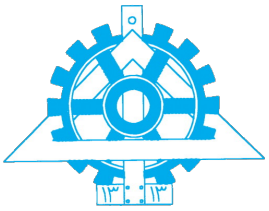
دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر



یادگیری ماشین

تمرین شماره ۱

نام و نام خانوادگی

مرتضی ملکی‌نژاد شوشتری

شماره دانشجویی

۸۱۰۱۰۴۲۵۶

۶ آذر ۱۴۰۴

فهرست مطالب

فهرست تصاویر

فهرست جداول

چکیده

هدف این تمرین آشنایی بیشتر با روش‌های کلاسیک PDF Estimation و Classification با تمرکز بر دو روش MLE و MAP می‌باشد. به غیر از سوال اول که به Minimum Risk Classification می‌پردازد، باقی سوالات عموما با ایده‌های مربوط به Likelihood و بعضا (مثل سوال ۵) به ارتباط بین Prior و Posterior می‌پردازند.

سوال ۱

در Minimum Risk Classification تابع ریسک بصورت زیر تعریف می‌شود:

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|x)$$

در این‌جا حالت رد [1] عملا یک کلاس دیگر در کلاس‌بندی حساب می‌شود؛ با این تفاوت که احتمال تعلق به این کلاس برای هر برداری برابر صفر می‌باشد. ریسک رد برابر است با:

$$R(\alpha_{c+1}|x) = \lambda_r \sum_{j=1}^{c} P(\omega_j|x) = \lambda_r$$

ریسک انتخاب، برای مثال انتخاب کلاس i برابر است با:

$$R(\alpha_i|x) = \lambda_s(1 - P(\omega_i|x))$$

واضح است که بین کلاس‌ها درصورتی که $P(\omega_i|x)$ بیشینه باشد کمترین ریسک حاصل می‌شود، پس اگر ورودی رد نشود، باید کلاس با بالاترین احتمال را انتخاب کرد ($\forall j\ P(\omega_i|x) \geq P(\omega_j|x)$) برای کلاس رد نیز،با مقایسه ریسک انتخاب با ریسک رد داریم:

$$R(\alpha_{c+1}|x) \geq R(\alpha_i|x) \Rightarrow \lambda_r \geq \lambda_s(1 - P(\omega_i|x)) \Rightarrow -P(\omega_i|x) \leq \frac{\lambda_r}{\lambda_s} - 1 \Rightarrow$$

$$P(\omega_i|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$$

همه مراحل برگشت‌پذیر هستند و می‌توان به این رسید که ریسک انتخاب از ریسک رد کمتر است اگر و تنها اگر $P(\omega_i|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$ باشد.

[1] Reject

سوال ۲

۱.۳ بخش a

با استفاده از قانون بیز داریم:

$$P(Y=c|X_1=0,X_2=0)=\frac{P(X_1=0,X_2=0|Y=c)P(Y=c)}{\sum_{i=0}^{2}P(X_1=0,X_2=0|Y=i))P(Y=i)}$$

Naive Bayes می‌باشد، فرض شده است که فیچرها از هم مستقل هستند. پس می‌توان گفت:

$$P(X_1=0,X_2=0|Y=c)=P(X_2=0|Y=c)P(X_1=0|Y=c)$$

پس داریم (چون اعداد ثابت در صورت و مخرج ساده می‌شوند از نوشتن آن‌ها صرف نظر شده است):

$$P(X_1=0,X_2=0|Y=0)P(Y=0)=P(X_2=0|Y=0)P(X_1=0|Y=0)P(Y=0)$$

$$=e^{\frac{-\mu_c^2}{2}}*(1-\theta_c)*prior_c=e^{-0.5}*0.5*0.5\approx 0.1516$$

$$P(X_1=0,X_2=0|Y=1)P(Y=1)=1*0.25*0.25=0.0625$$

$$P(X_1=0,X_2=0|Y=2)P(Y=2)=e^{-0.5}*0.5*0.25\approx 0.0758$$

$$\Rightarrow \sum_{i=0}^{2}P(X_1=0,X_2=0|Y=i)P(Y=i)\approx 0.2899$$

$$\Rightarrow \begin{cases} P(Y=0|X_1=0,X_2=0)\approx \frac{1516}{2899}\approx 0.52 \\ P(Y=1|X_1=0,X_2=0)\approx \frac{625}{2899}\approx 0.22 \\ P(Y=2|X_1=0,X_2=0)\approx \frac{758}{2899}\approx 0.26 \end{cases}$$

پس بردار خواسته شده برابر با مقدار زیر است:

$$\begin{pmatrix} 0.52 \\ 0.22 \\ 0.26 \end{pmatrix}$$

۲.۳ بخش b

طبق قانون بیز داریم:

$$P_{Y|X_1} = \frac{P(X_1|Y) * P(Y)}{P(X_1)}$$

$$\Rightarrow P(Y = c|X_1 = 0) = \frac{P(X_1 = 0|Y = c)P(Y = c)}{\sum_{i=0}^{2} P(X_1 = 0|Y = i)P(Y = i)}$$

$$h(c) = P(X_1 = 0|Y = c)P(Y = c) = (1 - \theta_c) * prior_c$$

$$\Rightarrow \begin{cases} h(0) = 0.5 * 0.5 = \frac{4}{16} \\ h(1) = 0.25 * 0.25 = \frac{1}{16} \\ h(2) = 0.25 * 0.5 = \frac{2}{16} \end{cases} \Rightarrow \begin{cases} P(Y = 0|X_2 = 0) = \frac{h(0)}{\sum_{i=0}^{2} h(i)} = \frac{4}{7} \\ P(Y = 1|X_2 = 0) = \frac{h(1)}{\sum_{i=0}^{2} h(i)} = \frac{1}{7} \\ P(Y = 2|X_2 = 0) = \frac{h(2)}{\sum_{i=0}^{2} h(i)} = \frac{2}{7} \end{cases}$$

و بردار خواسته شده برابر خواهد بود با:

$$\begin{pmatrix} \frac{4}{7} \\ \frac{1}{7} \\ \frac{2}{7} \end{pmatrix}$$

۳.۳ بخش c

مشابه قسمت‌های قبل عمل می‌کنیم (در اینجا منظور از $P(X_2)$ مقدار تابع PDF در آن نقطه است و نه احتمال آن نقطه)

$$h(c) = P(X_2 = 1|Y = c)P(Y = c) = \frac{1}{\sigma_c\sqrt{2\pi}} * e^{(\frac{-1}{2}*(\frac{1-\mu_c}{\sigma_c})^2)} * prior_c$$

چون σ_c در همه موارد برابر 1 است می‌توان آن را جایگذاری کرد.

$$K = \frac{1}{\sqrt{2\pi}}$$

$$h(c) = K*e^{\frac{-(1-\mu_c)^2}{2}} *prior_c \Rightarrow \begin{cases} h(0) = K * e^{-2} * 0.5 \approx 0.068K \\ h(1) = K * e^{-0.5} * 0.25 \approx 0.152K \\ h(0) = K * e^{0} * 0.25 = 0.25K \end{cases} \Rightarrow$$

$$\sum_{i=0}^{2} h(i) = 0.47K$$

$$\Rightarrow \begin{cases} P(Y = 0|X_2 = 1) = \frac{h(0)}{\sum_{i=0}^{2} h(i)} = \frac{68}{470} \\ P(Y = 1|X_2 = 1) = \frac{h(1)}{\sum_{i=0}^{2} h(i)} = \frac{152}{470} \\ P(Y = 2|X_2 = 1) = \frac{h(2)}{\sum_{i=0}^{2} h(i)} = \frac{250}{470} \end{cases}$$

و بردار خواسته شده برابر خواهد بود با:

$$\begin{pmatrix} \frac{68}{470} \\ \frac{152}{470} \\ \frac{250}{470} \end{pmatrix}$$

۴.۳ بخش d

برای قسمت c می‌توان دید که احتمال Y رابطه مستقیمی دارد با اینکه چقدر میانگین توزیع $P(X_2|Y)$ که یک توزیع نمایی است به 0 نزدیک‌تر است. یعنی می‌توان رابطه مستقیم بین $P(Y|X_2)$ و $P(X_2|Y)$ را مشاهده کرد. برای مثال برای $Y = 0$ با این که احتمال prior برابر ۵.۰ است،ولی چون توزیع $X_2|Y = 0$ یک توزیع نمایی با میانگین -1 است، احتمال آن کمتر است. در بخش b این اتفاق با شدت کمتری می‌افتد. زیرا مقادیر θ نسبت به مقادیر prior تغییرات بسیار جدی‌ای ندارند (به نسبت توزیع نمایی که با فاصله گرفتن چند انحراف معیار از میانگین،

تابع pdf کاهش زیادی خواهد داشت.) برای همین می‌توان دید که احتمال‌های بدست آمده به نسبت به احتمال prior نزدیک ترند.

سوال ۳

مدل فرض شده یک مدل Naive Bayes است که توزیع کلمات را یک توزیع چندجمله‌ای فرض کرده است. [۴] تابع توزیع احتمال برای توزیع چندجمله‌ای برابر زیر است [۵] :

$$X \sim Multinomial(n, p_1, p_2, ..., p_k)$$

$$\Rightarrow P_X(x_1, x_2, ..., x_k) = \frac{n!}{x_1!x_2!x_3!...x_k!} p_1^{x_1} p_2^{x_2} ... p_k^{x_k}$$

برای محاسبه احتمال خواسته شده با استفاده از قانون بیز داریم:

$$P(Y = 1|x) = \frac{P(x|Y = 1) * P(Y = 1)}{\sum_{i=0}^{1} P(x|Y = i)P(Y = i)}$$

برای محاسبه $P(x|Y = 1)$ می‌توان از تابع توزیع احتمال توزیع چندجمله‌ای استفاده کرد.

پس:

$$P(x|Y = 1) = \frac{n!}{c_1!c_2!c_3!...c_d!} \prod_{i=1}^{d} (P_{i|y=1})^{c_i}$$

پس داریم:

$$P(Y = 1|x) = \frac{P(x|Y = 1)P(Y = 1)}{\sum_{i=0}^{1} P(x|Y = i)P(Y = i)}$$

$$= \frac{P(Y = 1) * \prod_{i=1}^{d} (P_{i|y=1})^{c_i}}{P(Y = 1) * \prod_{i=1}^{d} (P_{i|y=1})^{c_i} + P(Y = 0) * \prod_{i=1}^{d} (P_{i|y=0})^{c_i}}$$

$$= \frac{P_y * \prod_{i=1}^{d} (P_{i|y=1})^{c_i}}{P_y * \prod_{i=1}^{d} (P_{i|y=1})^{c_i} + (1 - P_y) * \prod_{i=1}^{d} (P_{i|y=0})^{c_i}}$$

۱.۴ بخش b

برای محاسبه مرز تصمیم‌گیری می‌توان مقدار Posterior ها را باهم برابر قرار داد، چون مخرج ها باهم برابر هستند،می‌توان فقط صورت‌ها را باهم برابر قرار داد:

$$P(Y=0|x) = P(Y=1|x) \Rightarrow P_y \prod_{i=1}^{d} (P_{i|y=1})^{c_i} = (1-P_y) * \prod_{i=1}^{d} (P_{i|y=0})^{c_i}$$

چون می‌خواهیم درنهایت به یک ترکیب خطی از c_i ها برسیم،برای تبدیل ضرب به جمع می‌توان از آن لگاریتم گرفت (چون لگاریتم یک تابع یکنوا است،خطی بودن را تغییر نمی‌دهد)

$$log(P_y) + \sum_{i=1}^{d} c_i log(P_{i|Y=1}) = log(1-P_y) + \sum_{i=1}^{d} c_i log(P_{i|Y=0}) \Rightarrow log(\frac{P_y}{1-P_y}) +$$

$$\sum_{i=1}^{d} c_i log(\frac{P_{i|Y=1}}{P_{i|Y=0}}) = 0$$

از آنجایی که در مدل مقادیر P_y و $P_i|Y=0$ و $P_i|Y=1$ از پیش تعریف شده‌اند و مقادیر ثابتی هستند، می‌توان دید که مرز تصمیم درواقع برابر مقدار زیر است:

$$b + \sum_{i=1}^{d} w_i c_i = 0$$

$$b = log(\frac{P_y}{1-P_y}), w = log(\frac{P_{i|Y=1}}{P_{i|Y=0}})$$

و واضح است که این یک خط است.

۲.۴ بخش c

صورت و مخرج احتمال را به صورت تقسیم می‌کنیم:

$$P(Y=1|x) = \frac{1}{1 + \frac{(1-P_y) * \prod_{i=1}^{d} (P_{i|y=0})^{c_i}}{P_y * \prod_{i=1}^{d} (P_{i|y=1})^{c_i}}}$$

از کسر حاصل در مخرج می‌توان لگاریتم گرفت و e را به توان آن رساند:

$$P(Y=1|x) = \frac{1}{1 + e^{log(\frac{1-P_y}{P_y}) + \sum_{i=1}^{d} c_i log(\frac{P_{i|y=0}}{P_{i|y=1}})}}$$

مشابه قسمت b ضرایب ثابت هستند پس اگر θ_0 را برابر $log\frac{P_y}{1-P_y}$ بگیریم و θ_i را برابر $log(\frac{P_{i|y=1}}{P_{i|y=0}})$ بگیریم داریم:

$$P(Y=1|c) = \frac{1}{1+e^{-(\theta_0+\sum_{i=1}^{d} c_i\theta_i)}} = \frac{1}{1+e^{-(\theta_0+\theta^T c)}}$$

سوال ۴

۱.۵ بخش a

چون داریم بیشینه می‌گیریم، هیچگاه نباید $P(data|\theta)$ صفر بشود (زیرا کمینه می‌شود) پس می‌توان فرض کرد که همه داده‌ها در ناحیه لوزی شکل هستند:

$$\theta \geq ||x||_1 \; for \; all \; x_i$$

$$\hat{\theta} = argmax P(data|\theta) = argmax(\frac{1}{2\theta^2})^N$$

واضح است که $(\frac{1}{2\theta^2})^N$ وقتی بیشینه است که θ کمترین مقدار ممکن را داشته باشد پس یعنی پاسخ MLE برابر با کمترین θ ای است که در شرط اول صدق کند پس:

$$\hat{\theta} = min(||x||_i) = min(|x_{i_1}| + |x_{i_2}|) \forall i$$

۲.۵ بخش b

تابع چگالی احتمال توزیع گاما برابر است با:

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}}$$

حال MLE به دنبال مقادیری برای α و β است که حاصل ضرب f های آن‌ها بیشینه شود.

برای راحتی می‌توان اینجا لگاریتم گرفت و سپس جمع این موارد را مورد بررسی قرار داد:

$$log(L) = (\alpha - 1)log(x) + \frac{-x}{\beta} log(e) - \alpha log(\beta) - log(\Gamma(\alpha))$$

$$= (\alpha - 1)log(x) + \frac{-x}{\beta} - \alpha log(\beta) - log(\Gamma(\alpha))$$

با مشتق گرفتن نسبت به β داریم:

$$\frac{x}{\beta^2} - \alpha \frac{1}{\beta} = 0 \Rightarrow x = \alpha\beta \Rightarrow \beta = \frac{x}{\alpha}$$

در حل برای راحتی برای یک مقدار x حل شد. چون در اینجا x ها iid هستند، مقدار likelihood آن‌ها باهم جمع می‌شود و چون مشتق می‌تواند وارد جمع شود، می‌توان این جمع را در مرحله آخر انجام داد پس داریم:

$$\sum_{i=1}^{n}(\frac{x}{\beta^2}-\alpha\frac{1}{\beta})=0\Rightarrow\sum_{i=1}^{n}x_i=n\alpha\beta\Rightarrow\beta=\frac{\bar{x}}{\alpha}$$

سوال ۵

۱.۶ بخش a

توزیع دیریکله درواقع یک توزیع با PDF زیر است: [۳]

$$\frac{1}{B(\alpha)}\prod_{i=1}^{k} x_i^{\alpha_i-1}, B(\alpha)=\frac{\prod_{i=1}^{k}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k}\alpha_i)}$$

برای اثبات conjugate prior باید درواقع ثابت کنیم که اگر توزیع prior یک توزیع دیریکله باشد و توزیع likelihood یک توزیع چندجمله‌ای باشد، توزیع posterior یک توزیع دیریکله است. [۲]

$$\text{prior: } P(p|\alpha)=\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma\alpha_k}\prod_{i=1}^{k} {p_k}^{\alpha_i-1}$$

$$\text{likelihood: } P(x|p)=\frac{N!}{\prod_{i=1}^{k} x_k!}\prod_{i=1}^{k} p_i^{x_i}$$

$$\Rightarrow \text{posterior: } P(p|x,\alpha)=\frac{P(x|p)P(p|\alpha)}{P(x|\alpha)}=\frac{P(x|p)P(p|\alpha)}{\int P(x|p)P(p|\alpha)dp}$$

$$P(x|p)P(p|\alpha)=\frac{N!}{\prod_{i=1}^{k} x_k!}\prod_{i=1}^{k} p_i^{x_i}*\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma\alpha_i}\prod_{i=1}^{k} {p_k}^{\alpha_i-1}=\frac{N!\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k} x_k!\prod_{i=1}^{k}\Gamma\alpha_k}\prod_{i=1}^{k} p_i^{x_i+\alpha_i-1}$$

چون قرار است مقدار حاصل به روی انتگرال نسبت به p تقسیم شود، می‌توان عواملی که در آن p نیست را درنظر نگرفت:

$$P(p|x,\alpha) = \frac{\prod_{i=1}^{k} p_i^{x_i+\alpha_1-1}}{\int \prod_{i=1}^{k} p_i^{x_i+\alpha_k-1} dp}$$

چون شیوه محاسبه این انتگرال ساده نبود از یکی از سایتهای محاسبه [۱] [۲] کمک گرفته شد که به خروجی زیر رسید:

$$\int \prod_{i=1}^{k} p_i^{x_i+\alpha_k-1} dp = B(\alpha+x) \Rightarrow P(p|x) = \frac{1}{B(\alpha+x)} * \prod_{i=1}^{k} p_i^{x_i+\alpha_i-1} = Dir(\alpha+x)$$

پس می‌توان دید اگر likelihood از توزیع چندجمله‌ای و prior از توزیع دیریکله باشد توزیع posterior از نوع دیریکله است.

۲.۶ بخش b

در اثبات بخش قبل عملا ثابت شد که توزیع Posterior یک توزیع دیریکله است که نقاط تمرکز آن باتوجه به داده شیفت پیدا کرده‌اند. باتوجه به این‌که مقادیر اولیه α برابر با $u_i = i$ بوده‌اند و مقادیر مشاهده‌شده در داده برابر n_i بوده است، توزیع Posterior یک توزیع دیریکله با پارامتر $\alpha_i = u_i + n_i$ می‌باشد. حال باید دید با کدام مقادیر p این مقدار بیشینه می‌شود. دوباره چون مشتق گرفتن از این توزیع کار ساده‌ای نیست از منابع خارجی کمک گرفته شد. [۱] باتوجه به نتیجه می‌توان دید که احتمال وقتی بیشینه می‌شود که:

$$p_i = \frac{\alpha_i - 1}{\sum_{j=1}^{k} \alpha_j - k} = \frac{n_i + i - 1}{1000 + 7 * 3} = \frac{n_i + i - 1}{1021}$$

[۲] توزیع دیریکله در درس و کتاب rice نبود و جستجو برای یافتن پاسخ این انتگرال به جایی نرسید. برای محاسبه انتگرال و مشتق این مقدار حدود دوساعت زمان صرف شد و پس از به نتیجه نرسیدن صرفا برای محاسبه انتگرال (و مشتق در قسمت بعد) از منابع خارجی استفاده شد که cite شد.

سوال ۶

۱.۷ بخش a

توزیع نرمال برای k متغیر بصورت زیر است:

$$f = (2\pi)^{-(\frac{k}{2})} * \frac{1}{\sqrt{|\Sigma|}} * e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

$\mu =$ matrix of means of dimension k * 1)

$\Sigma =$ Covariance matrix of dimension k * k

پس ابتدا باید ماتریس میانگین و کواریانس را بدست آورد که محاسبه شده‌اند. با توجه به نتایج برای کلاس اول ($y = 0$) داریم:

$$\mu \approx \begin{pmatrix} -2.64 \\ 0.19 \end{pmatrix}, \Sigma \approx \begin{pmatrix} 0.05 & 0.05 \\ 0.05 & 0.21 \end{pmatrix} \Rightarrow |\Sigma| \approx 0.008 \Rightarrow \Sigma^{-1} \approx \begin{pmatrix} 26.25 & -6.25 \\ -6.25 & 6.25 \end{pmatrix}$$

و برای کلاس دوم خواهیم داشت:

$$\mu \approx \begin{pmatrix} 0.53 \\ -0.25 \end{pmatrix}, \Sigma \approx \begin{pmatrix} 0.35 & 0.19 \\ 0.19 & 0.18 \end{pmatrix} \Rightarrow |\Sigma| \approx 0.027 \Rightarrow \Sigma^{-1} \approx \begin{pmatrix} 6.69 & -7.06 \\ -7.06 & 13.01 \end{pmatrix}$$

و برای کلاس سوم داریم:

$$\mu \approx \begin{pmatrix} 2.1 \\ 0.05 \end{pmatrix}, \Sigma \approx \begin{pmatrix} 0.48 & 0.27 \\ 0.27 & 0.22 \end{pmatrix} \Rightarrow |\Sigma| \approx 0.038 \Rightarrow \Sigma^{-1} \approx \begin{pmatrix} 5.92 & -7.02 \\ -7.02 & 12.66 \end{pmatrix}$$

که باید در رابطه زیر جایگذاری شوند

$$P_X(x) = \frac{1}{2\pi * |\Sigma|} * e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

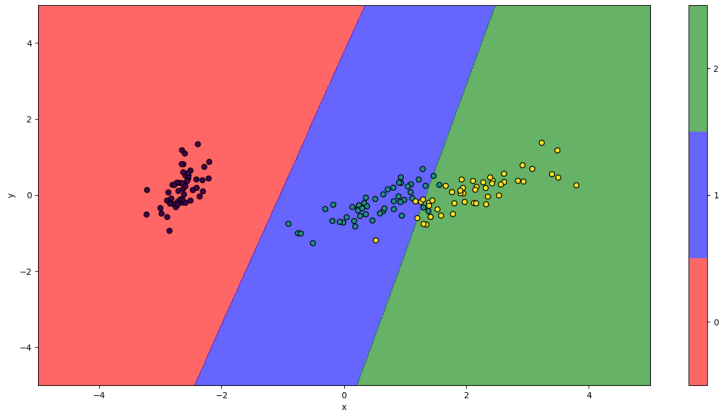
چون هدف بدست آوردن پارامترهای توزیع است این مقدار از محاسبه کافیست.

۲.۷ بخش b

برای محاسبه تابع تفکیک [3] چون احتمال prior تعیین نشده می‌توان آن را برای کلاس ها یکسان در نظر گرفت و صرفا likelihood را حساب کرد.

۳.۷ بخش c

باتوجه به اینکه ماتریس کواریانس یکسان گرفته شد، نواحی بصورت خطی جداپذیر هستند.



شکل ۱: کانتورها باتوجه به Discriminant Function

۴.۷ بخش d

باید دو تابع PDF را برابر هم قرار داد. در قسمت‌های قبل به دلیل اینکه محاسبه خیلی سخت نبود،ماتریس‌های کواریانس جداگانه حساب شد ولی در اینجا با فرض یکی بودن ماتریس کواریانس حل می‌شود زیرا در غیر اینصورت ناحیه خطی نمی‌شود (در نمودار ها هم فرض یکی بودن نشان داده شده و هم فرض متفاوت بودن):

$$m_1 = x - \mu_1, m_2 = x - \mu_2 \; (x-\mu_1)^T\Sigma^{-1}(x-\mu_1) = (x-\mu_2)^T\Sigma^{-1}(x-\mu_2)$$

با باز کردن حاصل داریم:

$$(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) = x^T\Sigma^{-1}x - \mu_1^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_1 + \mu_1^T\Sigma^{-1}\mu_1$$

بخش‌های این معادله هرکدام عدد هستند می‌توان بجای $x^T\Sigma^{-1}\mu_1$ ترانهاد آن را قرار داد.
همچنین چون Σ یک ماتریس متقارن است، ترانهاد آن با خودش برابر است پس داریم:

$$x^T\Sigma^{-1}\mu_1 = (x^T\Sigma^{-1}\mu_1)^T = \mu_1^T\Sigma^{-1}x$$

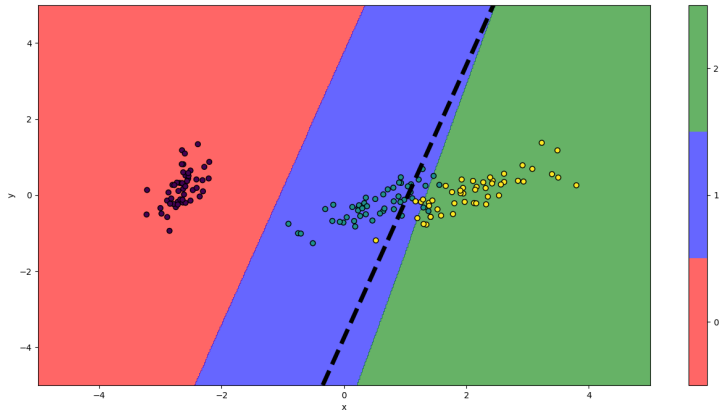
پس در فرمول ناحیه تصمیم داریم:

[3] Discriminant Function

$$x^T\Sigma^{-1}x - 2\mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1 = x^T\Sigma^{-1}x - 2\mu_2^T\Sigma^{-1}x + \mu_2^T\Sigma^{-1}\mu_2$$
$$\Rightarrow \mu_2^T\Sigma^{-1}\mu_2 + 2\mu_1^T\Sigma^{-1}x = \mu_1^T\Sigma^{-1}\mu_1 + 2\mu_2^T\Sigma^{-1}x \Rightarrow 2(\mu_2^T\Sigma^{-1} - \mu_1^T\Sigma^{-1})x = \mu_2^T\Sigma^{-1}\mu_2 - \mu_1^T\Sigma^{-1}\mu_1$$

در اینجا می‌توان جایگذاری کرد و سپس با محاسبه معکوس ماتریس ضریب x جواب را حساب کرد:

$$\Sigma^{-1} = \begin{pmatrix} 29.49 & -7.6 \\ -7.6 & 6.64 \end{pmatrix}, \mu_1 = \begin{pmatrix} -2.65 \\ 0.19 \end{pmatrix}, \mu_2 \begin{pmatrix} 0.53 \\ -0.25 \end{pmatrix}$$
$$\begin{pmatrix} 96.98 & -27.02 \end{pmatrix} x = -101.53 \Rightarrow 96.98x_1 - 27.02x_2 = 101.53$$



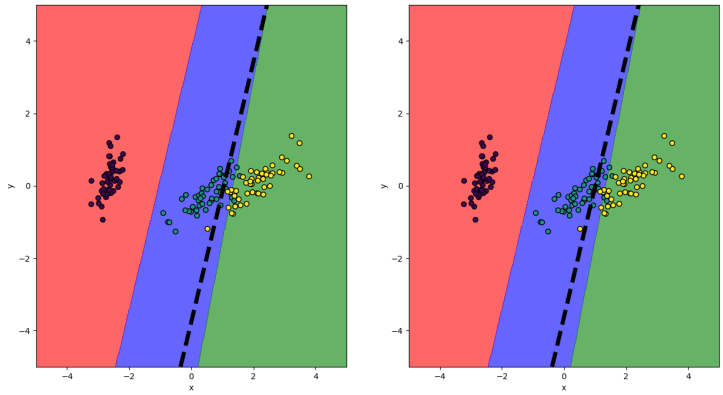
شکل ۲: به دلیل ساده‌سازی در محاسبات، خط مرز تصمیم منطبق نیست اما کلیت آن واضح است.

۵.۷ بخش e

در اینصورت در معادله Discriminant Function بعد از گرفتن لگاریتم، یک طرف معادله با $log(\frac{p_1}{p_2})$ جمع می‌شود و عملا درنهایت مرز تصمیم به سمت کلاسی که احتمال prior بیشتری دارد شیفت پیدا می‌کند. از آنجایی که شیب خط به این مقدار وابسته نیست (زیرا در ضریب x نمود پیدا نمی‌کند) صرفا عرض از مبدا مرز تصمیم فرق می‌کند و شیب خط فرقی نمی‌کند. همچنین چون این ضریب در چیزی ضرب نمی‌شود،عملا تاثیر کمی دارد.

برای مثال اگر داشته باشیم $p_1 = 1000p_2$ معادله نهایی برابر خواهد شد با:

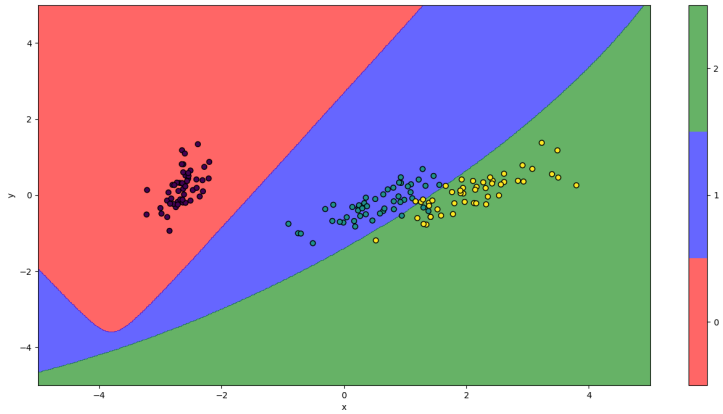
$$96.98x_1 - 27.02x_2 = 101.53 - 0.5ln(1000) = 98.07$$



شکل ۳: می‌توان دید که در عکس سمت راست، مرز تصمیم کمی بیش‌تر به سمت چپ رفته است.

۶.۷ بخش f

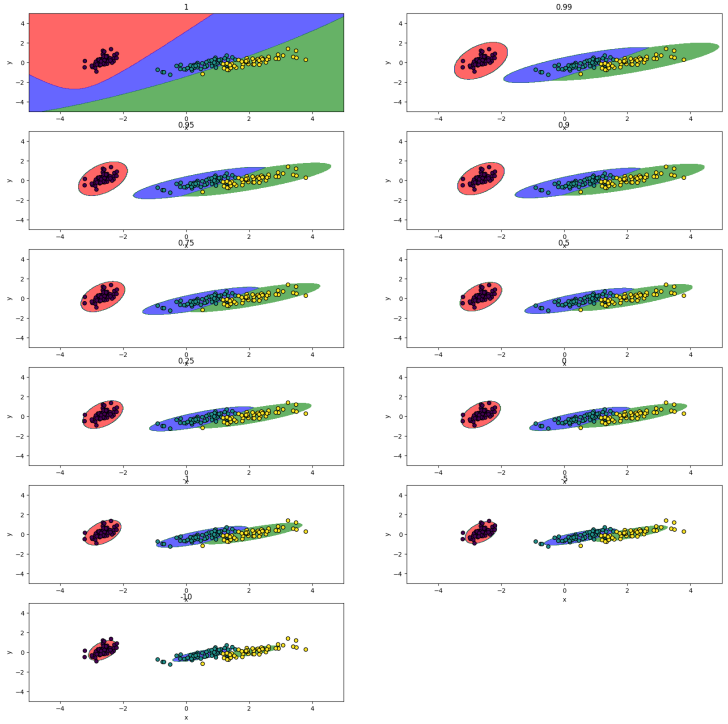
تنها برای بخش d برای خطی بودن نیاز به فرض یکی بودن کواریانس بود. در باقی موارد یکی بودن و نبودن تفاوتی در نمایش داده ندارد (مرز تصمیم خطی نمی‌شود صرفا)



شکل ۴: مهم ترین تغییری که با استفاده مجزا از کواریانس رخ می‌دهد، خطی نبودن مرز تصمیم است. به طور دقیق تر، مرز تصمیم از خط به سطح مشترک دو تابع گاوسی تبدیل می‌شود.

۷.۷ بخش i

کافیست تا Discriminant Function مربوط به Rejection محاسبه شود. در سوال اول این حساب شد که مقدار آن درصورت برابر بودن ریسک همه کلاس‌ها برابر است با: $1 - \frac{\lambda_r}{\lambda_s} = 1 - \lambda_r$
عملا یک بیان دیگر این مسئله این است که اگر احتمال از حدی پایین‌تر است (از $1 - \lambda_r$) و مدل با اطمینان بالایی خروجی نمی‌دهد، خروجی ندهد!



شکل ۵: با کم شدن خطای رد، نواحی بیشتری بدون تصمیم می‌مانند. البته چون در اینجا بجای احتمال تابع چگالی احتمال موجود است، حتی می‌توان مقادیر منفی را بررسی کرد زیرا PDF ممکن است بیشتر از یک باشد.

با توجه به عکس به نظر انتخاب ریسک رد بین ۹.۰ تا ۹۹.۰ انتخابی منطقی باشد.

سوال ۷

چون Naive Bayes است، فرض شده فیچرها مستقل هستند. برای هر ورودی چک می‌شود که کدام کلاس احتمال $likelihood * prior$ بیشتری را نتیجه می‌دهد و سپس بهترین کلاس انتخاب می‌شود.

برای محاسبه متریک‌ها ابتدا باتوجه به نوع متریک، ماسک انتخاب می‌شود که نشان می‌دهد کدام داده‌ها روی متریک تاثیر گذارند، سپس نسبت True positive به کل داده سنجیده می‌شود.

موارد خواسته شده برابرند با:

جدول ۱: دقت مدل

Fraction of test samples classified correctly	0.835
Class 1 Precision	0.951
Class 1 Recall	0.975
Class 5 Precision	0.875
Class 5 Recall	0.778

سوال ۸

۱.۹ بخش a

توزیع کلمات بصورت چندجمله‌ای درنظر گرفته شد. برای تخمین MLE توزیع چندجمله‌ای، یک شهود این است که احتمال هر فیچر متناسب باشد با تعداد آن داده که بصورتی نرمالایز شده باشد که جمع احتمالات یک باشد. برای اثبات این موضوع می‌توان ثابت کرد که اگر x_i از x_j بزرگ‌تر باشد باید p_i از p_j بزرگ‌تر باشد.

$$p_i < p_j, x_i > x_j, c = x_j - x_i$$

$$p_i^{x_i} p_j^{x_j} = (p_i)^{x_j+c}(p_j)^{x_j} = p_i^c p_i^{x_j} p_j^{x_j} < p_j^c p_i^{x_j} p_j^{x_j} \Rightarrow p_i^{x_i} p_j^{x_j} < p_j^{x_j+c} p_i^{x_j} \Rightarrow p_i^{x_i} p_j^{x_j} < p_j^{x_i} p_i^{x_j}$$

یعنی بدون تغییر در مقدار $p_i + p_j$ اگر نسبت p ها متناظر با نسبت x ها نباشد، مقدار likelihood بیشینه نخواهد شد (می‌توان با یک ترکیب دیگر مقدار بهتری یافت.) پس می‌توان دید که نسبت p ها باید متناظر با x باشد. البته اثبات ارائه شده تنها برای نسبت دومتغیر است، ولی با وارد شدن چند متغیر نیز اثبات همچنان کار می‌کند (صرفا کافیست که x_i از همه متغیر های دیگر بزرگ‌تر باشد.)

پس عملا تخمین MLE برای توزیع چندجمله این می‌باشد که نسبت p_i ها برابر نسبت x_i ها باشد. همچنین چون ترم اول تابع چگالی احتمال توزیع چندجمله‌ای نسبت به p ثابت است در کلاس‌های متفاوت برابر است و می‌توان از آن صرف نظر کرد.

۱.۱.۹ سوالات

۱. روی مجموعه داده تست ۸۵ مورد درست حدس زده شدند.

۲. درواقع classifier از روی likelihood احتمال رخ دادن داده درصورتی که در هر کلاس باشد را حساب می‌کند و کلاسی که بیشینه آن را می‌دهد خروجی می‌دهد.

۳. چون در این‌جا بسیاری از فیچرها صفر هستند و عملکرد MLE را تحت تاثیر قرار می‌دهند.

MLE در حالتی که همه فیچرها دخیل باشند بهترین عملکرد را دارد. ولی اگر بسیاری از فیچرها تاثیر نداشته باشند، نتایج PDF ها می‌تواند نزدیک به هم باشند و دچار اشتباه شود. هرچند به نسبت خروجی قابل قبول است.

۲.۹ بخش b

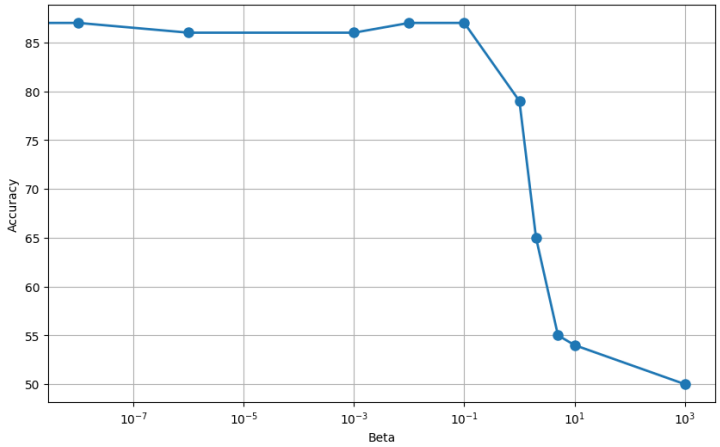
چون توزیع prior برای همه کلاس‌ها یکی می‌شود، می‌توان آن را درنظر نگرفت، اما توزیع دیریکله در likelihood نیز تاثیر می‌گذارد. چون فرض گرفته می‌شود که از قبل هر داده β بار دیده شده، موقع تخمین likelihood بجای محاسبه count باید از count + β استفاده کرد.

۳.۹ بخش c

باتوجه به نمودار مقادیر بالای β خوب نیست زیرا عملا باعث می‌شود تا در classifier تاثیر داده کمرنگ شود و همه کلاس‌ها هم احتمال شوند. اما مقادیر پایین آن باعث می‌شود تا دقت بهبود یابد. همچنین احتمال صفر شدن لگاریتم دیگر پیش نمی‌آید زیرا هر کلمه یک احتمال هرچند خیلی کم ولی غیر صفر دارد.

جدول ۲: تاثیر β بر دقت

β	**Accuracy**
0	83
10^{-8}	87
10^{-6}	86
0.001	86
0.01	87
0.1	87
1	79
2	65
5	55
10	54
1000.0	50

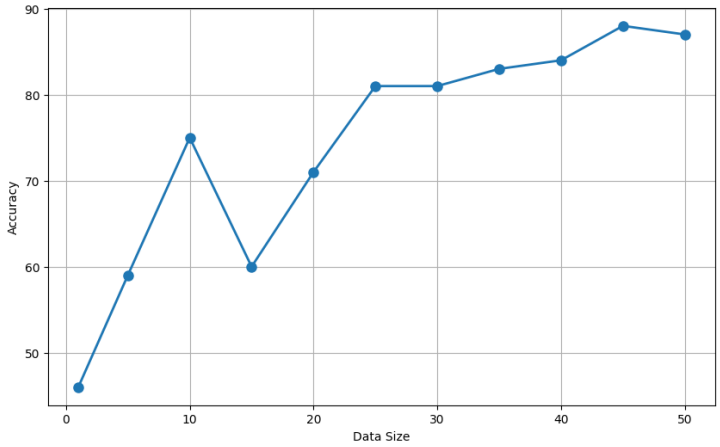


شکل ۶: تاثیر β در دقت. بهترین دقت در ۱.۰ بدست می‌آید.

۴.۹ بخش d

جدول ۳: تاثیر تعداد داده بر دقت

تعداد داده	درصد دقت
۱	۴۶
۵	۵۹
۱۰	۷۵
۱۵	۶۰
۲۰	۷۱
۲۵	۸۱
۳۰	۸۱
۳۵	۸۳
۴۰	۸۴
۴۵	۸۸
۵۰	۸۷



شکل ۷: می‌توان دید با بیش‌تر شدن داده دقت بهبود می‌یابد.

مراجع

[1] https://chat.deepseek.com/share/s1e5ijwd7xek7eoegg. Accessed: 2025-11-23.

[2] *Conjugate Prior*. https://en.wikipedia.org/wiki/Conjugate_prior. Accessed: 2025-11-23.

[3] *Dirichlet Distribution*. https://en.wikipedia.org/wiki/Dirichlet_distribution. Accessed: 2025-11-23.

[4] Geek for Geeks. *Multinomial Naive Bayes*. https://www.geeksforgeeks.org/machine-learning/multinomial-naive-bayes/. Accessed: 2025-11-23.

[5] *Multinomial*. https://en.wikipedia.org/wiki/Multinomial_distribution. Accessed: 2025-11-23.