1\\0.030246340 56 & if\ c=2\end{cases} \quad (19)$$

در مجموع احتمال مشترک ویژگی ها از جمع مقادیر رابطه‌ی (۱۹) محاسبه می شود.

$$P(X_1=0,X_2=0)=0.115672914 2 \quad (20)$$

در جزوه‌ی درس و منبع [۱] دیدیم که رابطه‌ی posterior بر اساس قانون بیز به صورت زیر است. مساوی دوم در رابطه ی (۲۰) بر اساس naive بودن نوشته شده.

$$P(Y|X_1,X_2)=\frac{P(X_1,X_2|Y).P(Y)}{P(X_1,X_2)}=\frac{P(X_1|Y).P(X_2|Y).P(Y)}{P(X_1,X_2)} \quad (۲۱)$$

حال با جایگذاری داریم.

$$P(Y|x_1=0,x_2=0)=\frac{P(x_1=0|Y).P(x_2=0|Y).P(Y)}{P(x_1=0,x_2=0)} \quad (۲۲)$$

$$=\begin{cases}\frac{0.5*0.2419707245*0.5}{0.1156729142} & if\ c=0\\ \frac{0.25*0.3989422804*0.25}{0.1156729142} & if\ c=1\\ \frac{0.5*0.2419707245*0.25}{0.1156729142} & if\ c=2\end{cases}$$

$$=\begin{cases}۰.۵۲۲۹۶۳۲۳۲۵ & if\ c=۰\\ ۰.۲۱۵۵۵۵۱۵۱۳ & if\ c=۱\\ ۰.۲۶۱۴۸۱۶۱۶۳ & if\ c=۲\end{cases}$$

که اگر بخواهیم در فضای $\mathbb{R}^3$ نمایش بدهیم، داریم.

$$P(Y|x_1=0,x_2=0)=[۰.۵۲۲۹۶۳۲۳۲۵,۰.۲۱۵۵۵۵۱۵۱۳,۰.۲۶۱۴۸۱۶۱۶۳] \quad (۲۳)$$

جمع این مقادیر برابر ۱ می شود. (برای اینکه این جمع برابر یک باشد، مقادیر را تا ده الی یازده رقم اعشار محاسبه کردم.)

۲-۲. بخش b: $P_{Y|X_1}(y|x_1=0)$

با استفاده از قانون بیز داریم:

$$P_{Y|X_1}(Y|x_1=0)=\frac{P(x_1=0|Y)P(Y)}{P(x_1=0)} \quad (۲٤)$$

عبارت $P(x_1=0|Y)$ را در رابطه‌ی (۱۵) محاسبه کردم. عبارت $P(Y)$ را هم در صورت مسئله داریم. اما مخرج کسر با استفاده از قانون احتمال کل محاسبه می شود.

$$P(x_1=0)=\sum_{c=0}^{2}P(x_1=0|Y=c)P(Y=c) \quad (۲۵)$$
$$=۰.۵*۰.۵+۰.۲۵*۰.۲۵+۰.۵*۰.۲۵=۰.۴۳۷۵$$

حال با جایگذاری در (۲۴) خواهیم داشت:

$$P_{Y|X_1}(Y|x_1 = ۰) = \frac{P(x_1 = ۰\ |Y)P(Y)}{P(x_1 = ۰)} = \begin{cases} \frac{۰.۵ * ۰.۵}{۰.۴۳۷۵} & if\ c = ۰ \\ \frac{۰.۲۵ * ۰.۲۵}{۰.۴۳۷۵} & if\ c = ۱ \\ \frac{۰.۵ * ۰.۲۵}{۰.۴۳۷۵} & if\ c = ۲ \end{cases} \quad (۲۶)$$

$$= \begin{cases} ۰.۵۷۱۴۲۸۵۷۱۴ & if\ c = ۰ \\ ۰.۱۴۲۸۵۷۱۴۲۹ & if\ c = ۱ \\ ۰.۲۸۵۷۱۴۲۸۵۷ & if\ c = ۲ \end{cases}$$

اگر بخواهیم این مقادیر را در فضای $\mathbb{R}^3$ بنویسیم:

$$P_{Y|X_1}(Y|x_1 = ۰) = [۰.۵۷۱۴۲۸۵۷۱۴,۰.۱۴۲۸۵۷۱۴۲۹,۰.۲۸۵۷۱۴۲۸۵۷] \quad (۲۷)$$

جمع این مقادیر برابر ۱ می شود. (برای اینکه این جمع برابر یک باشد، مقادیر را تا ده رقم اعشار محاسبه کردم.)

۲-۳. بخش c: $P_{Y|X_2}(y|x_2 = 1)$

با استفاده از قانون بیز داریم:

$$P_{Y|X_۲}(Y|x_۲ = ۱) = \frac{P(x_۲ = ۱\ |Y)P(Y)}{P(x_۲ = ۱)} \quad (۲۸)$$

عبارت $P(Y)$ در صورت مسئله داده شده. $P(x_۲ = ۱\ |Y)$ را به صورت زیر محاسبه می کنم.

$$P(x_۲ = ۱\ |Y) = \frac{۱}{\sqrt{۲\pi\ \sigma_c^۲}} e^{-\frac{(۱-\mu_c)^۲}{۲\sigma_c^۲}} = \begin{cases} ۰.۰۵۳۹۹۰۹۶۶۵۱ & if\ c = ۰ \\ ۰.۲۴۱۹۷۰۷۲۴۵ & if\ c = ۱ \\ ۰.۳۹۸۹۴۲۲۸۰۴ & if\ c = ۲ \end{cases} \quad (۲۹)$$

مخرج کسر با استفاده از قانون احتمال کل محاسبه می شود

$$P(x_۲ = ۱) = \sum_{c=۰}^{۲} P(x_۲ = ۱|Y = c)P(Y = c) \quad (۳۰)$$

$$= ۰.۵ * ۰.۰۵۳۹۹۰۹۶۶۵۱ + ۰.۲۵ * ۰.۲۴۱۹۷۰۷۲۴۵ + ۰.۲۵ * ۰.۳۹۸۹۴۲۲۸۰۴ = ۰.۱۸۷۲۲۳۷۳۴۵$$

حال مقادیر (۳۰) و (۲۹) را در (۲۸) قرار می دهیم.

$$P_{Y|X_2}(Y|x_2 = 1) = \frac{P(x_2 = 1\,|Y)P(Y)}{P(x_2 = 1)} = \begin{cases} \frac{0.053990966651 * 0.5}{0.187223373345} & if\ c = 0 \\ \frac{0.241970724245 * 0.25}{0.187223373345} & if\ c = 1 \\ \frac{0.398942280404 * 0.25}{0.187223373345} & if\ c = 2 \end{cases} \quad (31)$$

$$= \begin{cases} 0.144188836001 & if\ c = 0 \\ 0.323103698808 & if\ c = 1 \\ 0.532707940409 & if\ c = 2 \end{cases}$$

اگر بخواهیم این مقادیر را در فضای $\mathbb{R}^3$ بنویسیم:

$$P_{Y|X_2}(Y|x_2 = 1) = [0.144188836001, 0.323103698808, 0.532707940409] \quad (32)$$

جمع این سه مقدار با تقریب بسیار خوبی برابر ۱ می شود.

۲-۴. بخش d: تحلیل بخش های قبلی

از هر بخش می توان چیزهایی یافت و تحلیل هایی ارائه کرد. ابتدا باید اشاره کرد که در بخش های قبلی از قانون بیز استفاده شد. مخرج کسر در قانون بیز با توجه به ثبات بودن برای تمامی کلاس ها (بسته به هر بخش) می تواند از تحلیل کنار گذاشته شود زیرا صرفا به عنوان یک ضریب نرمالیزاسیون عمل می کند.

یکی از راه های تحلیل مقادیر هر بخش این است که ببینیم هر prior چه تاثیری بر likelihood گذاشته تا خروجی posterior را داده است.

تحلیل بخش a

از محاسبات پایانی بخش a یعنی روابط (۲۲) و (۲۳) می بینیم که کلاس ۰ بیشترین و کلاس ۱ کمترین احتمال پسین را دارد. علت آن این است که توزیع احتمال $X_1 = 0$ بیشترین احتمال را به کلاس های ۰ و۲ و کمترین را به کلاس ۰ می دهد و در مقابل برای $X_2 = 0$ با توجه به مقدار μ احتمال کلاس ۱ کمی بیشتر می شود. از طرفی prior بیشترین احتمال را به کلاس ۰ می دهد.

ترکیب این ها باعث می شود که کلاس ۰ همچنان محتمل ترین کلاس باشد. کلاس ۱ به دلیل احتمال کمی که $X_1 = 0$ و prior به آن می دهند، تاثیر احتمال بالایی که ویژگی $X_2 = 0$ به کلاس ۱ می دهد را کم می کند و باعث می شود کم احتمال ترین کلاس باشد.

تحلیل بخش b

در این بخش تنها ویژگی $X_1 = 0$ مورد استفاده قرار می گیرد. در محاسبات احتمال پسین در (۲۶) و (۲۷) می بینیم که کلاس ۰ بیشترین احتمال و کلاس ۱ کمترین احتمال را می گیرند. می توان گفت که کلاس ۰ قوی تر و کلاس ۱ ضعیف تر از بخش a ظاهر می شوند و کلاس ۲ کمترین تغییر (حدود ۰.۲ افزایش) را دارد. حضور

$X_1 = 0$ به نفع کلاس ۰ و ۲ است و به ضرر کلاس ۱ می باشد. در مقابل می توان گفت، حضور $X_2 = 0$ کمی به نفع کلاس ۱ بود.

تحلیل بخش c

در این بخش تنها ویژگی $X_2 = 1$ مورد استفاده قرار می گیرد. در محاسبات احتمال پسین در (۳۱) و (۳۲) می بینیم که کلاس ۲ بیشترین و کلاس ۰ کمترین احتمال را دارد. مقدار $X_2 = 1$ دقیقا برابر با میانگین کلاس ۲ می باشد و باعث می شود احتمال کلاس ۲ به شدت افزایش یابد و به نفع آن باشد. در مقابل هر چی از مقدار میانگین $\mu = 1$ دور می شیم (به علت ساختار نمودار توزیع گوسی) احتمال کلاس های دیگر کمتر می شود. کمااینکه می بینیم کمترین احتمال به کلاس ۰ که میانگینش ۱- است می رسد.

پس به طور خلاصه $X_2 = 1$، به نفع کلاس ۲ بوده و آن را غالب می کند. کلاس ۱ را در رتبه ی دوم و کلاس ۰ را به پایین ترین مقدار می رساند.

پرسش ۳– Multinomial Naive Bayes

این سوال هم مانند سوال قبل درباره‌ی naive bayes می باشد ولی محاسبات سوال قبل را ندارد و صرفا اثبات روابط است. مباحث مربوط به این سوال را از جزوه ی درس و تدریس دکتر ابولقاسمی و منبع [۱] مرور کردم.

۳-۱. بخش a: $P(y=1|x)$

با استفاده از قانون بیز داریم:

$$P(y=۱|x)=\frac{P(x|y=۱)P(y=۱)}{P(x|y=۰)P(y=۰)+P(x|y=۱)P(y=۱)} \quad (۳۳)$$

اگر document مورد بررسی L کلمه داشته باشد. آنگاه $L=\sum_{i=۱}^{d} c_i$ است. یعنی دقیقا برابر جمع تعداد دفعات تکرار همه ی کلمات است (بدیهی است). توزیع این document مطابق توزیع چندجمله‌ای (multinomial) به صورت زیر است.

برای کلاس ۰:

$$P(x|y=۰)=\frac{(\sum_{i=۱}^{d} c_i)!}{\prod_{i=۱}^{d} c_i!}\prod_{i=۱}^{d} P_{i|۰}^{c_i} \quad (۳۴)$$

برای کلاس ۱:

$$P(x|y=۱)=\frac{(\sum_{i=۱}^{d} c_i)!}{\prod_{i=۱}^{d} c_i!}\prod_{i=۱}^{d} P_{i|۱}^{c_i} \quad (۳۵)$$

توجه کنید که notation های $P_{i|۱}$ و $P_{i|۰}$ را خود سوال معرفی کرده است.

حال (۳۴) و (۳۵) را در (۳۳) جایگذاری می کنم.

$$P(y=۱|x)=\frac{\frac{(\sum_{i=۱}^{d} c_i)!}{\prod_{i=۱}^{d} c_i!}\prod_{i=۱}^{d} P_{i|۱}^{c_i}\,.P_۱}{\frac{(\sum_{i=۱}^{d} c_i)!}{\prod_{i=۱}^{d} c_i!}\prod_{i=۱}^{d} P_{i|۱}^{c_i}\,.P_۱+\frac{(\sum_{i=۱}^{d} c_i)!}{\prod_{i=۱}^{d} c_i!}\prod_{i=۱}^{d} P_{i|۰}^{c_i}.P_۰} \quad (۳۶)$$

حال با ساده سازی عبارات مشترک صورت و مخرج داریم:

$$P(y=۱|x)=\frac{\prod_{i=۱}^{d} P_{i|۱}^{c_i}\,.P_۱}{\prod_{i=۱}^{d} P_{i|۱}^{c_i}\,.P_۱+\prod_{i=۱}^{d} P_{i|۰}^{c_i}.P_۰} \quad (۳۷)$$

تا اینجا ساده سازی این عبارت تمام شده؛ اما من در درس تئوری آشکارسازی تخمین (دانشجوی مخابرات هستم) دیدم که برای راحتی بیشتر از معادل های $P_۱=\pi$ و $P_۰=۱-\pi$ در اکثر مواقع استفاده می شود. برای همین رابطه ی (۳۷) را به این صورت هم می توان نوشت.

$$P(y=1|x)=\frac{\pi\,.\prod_{i=1}^{d}P_{i|1}^{c_i}}{\pi\,.\prod_{i=1}^{d}P_{i|1}^{c_i}+(1-\pi)\prod_{i=1}^{d}P_{i|.}^{c_i}} \quad (۳۸)$$

واضحا، روابط (۳۷) و (۳۸) عین یکدیگرند.

۳-۲. بخش b: Linearity of Decision Boundary

در مبحث Two-Category Classification از جزوه ی درس برای likelihood ratio نامساوی زیر را دیدیم.

$$\frac{P(x|\omega_1)}{P(x|\omega_2)}>\frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}\frac{P(\omega_2)}{P(\omega_1)} \quad (۳۹)$$

در درس تئوری آشکارسازی و تخمین بارها دیدیم که وقتی مقدار loss ها مشخص نشده اند، باید آن ها را اصطلاحا uniform loss در نظر بگیریم (از اصطلاحات تئوری آشکارسازی تخمین). به این معنی که کسر مربوط به loss ها در نامساوی (۳۹) برابر ۱ می شود.

حالا مفروضات مسئله را در رابطه ی (۳۹) اعمال می کنیم.

$$\frac{P(x|y_1)}{P(x|y.)}>\frac{1-\pi}{\pi} \quad (۴۰)$$

با استفاده از (۳۴) و (۳۵) در (۴۰) داریم:

$$\prod_{i=1}^{d}\frac{P_{i|1}^{c_i}}{P_{i|.}^{c_i}}>\frac{1-\pi}{\pi} \quad (۴۱)$$

از دو طرف این رابطه لگاریتم می گیریم.

$$\sum_{i=1}^{d}c_i\log\left(\frac{P_{i|1}}{P_{i|.}}\right)>\log\left(\frac{1-\pi}{\pi}\right) \quad (۴۲)$$

$$\sum_{i=1}^{d}c_i\log\left(\frac{P_{i|1}}{P_{i|.}}\right)+\log\left(\frac{\pi}{1-\pi}\right)>0 \quad (۴۳)$$

حال اگر $x=\sum_{i=1}^{d}c_i$ و $\theta_i=\log\left(\frac{P_{i|1}}{P_{i|.}}\right)$ و $\theta_.=\log\left(\frac{\pi}{1-\pi}\right)$ و همچنین $\theta=(\theta_1,\theta_2,\dots,\theta_d)$ می توانیم (۴۳) را به صورت زیر بازنویسی کنیم.

$$\sum_{i=1}^{d}c_i\log\left(\frac{P_{i|1}}{P_{i|.}}\right)+\log\left(\frac{\pi}{1-\pi}\right)=\theta^Tx+\theta_0>0 \quad (۴۴)$$

که با در نظر گرفتن مرز تصمیم گیری به صورت زیر خواهد بود.

$$\theta^Tx+\theta_0=0 \quad (۴۵)$$

که واضحا این یک رابطه ی خطی نسبت به x بوده که بایاس θ_0 دارد. (توجه کنید تغییر متغیر های $x, \theta_i, \theta_., \theta$ با توجه به بخش بعدی انتخاب شده‌اند.)

۳-۳. بخش c: Logistic Function

در بخش a رابطه‌ی (۳۴) را دیدیم. می خواهم به نحوی رابطه‌ی (۴۵) را در (۳۸) ایجاد کنم. با تقسیم صورت و مخرج آن بر $(۱ - \pi) \prod_{i=۱}^{d} P_{i|.}^{c_i}$ خواهیم داشت:

$$P(y = ۱|x) = \frac{\frac{\pi}{۱ - \pi} \prod_{i=۱}^{d} \left(\frac{P_{i|۱}}{P_{i|.}}\right)^{c_i}}{۱ + \frac{\pi}{۱ - \pi} \prod_{i=۱}^{d} \left(\frac{P_{i|۱}}{P_{i|.}}\right)^{c_i}} \quad (٤٦)$$

با توجه به تعریف متغیر ها در (۴۴) می توان نوشت:

$$\prod_{i=۱}^{d} \left(\frac{P_{i|۱}}{P_{i|.}}\right)^{c_i} = e^{\theta^T x} \quad , \quad \frac{\pi}{۱ - \pi} = e^{\theta_0} \quad (٤٧)$$

پس می توانیم (۴۶) را اینگونه بنویسیم.

$$P(y = ۱|x) = \frac{e^{\theta^T x + \theta_0}}{۱ + e^{\theta^T x + \theta_0}} = \frac{۱}{۱ + e^{-(\theta^T x + \theta_0)}} \quad (٤٨)$$

پس اثبات شد این احتمال شرطی یک تابع لجستیک است.

پرسش 4 – MLE

۴-۱. بخش a:

در جزوه ی درس دیدیم که likelihood function به صورت زیر است.

$$\mathcal{L}(\theta|D) = \prod_{i=1}^{N} p(X_i|\theta) \qquad (۴۹)$$

با توجه به توزیعی که در صورت سوال داده شده، اگر حتی یکی از نمونه ها خارج از لوزی باشد مقدار likelihood function صفر می شود. پس باید تک تک نمونه ها داخل لوزی باشند.

$$\|X_i\|_۱ \leq \theta \quad , \quad \forall i \qquad (۵۰)$$

واضح است که اگر بزرگ ترین نمونه در شرط بالا صدق کند، بقیه ی نمونه ها هم صدق خواهند کرد. پس می توان این شرط به صورت زیر نوشت.

$$\max_i \|X_i\|_۱ \leq \theta \qquad (۵۱)$$

با توجه به مستقل بودن نمونه ها و با توجه به (۴۹) می توان نوشت.

$$\mathcal{L}(\theta|D) = \prod_{i=۱}^{N} p(X_i|\theta) = \prod_{i=۱}^{N} \frac{۱}{۲\theta^۲} = \frac{۱}{(۲\theta^۲)^N} \qquad (۵۲)$$

تابع بدست آمده یک تابع اکیدا نزولی است که بیشترین مقدار آن به ازای کمترین مقدار θ بدست می آید. از طرفی در (۵۱) کمترین مقدار θ را یافتیم. پس:

$$\hat{\theta} = \max_{i=۱,\dots,N} \|X_i\|_۱ \qquad (۵۳)$$

این تخمین MLE می باشد.

۴-۲. بخش b:

توزیع $gamma(\alpha, \beta)$ به صورت زیر است.

$$f(x) = \frac{۱}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-۱} e^{-\frac{x}{\beta}} \qquad (۵۴)$$

تابع likelihood آن مطابق (۴۹) به صورت زیر محاسبه می شود.

$$\mathcal{L}(\beta|D) = \prod_{i=1}^{N} p(X_i|\beta) = \prod_{i=1}^{N} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} X_i^{\alpha-1} \exp\left(-\frac{X_i}{\beta}\right) \quad (۵۵)$$
$$= \left(\frac{۱}{\Gamma(\alpha)}\right)^{N} \beta^{-\alpha N} \left(\prod_{i=۱}^{N} X_i^{\alpha-۱}\right) \exp\left(-\frac{۱}{\beta}\sum_{i=۱}^{N} X_i\right)$$

در این رابطه D بردار شامل نمونه هاست. می خواهیم ماکزیمم این تابع را پیدا کنیم. می دانیم لگاریتم تابعی اکیدا صعودی است پس می توانیم به جای خود $\mathcal{L}(\beta|D)$ ماکزیمم لگاریتم آن را بیابیم.

$$\log(\mathcal{L}(\beta|D)) = N\log\left(\frac{۱}{\Gamma(\alpha)}\right) - \alpha N\log(\beta) + (\alpha-۱)\sum_{i=۱}^{N}\log(X_i) - \frac{۱}{\beta}\sum_{i=۱}^{N} X_i \quad (۵۶)$$

حال از (۵۶) مشتق می گیریم و مساوی صفر قرار می دهیم تا نقطه ای که ماکزیمم در آن رخ می دهد را بیابیم. این مشتق را با A نمایش می دهیم.

$$A = \frac{d\log(\mathcal{L}(\beta|D))}{d\beta} = -\frac{\alpha N}{\beta} + \frac{۱}{\beta^{۲}}\sum_{i=۱}^{N} X_i = ۰ \quad (۵۷)$$

$$-\alpha N\beta + \sum_{i=۱}^{N} X_i = ۰ \quad (۵۸)$$

در نتیجه خواهیم داشت:

$$\hat{\beta} = \frac{۱}{\alpha N}\sum_{i=۱}^{N} X_i \quad (۵۹)$$

طبق (۵۹) مقدار تخمین ML برای β بدست آمد. بخشی که در ادامه می نویسم نه در صورت سوال خواسته شده و نه در کلاس درس مطرح شده اما من می نویسم. در درس تئوری آشکارسازی و تخمین، خواندیم که حتما باید مشتق دوم را حساب کنیم تا از مقعر بودن تابع و ماکزیمم بودن نقطه ی بدست آمده در تخمین مطمئن شویم.

مشتق دوم با استفاده از (۵۷) به صورت زیر بدست می آید.

$$\frac{dA}{d\beta} = \frac{d^{۲}\log(\mathcal{L}(\beta|D))}{d\beta^{۲}} = \frac{\alpha N}{\beta^{۲}} - \frac{۲}{\beta^{۳}}\sum_{i=۱}^{N} X_i \quad (۶۰)$$

مقدار تخمین (۵۹) را در (۶۰) قرار می دهیم.

$$\frac{\alpha N}{\beta^{۲}} - \frac{۲}{\beta^{۳}}(\alpha N\beta) = -\frac{\alpha N}{\beta^{۲}} < ۰ \quad (۶۱)$$

توجه کنید $\beta^{۲}$ و N ذاتا مقادیر مثبت هستند و α هم فقط مقادیر مثبت می گیرید. (می توانید در مورد توزیع گاما در منبع [۲] بخوانید.) پس مشتق دوم منفی شد در نتیجه تابع مقعر بوده و نقطه ی بدست آمده در تخمین، ماکزیمم است.

پرسش ۵ – MAP

در این سوال مفهوم مزدوج بودن توزیع پیشین را مرور و از تخیمن MAP استفاده می کنیم.

۵-۱. بخش a: Conjugacy of the Dirichlet Prior

توزیع دیریکله با پارامتر u_i به صورت زیر تعریف شده.

$$\mathbb{P}(p_1, p_2, \dots, p_6) = \frac{\Gamma(\sum_{i=1}^{6} u_i)}{\prod_{i=1}^{6} \Gamma(u_i)} \prod_{i=1}^{6} p_i^{u_i - 1} \ \delta\left(\sum_{i=1}^{6} p_i - 1\right) \tag{۶۲}$$

توزیع دیریکله توزیع پیشین است. توزیع likelihood (multinomial) به صورت زیر است.

$$P(n_1, \dots, n_6 | p) = \frac{N!}{\prod_{i=1}^{6} n_i!} \prod_{i=1}^{6} p_i^{n_i} \tag{۶۳}$$

در (۶۳) مقدار $N = 1000$ تعداد دفعات پرتاب تاس است.

هدف این است که نشان دهیم توزیع پسین هم توزیع دیریکله دارد، پس روابط را در قانون بیز می گذاریم و ادامه می دهیم.

$$P(p|\{n_i\}) = \frac{p(\{n_i\}|p)\mathbb{P}(p)}{p(\{n_i\})} \tag{۶۴}$$

توجه کنید که مخرج در (۶۴) صرفا نقش یک ضریب نرمالیزاسیون دارد. پس می توانیم از آن صرف نظر کنیم.

$$\begin{aligned} p(\{n_i\}|p)\mathbb{P}(p) &= \left(\frac{N!}{\prod_{i=1}^{6} n_i!} \prod_{i=1}^{6} p_i^{n_i}\right)\left(\frac{\Gamma(\sum_{i=1}^{6} u_i)}{\prod_{i=1}^{6} \Gamma(u_i)} \prod_{i=1}^{6} p_i^{u_i - 1} \ \delta\left(\sum_{i=1}^{6} p_i - 1\right)\right) \\ &= \frac{N!\,\Gamma(\sum_{i=1}^{6} u_i)}{(\prod_{i=1}^{6} n_i!)(\prod_{i=1}^{6} \Gamma(u_i))}\left(\prod_{i=1}^{6} p_i^{n_i + u_i - 1}\right)\delta\left(\sum_{i=1}^{6} p_i - 1\right) \end{aligned} \tag{۶۵}$$

در رابطه ی (۶۵) کسر پایانی تنها یک ضریب است که نسبت به p ها ثابت است. پس می توانیم از آن ها صرف نظر کنیم.

$$P(p|\{n_i\}) \propto \left(\prod_{i=1}^{6} p_i^{n_i + u_i - 1}\right)\delta\left(\sum_{i=1}^{6} p_i - 1\right) \tag{۶۶}$$

رابطه‌ی بدست آمده مشابه توزیع دیریکله در (۶۲) است. پارامتر توزیع دیریکله بدست آمده برای توزیع پسین، به صورت زیر است.

$$\alpha_i = u_i + n_i \tag{۶۷}$$

پس در کل توزیع پسین یک توزیع دیریکله به صورت زیر دارد.

$$P(p|\{n_i\}) = \frac{\Gamma(\sum_{i=1}^{6}\alpha_i)}{\prod_{i=1}^{6}\Gamma(\alpha_i)}\prod_{i=1}^{6}p_i^{\alpha_i-1}\ \delta\left(\sum_{i=1}^{6}p_i - 1\right) \quad (68)$$
$$= \frac{\Gamma(\sum_{i=1}^{6}(u_i+n_i))}{\prod_{i=1}^{6}\Gamma(u_i+n_i)}\prod_{i=1}^{6}p_i^{u_i+n_i-1}\ \delta\left(\sum_{i=1}^{6}p_i - 1\right)$$

۵-۲. بخش b: Posterior Distribution

با جایگذاری مقادیر مشخص شده در صورت سوال در (۶۸)، می توانیم به صورت زیر احتمال خواسته شده را بنویسیم.

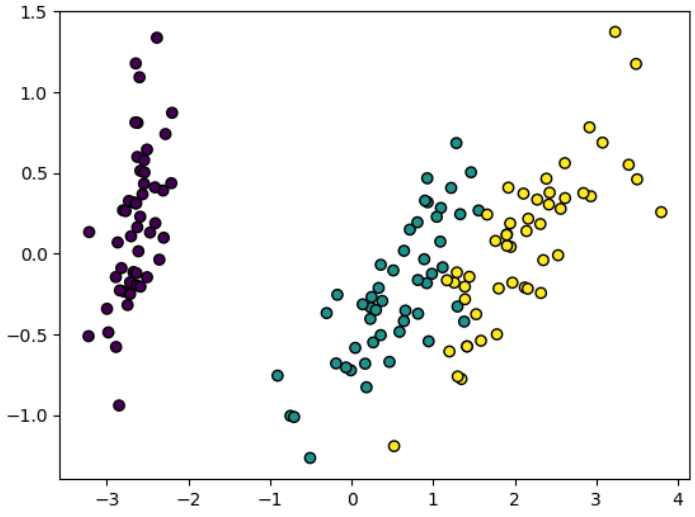
$$P(p_1,\dots,p_6|n_1,\dots,n_6) = \frac{\Gamma(\sum_{i=1}^{6}(i+n_i))}{\prod_{i=1}^{6}\Gamma(i+n_i)}\prod_{i=1}^{6}p_i^{i+n_i-1}\ \delta\left(\sum_{i=1}^{6}p_i - 1\right) \quad (69)$$
$$= \frac{\Gamma(1021)}{\prod_{i=1}^{6}\Gamma(u_i+n_i)}\prod_{i=1}^{6}p_i^{i+n_i-1}\ \delta\left(\sum_{i=1}^{6}p_i - 1\right)$$

توجه کنید که خود سوال گفته $u_i = i$.

پرسش ۶ – Gaussian Discriminant Analysis

در این تمرین کامپیوتری پارامترهای توزیع شرطی (روش پارامتریک تخمین توزیع) را محاسبه می کنیم، داده ها را با استفاده از یک discriminant function به کلاس های مختلف کلاس‌بندی می کنیم و مرز تصمیم گیری را تحلیل می کنیم.

داده های تولید شده برای دیتاست در نمودار زیر رسم شده اند.



شکل ۱. نمودار دیتاست سوال ۶

۶-۱. بخش a: محاسبه ی میانگین و کوواریانس

در این بخش باید با استفاده از داده ها، پارامتر های توزیع گوسی bivariate را محاسبه کنیم. خود سوال گفته که ماتریس کوواریانس برای هر دو کلاس مشترک است پس باید برای ساخت ماتریس کوواریانس مشترک، از pooling covariance استفاده شود. یعنی به این شکل:

$$\Sigma = \frac{1}{N_1 + N_2}\left(\sum_{i \in C_1}(x_i - \mu_1)(x_i - \mu_1)^T + \sum_{i \in C_2}(x_i - \mu_2)(x_i - \mu_2)^T\right) \quad (۷۰)$$

بردار میانگین و ماتریس کوواریانس هر کلاس به صورت زیر است.

Mean of class ۱: [۰,۵۳۳۲۰۶۵۷ -۰,۲۴۵۵۴۹۸۳]

Mean of class ۲: [۲,۱۰۹۲۰۸۸۹ ۰,۰۵۴۶۶۴۷۹]

Covariance of class ۱:

[[۰,۳۴۸۳۶۱۵۵ ۰,۱۹۴۷۶۷۳۲]

[٠,١٩٤٧٦٧٣٢ ٠,١٨١٤٦٨٧٨]]

Covariance of class ٢:

[[٠,٤٨٨٣٢٦٣٤ ٠,٢٧٠٣٣٧٧٦]

[٠,٢٧٠٣٣٧٧٦ ٠,٢٢٨٦٢٦٦٣]]

و نهایتا کوواریانس مشترک به صورت زیر است.

Shared covariance matrix (Σ):

[[٠,٤١٨٣٤٣٩٤ ٠,٢٣٢٥٥٢٥٤]

[٠,٢٣٢٥٥٢٥٤ ٠,٢٠٥٠٤٧٧١]]

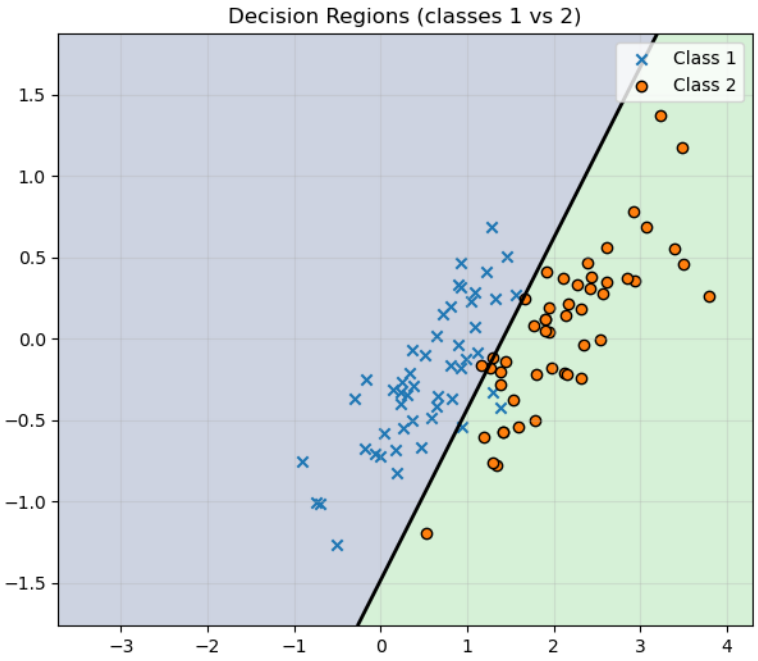
۶-۲. بخش b: Discriminant function

با توجه به توزیع چند متغیره و یکسان بودن کوواریانس برای هر دو کلاس، تابع discriminant به صورت زیر تعریف می شود.

$$g_i(x) = -\frac{١}{٢}(x-\mu_i)^T\Sigma^{-١}(x-\mu_i) + \ln P(\omega_i) \quad (٧١)$$

۶-۳. بخش c: Decision Regions

ناحیه های تصمیم گیری برای هر کلاس در تصویر مشخص شده اند.



شکل ۲. ناحیه های تصمیم گیری دو کلاس

۶-۴. بخش d: خط مرز تصمیم گیری

مرز ناحیه‌ی تصمیم گیری دقیقا جایی است که مقدار discriminant function به ازای هر دو کلاس برابر می شود.

$$g_1(x) = g_2(x) \tag{۷۲}$$

$$-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + \ln P(\omega_1) = -\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) + \ln P(\omega_2) \tag{۷۳}$$

به علت یکسان بودن ماتریس کوواریانس برای هر دو کلاس، جملات مرتبه دوم x در دو طرف تساوی، برابر شده و حذف می شوند. در نهایت به رابطه ی خط زیر می رسیم.

$$w^T x + b = 0 \tag{۷۴}$$

در این رابطه w^T برداری وزن ها و b بایاس (عرض از مبدا) خط است.

$$w = \Sigma^{-1}(\mu_1 - \mu_2) \tag{۷۵}$$

$$b = -\frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2) + \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) \tag{۷۶}$$

با استفاده از کد پایتون نوشته شده هر کدام از مقادیر (۷۵) و (۷۶) به صورت زیر بدست آمد.

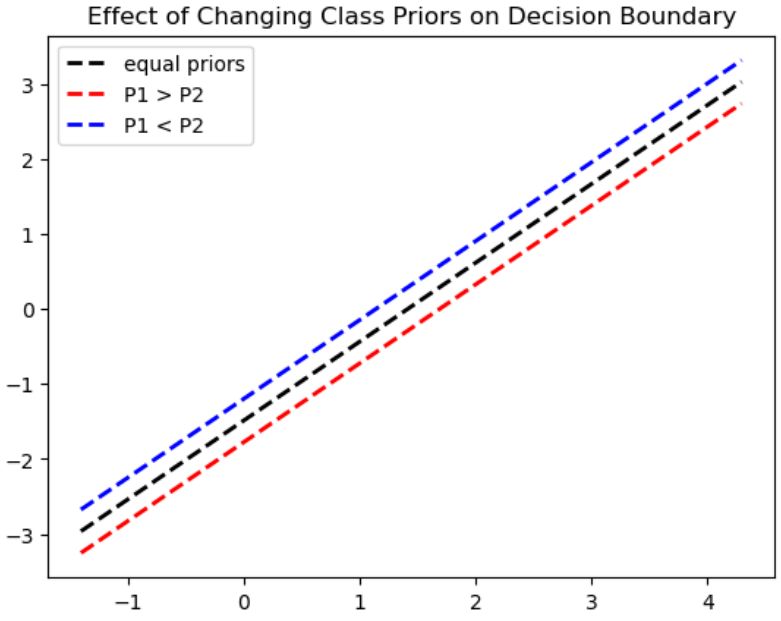
w (decision boundary vector): [-۷,۹۹۱۸۵۲۹۵ ۷,۵۹۹۷۴۹۰۲]

b (decision boundary bias): ۱۱,۲۸۴۲۳۷۱۲۷۳۶۰۱۵۴

۶-۵. بخش e: تاثیر تابع پیشین در نواحی تصمیم گیری

تصویر زیر مرز تصمیم گیری را به ازای زوج $(prior1, prior2)$ با مقادیر مختلف (۰.۵و۰.۵)، (۰.۱و۰.۹) و (۰.۹و۰.۱) نشان می دهد.

می بینیم که وقتی احتمال پیشین کلاس ۱ بیشتر است، مرز تصمیم به سمت کلاس ۲ می رود و ناحیه ی تصمیم گیری کلاس ۱ بزرگتر می شود (خطای نوع اول پیش می آید). وقتی احتمال پیشین کلاس ۲ بیشتر است، مرز تصمیم به سمت کلاس ۱ می رود و ناحیه ی تصمیم گیری کلاس ۲ بزرگتر می شود(خطای نوع دوم پیش می آید).



شکل ۳. تاثیر احتمال پیشین بر مرز تصمیم‌گیری

۶-۶. بخش f: تکرار سوال برای سه کلاس

حال باید بخش های a تا c را برای سه کلاس با فرض ماتریس کوواریانس نابرابر تکرار کرد. داده های سه کلاس در شکل ۱ رسم شده اند.

۶-۶-۱. بخش a: محاسبه ی میانگین و کوواریانس

بردار میانگین و ماتریس کوواریانس هر کلاس به صورت زیر است.

Mean of class ۰: [-۲,۶۴۲۴۱۵۴۶ ۰,۱۹۰۸۸۵۰۵]

Mean of class ۱: [۰,۵۳۳۲۰۶۵۷ -۰,۲۴۵۵۴۹۸۳]

Mean of class ۲: [۲,۱۰۹۲۰۸۸۹ ۰,۰۵۴۶۶۴۷۹]

Covariance of class ۰:

[[۰,۰۴۸۰۴۱۵ ۰,۰۵۴۹۲۱۸۵]

[۰,۰۵۴۹۲۱۸۵ ۰,۲۱۳۳۴۳۳۶]]

Covariance of class ۱:

[[۰,۳۴۸۳۶۱۵۵ ۰,۱۹۴۷۶۷۳۲]

[۰,۱۹۴۷۶۷۳۲ ۰,۱۸۱۴۶۸۷۸]]

Covariance of class ۲:

[[۰,۴۸۸۳۲۶۳۴ ۰,۲۷۰۳۳۷۷۶]

[۰,۲۷۰۳۳۷۷۶ ۰,۲۲۸۶۲۶۶۳]]

۶-۶-۲. بخش b: Discriminant function

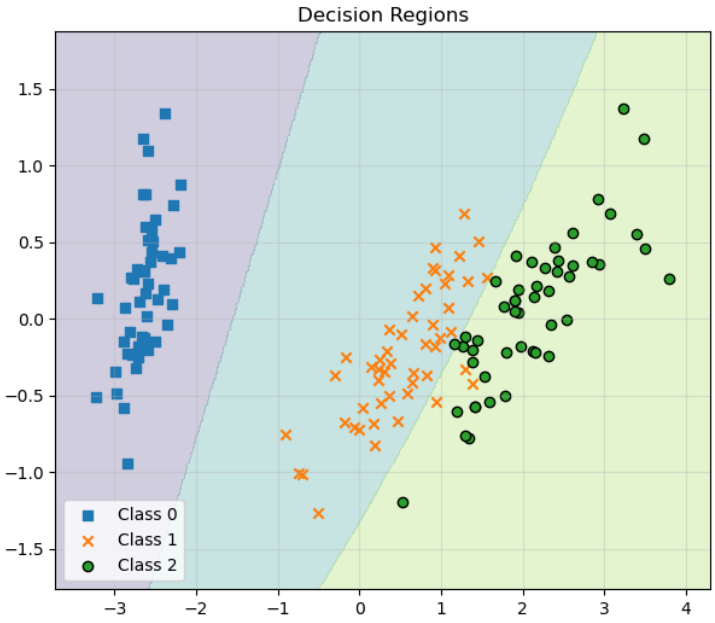
اینجا باید از discriminant function برای حالتی که ماتریس کوواریانس کلاس ها متفاوت است استفاده کنیم. رابطه ی آن به صورت زیر است.

$$g_i(x) = -\frac{۱}{۲}(x-\mu_i)^T\Sigma^{-۱}(x-\mu_i) - \frac{d}{۲}\ln ۲\pi - \frac{۱}{۲}\ln|\Sigma_i| + \ln P(\omega_i) \quad (۷۷)$$

توجه کنید که جمله ی $\frac{d}{۲}\ln ۲\pi$ در تمامی کلاس ها یکسان می شود پس می توانیم در نوشتن کد، از آن صرف نظر کنیم.

۶-۶-۳. بخش c: Decision Regions

ناحیه های تصمیم گیری سه کلاس به صورت زیر می شود.



شکل ۴. ناحیه های تصمیم‌گیری سه کلاس

۶-۷. بخش i: تکرار سوال با داده های کمتر

در اینجا باید بخش های a تا c را با فرض تفاوت ماتریس کوواریانس توزیع ها و همچنین با تقسیم داده ها و استفاده از ۸۰ درصد نمونه ها و در نظر گرفتم هزینه rejection تکرار کنیم.

اول از هر کاری باید داده ها را تقسیم کنیم و ۸۰ درصد مورد نظر را مشخص کنیم در غیر این صورت احتمالا leakage خواهیم داشت.

۶-۷-۱. بخش a: محاسبه ی میانگین و کوواریانس

بردار میانگین و ماتریس کوواریانس هر کلاس به صورت زیر است.

Mean of class ۰: [-۲,٦٦١٧٨٥٧٨ ۰,۲۰٤٥۹۱۸۹]

Mean of class ۱: [۰,٥۲٥۷۰۷٦ -۰,۲٤۰۹۰٦٤۸]

Mean of class ۲: [۲,۱٥۹۱۱٥۷٤ ۰,۰۹۰٥۹٥۰۸]

Covariance of class ۰:

[[۰,۰٤۲۳۲۱٦۷ ۰,۰٤٤۲۹۰٤۳]

[۰,۰٤٤۲۹۰٤۳ ۰,۲۱٥٦٦٥۰٤]]

Covariance of class ۱:

[[۰,۳۳۳۱۰٤٤۸ ۰,۱۹۱۱۸٤۳۷]

[۰,۱۹۱۱۸٤۳۷ ۰,۱۹۸۳۰٤۱٥]]

Covariance of class ۲:

[[۰,٥۱۲۱۹٦۲۱ ۰,۲٦۰۳٤۸۳٥]

[[٠,٢٦٠٣٤٨٣٥ ٠,٢٠٥٠٨٤٠٥]]

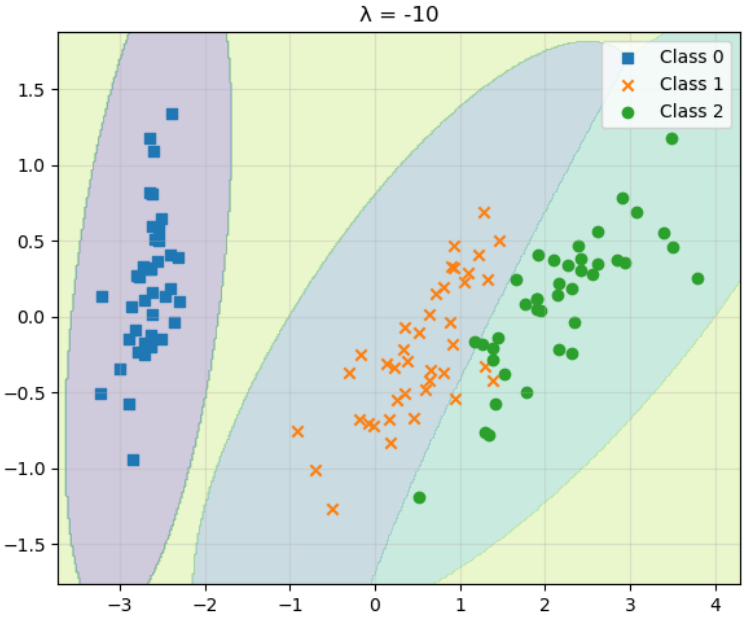
۶-۷-۲. بخش b: Discriminant function

اینجا باید از discriminant function برای حالتی که ماتریس کوواریانس کلاس ها متفاوت است استفاده کنیم (رابطه‌ی (۷۷)). همان طور که صورت مسئله خواسته، هزینه ی rejection در نظر گرفته می شود.

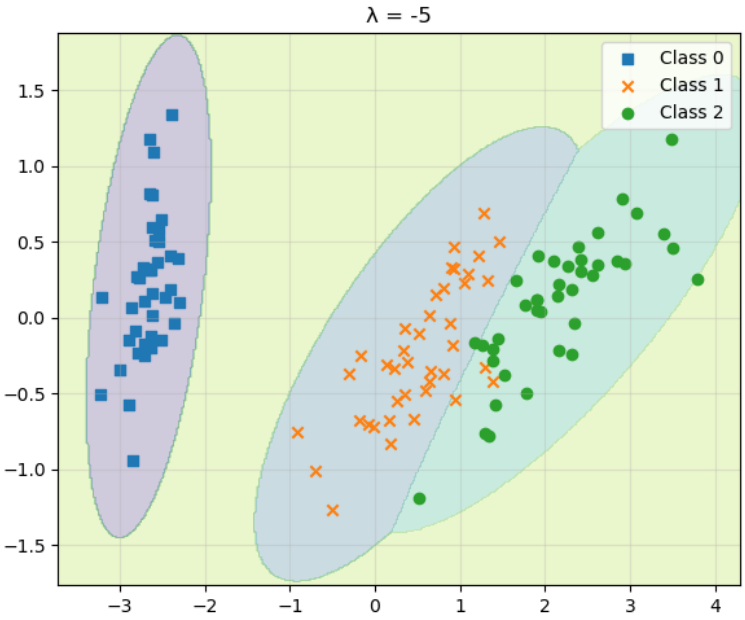
۶-۷-۳. بخش c: Decision Regions

ناحیه های تصمیم گیری سه کلاس به ازای λ های مختلف به صورت زیر می شود. با توجه به اشکال ۵ الی ۱۱،هر چه مقدار لاندا بزرگ تر می شود ناحیه های reject هم بزرگ تر شده و داده هایی که در این ناحیه قرار می گیرند بیشترند. به ازای مقادیر ۱۰- و ۵- برای لاندا، ناحیه های reject بسیار محافظه کارانه هستند. در مقابل به ازای ۲-، ۱- و ۰ نواحی reject بسیار بزرگ شده و بخش زیادی از فضای تصمیم از بین می رود.

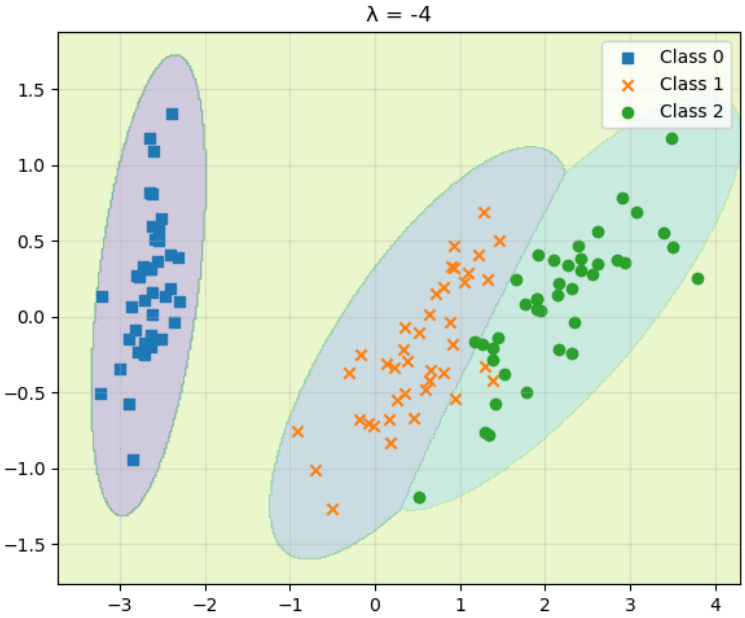
اما به نظر به ازای مقادیر ۴- و ۳- تعادل خوبی بین ناحیه‌ی reject و نواحی تصمیم گیری وجود دارد. اگر قرار بر انتخاب لاندا باشد، بهترین مقدار ها ۴- و ۳- هستند.



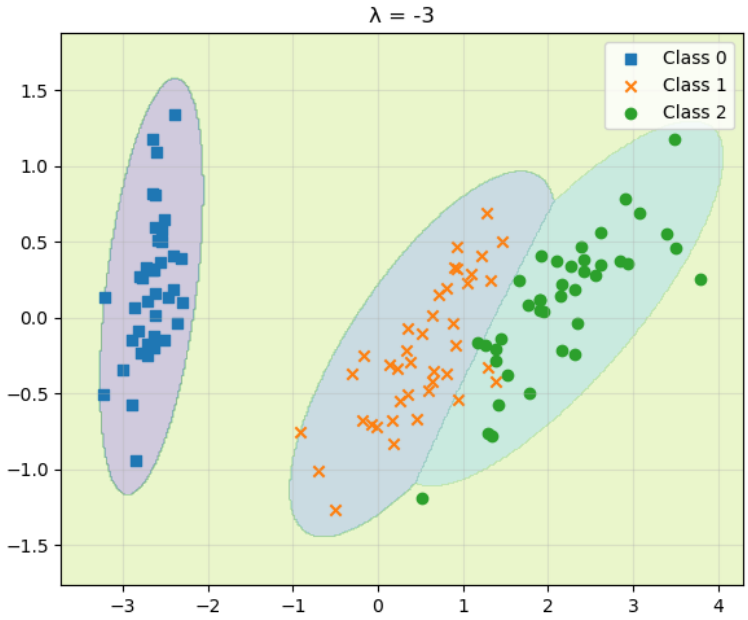
شکل ۵. ناحیه های تصمیم گیری سه کلاس به ازای $\lambda = -10$



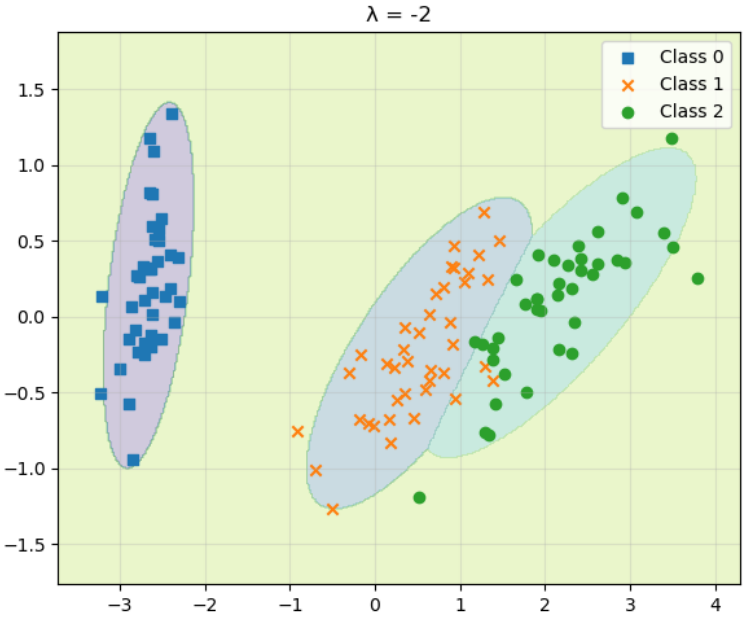
شکل ۶. ناحیه های تصمیم گیری سه کلاس به ازای $\lambda = -5$



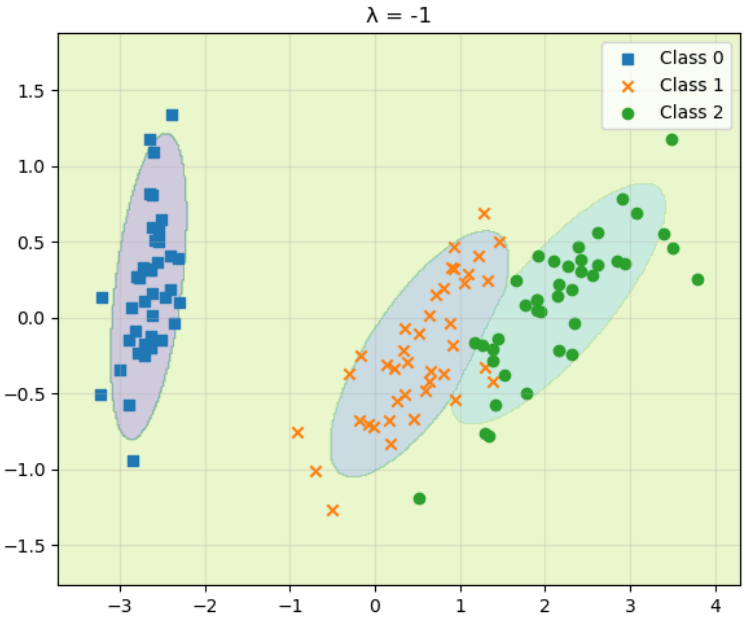
شکل ۷. ناحیه های تصمیم گیری سه کلاس به ازای $\lambda = -4$



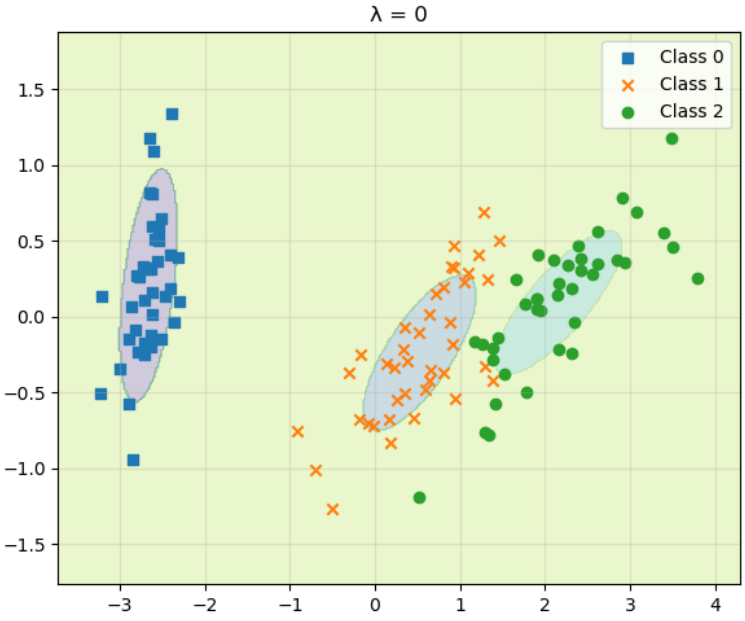
شکل ۸. ناحیه های تصمیم گیری سه کلاس به ازای $\lambda = -3$



شکل ۹. ناحیه های تصمیم گیری سه کلاس به ازای $\lambda = -2$



شکل ۱۰. ناحیه های تصمیم گیری سه کلاس به ازای $\lambda = -1$



شکل ۱۱. ناحیه های تصمیم گیری سه کلاس به ازای $\lambda = 0$

پرسش ۷– Gaussian Naive Bayes

در این تمرین یک مدل Gaussian Naive Bayes را کامل نوشته، آموزش می دهیم و عملکرد آن را بررسی می کنیم. بخش های a تا c این تمرین کاملا کامپیوتری بوده و جنبه پرسش تئوری ندارد که نیاز به توضیح داشته باشد.

۷-۱. بخش d: عملکرد مدل

تمامی مقادیر زیر در فایل evaluation.txt موجود است. در اینجا صرفا برای کامل بودن گزارش دوباره نوشته می شود.

Accuracy: ۰٫۸۳۵

Precision for class ۱: ۰٫۹۵۱

Recall for class ۱: ۰٫۹۷۵

Precision for class ۵: ۰٫۸۷۵

Recall for class ۵: ۰٫۷۷۸

پرسش ۸– Bag of Words — Naive Bayes Classifier

این تمرین درباره‌ی پیاده سازی طبقه بند naive bayes با MLE و MAP می شود. مفاهیم مهمی از سوالات تئوری این تمرین می توان دریافت.

۸-۱. بخش a: Naive Bayes Classifier — MLE

در MLE با احتمال زیر کار می کنیم.

$$p(w|c) = \frac{count_{w,c}}{N_c} \quad (78)$$

به مفهوم تعداد ظاهر شدن هر کلمه در هر کلاس به کل کلمات آن کلاس است.

- دقت مجموعه تست شما چقدر است؟ مقدار دقت روی داده های تست:

Accuracy on baseball test: ۰٫۳۲

Accuracy on hockey test: ۰٫۰٤

Total accuracy: ۰٫۱۸

که این دقت بسیار بد است.

- در نهایت طبقه بند شما چه نتیجه‌ای پیش‌بینی کرد؟

در خروجی زیر می بینیم که بسیاری از پست های بیسبال را به اشتباه هاکی و بسیاری از پست های هاکی را به اشتباه بیسبال طبقه بندی کرده است. که این مورد نشان دهنده ی بایاس این طبقه بند است. علت اصلی این مورد را در سوال بعدی توضیح خواهم داد.

Predictions for baseball_test: ['hockey', 'hockey', 'baseball', 'baseball', 'hockey', 'baseball', 'baseball', 'hockey', 'hockey', 'hockey'] ...
Predictions for hockey_test: ['hockey', 'baseball', 'baseball', 'baseball', 'baseball', 'baseball', 'baseball', 'baseball', 'baseball', 'baseball'] ...

- چرا استفاده از MLE در این شرایط ایده‌ی خوبی نیست؟

۱) اگر در داده های آموزش یکی از کلاس ها، کلمه ای هیچ گاه ظاهر نشود آنگاه MLE احتمال آن کلمه را به شرط آن کلاس صفر در نظر می گیرد.

$$p(w|c) = o \quad (۷۹)$$

در نتیجه برای هر داده ی تستی که شامل آن کلمه ی دیده نشده باشد، خواهیم داشت:

$$p(w|c) = ۰ \Rightarrow \log\big(p(w|x)\big) = -\infty \quad (۸۰)$$

این باعث می شود مدل آن کلاس را رد کند و روی داده های تست دقت بدی داشته باشد.

۲) از آنجا که smoothing در این بخش انجام نشده (سوال نخواسته است) احتمالات تحت تاثیر چند کلمه پرتکرار قرار می گیرد و مدل به سمت یک کلاس متمایل می شود. این دقیقا در نتایج هم دیده شد.

۸-۲. بخش b: Naive Bayes Classifier — MAP with Dirichlet Prior

نتیجه ی محاسبه‌ی arg max (posterior) به صورت زیر است.

$$p(w|c) = \frac{coutn_{w,c} + \beta}{N_c + \beta V} \quad (۸۱)$$

که در آن V تعداد کلمات متمایز (۵۸۲۲) است و β تعداد دیده شدن کلمه در داده های آموزش است.

۸-۳. بخش c: Effect of β on Classification Accuracy

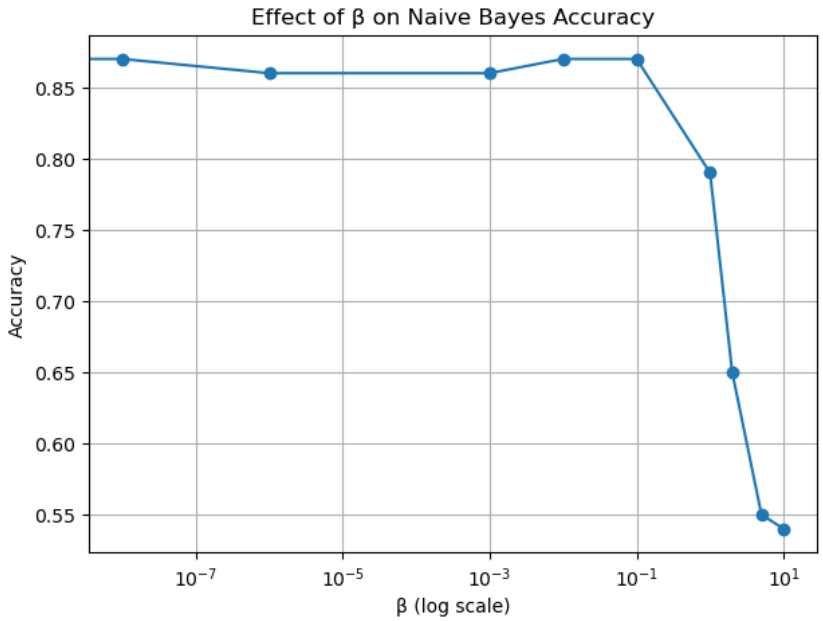
در ابتدا باید گفت که اضافه شدن β به کسر احتمال باعث smoothing می شود. یعنی حالا اگر کلمه ای هیچگاه در داده های آموزش دیده نشود، احتمال آن صفر نیست و ضریبی از β است. در جدول ۱ و شکل ۶ تاثیر مقادیر مختلف β روی دقت مدل را می بینیم. β های کوچک (از ۱e-۰۸ الی ۰) باعث smoothing می شوند. یعنی صرفا احتمال های صفر را اصلاح می کنند و باعث پایداری بیشتر و دقت بیشتر مدل روی داده های تست نسبت به MLE می شوند.

در مقابل β های بزرگ دقت مدل را به شدت کاهش می دهند. در واقع با زیاد شدن β شمارش هر کلمه بیش از حد smooth می شود و باعث می شود تفاوت ها واقعی بین کلاس های بیسبال و هاکی از بین برود. هدف ما این بود که با اضافه شد β مشکل ظاهر نشدن برخی کلمات در آموزش مدل را رفع کنیم. ولی اگر β مقدار زیادی داشته باشد به این مفهوم است که ما خود خواسته احتمال دیده شدن آن کلمه را بیش از حد افزایش دادیم در حالی که این مورد نه هدف ما بوده و نه مطلوب ما. بنابراین β های زیاد قدرت تمایز مدل را کاهش می دهد و دقت افت می کند.

پس β های کوچک مطلوب هستند.

جدول ۱. تاثیر بتا بر دقت مدل

β	Accuracy
۰	۰.۸۶۰۰
۱e-۰۸	۰.۸۷۰۰
۱e-۰۶	۰.۸۶۰۰
۰.۰۰۱	۰.۸۶۰۰
۰.۰۱	۰.۸۷۰۰
۰.۱	۰.۸۷۰۰
۱	۰.۷۹۰۰
۲	۰.۶۵۰۰
۵	۰.۵۵۰۰
۱۰	۰.۵۴۰۰



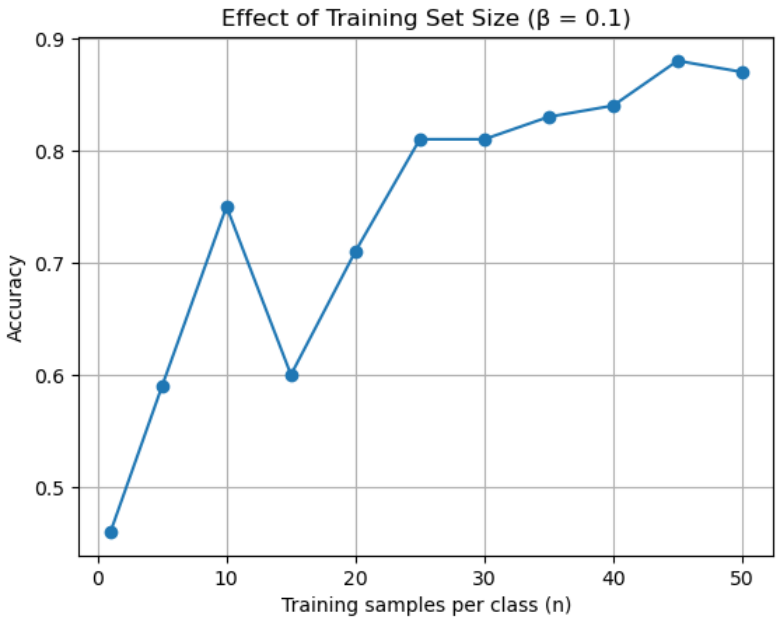
شکل ۱۲. تغییرات دقت نسبت به بتا

۸-۴. بخش d: Effect of Training Set Size

در این بخش تاثیر اندازه‌ی داده‌ی آموزش بر دقت مدل می سنجیم. در $n = ۱$ و $n = ۵$ دقت مدل پایین است زیرا با تعداد کمی پست، تقریبا اطلاعات قابل اعتمادی درباره‌ی توزیع کلمات بدست نمی آوریم. در نتیجه مدل بدون داده‌ی کافی قادر به یادگیری تمایز بین کلاس ها نیست. با افزایش n دقت مدل به مرور افزایش می یابد زیرا مدل تصویر کامل تری از توزیع کلمات در هر کلاس بدست می آورد و توانایی تمایزش بهتر می شود. افتی که در $n = ۵۰$ و $n =$ ۱۵ داریم می تواند به دلیل نویز یا خاص بودن بعضی از پست ها باشد.

n	Accuracy
۱	۰.۴۶۰۰
۵	۰.۵۹۰۰
۱۰	۰.۷۵۰۰
۱۵	۰.۶۰۰۰
۲۰	۰.۷۱۰۰
۲۵	۰.۸۱۰۰
۳۰	۰.۸۱۰۰
۳۵	۰.۸۳۰۰
۴۰	۰.۸۴۰۰
۴۵	۰.۸۸۰۰
۵۰	۰.۸۷۰۰

جدول ۲. تاثیر اندازه‌ی داده‌ی آموزش بر دقت مدل



شکل ۱۳. تاثیر اندازه‌ی داده‌ی آموزش بر دقت مدل

پس در کل هر چه اندازه ی داده های آموزش بیشتر باشد دقت مدل افزایش می یابد اما این افزایش دقت در n های بزرگ آهسته تر رخ می دهد. بهترین دقت هم در اینجا به ازای $n = ۴۵$ رخ داده است.

مراجع

References

[۱] faradars, "naive-bayes-classifier," ۲۳ ۰۱ ۱٤۰٤. [Online]. Available: https://blog.faradars.org/naive-bayes-classifier/#%d۸٪af%d۸٪b۳٪d۸٪aa%d۹٪۸۷-%d۸٪a۸٪d۹٪۸٦٪d۸٪af-%d۸٪a۸٪db%۸c%d۸٪b۲-%d۹٪۸۸-%d۹٪۸٥٪d۸٪af%d۹٪۸٤-%d۸٪a۷٪d۸٪ad%d۸٪aa%d۹٪۸٥٪d۸٪a۷٪d۹٪۸٤٪d۸٪a۷٪d۸٪aa%db%۸c-۱.

[۲] "Gamma distribution," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Gamma_distribution.

[۳] جزوه‌ی درس