به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس شناسایی آماری الگو

استاد رحمتی

تمرین دوم

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

سوال ۱

الف) جهت سادگی فرض می‌کنیم شماره دری که ما انتخاب کرده‌ایم برای ۱ است و دو در دیگر به ترتیب شـــماره ۲ و ۳ را دارند. X_i متغیر تصـــادفی باینری و به معنای آن است که پشت درب i لامبورگینی قرار گرفته است یا خیر. Y_i متغیر تصادفی باینری و به معنای آن اسـت که درب i توسـط مجری برنامه باز شـده اسـت یا خیر. طبیعتا پس از انتخاب در شماره ۱، دو حالت محتمل است؛ یا در ۲ توسط مجری باز می‌شود و یا در ۳. باتوجه به اینکه هر دوحالت مشـابه هم هسـتند بیایید فرض کنیم در شـماره ۲ توسـط مجری باز شـود و ما این انتخاب را داشـته باشـیم که در شـماره ۳ را به جای ۱ انتخاب کنیم (مشابه صورت سوال). حال باید ببینیم در این شرایط که مطمئن هستیم در شماره ۲ توسط مجری باز شده است (رخ دادن Y_2) احتمال آنکه لامبورگینی پشت در ۳ باشـد (X_3) چقدر اسـت. یعنی به دنبال محاسـبه $p(X_3|Y_2)$ هسـتیم. در این حالت برای محاسبه این مقدار می‌توان مدل بیز زیر را تعریف کرد:

$$p(X_3|Y_2) = \frac{p(Y_2|X_3) * p(X_3)}{p(Y_2)}$$

طبیعتـا در این رابطـه $p(X_3|Y_2)$ احتمـال posterior اســـت؛ $p(Y_2|X_3)$ احتمـال likelihood اســت؛ $p(X_3)$ احتمال prior اســت و $p(Y_2)$ احتمال Evidence. همانطور کـه پیش‌تر گفتیم معنـای احتمـال posterior در این مســـئلـه یعنی احتمـال آنکـه لامبورگینی پشـت در ۳ باشـد به شـرط آنکه در ۲ توسـط مجری باز شـده باشـد. معنای احتمال prior آن است که با چه احتمالی لامبورگینی پشت در ۳ باشد.

ب) بدیهی است که احتمال وجود لامبورگینی پشت هر در، بدون مشاهده هیچ پدیده دیگری تماما با یگدیگر برابر است. پس:

$$p(X_3) = \frac{1}{3}$$

ج) چنانچه در ۱ توسط ما انتخاب شده باشد و پشت در ۳ لامبورگینی باشد، مجری حتما در ۲ را باز خواهد کرد پس داریم:

$$p(Y_2|X_3) = 1$$

د)

$$p(Y_2) = p(Y_2|X_1) * p(X_1) + p(Y_2|X_2) * p(X_2) + p(Y_2|X_3) * \ p(X_2)$$
$$= \frac{1}{2} * \frac{1}{3} + 0 + 1 * \frac{1}{3} = \frac{1}{2}$$

$$p(X_3|Y_2) = \frac{p(Y_2|X_3) * p(X_3)}{p(Y_2)} = \frac{1 * \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

ه) احتمال آنکه در شماره ۱ بدون مشاهده هیچ پدیده‌ای دارای لامبورگینی باشد برابر است با $p(X_1) = \frac{1}{3}$ و احتمال آنکه پس از مشاهده بازشدن در شماره ۲، در شماره ۳ دارای لامبورگینی باشد $p(X_3|Y_2) = \frac{2}{3}$ است. پس تعویض در پس از مشاهده مدنظر معقول است.

و) در این حالت دیگر تحلیل‌های قبلی صحیح نخواهد بود؛ به عنوان مثال برای محاسبه احتمال likelihood ما فرض می‌کردیم که اگر لامبورگینی پشت در ۳ باشد، مجری حتما در ۲ را باز می‌کند که الان چنین فرضی برقرار نیست.

ز) برای این حالت هم فرض‌های اولیه قسمت‌های پیشین سوال را درنظر می‌گیریم. یعنی ما در شماره ۱ را انتخاب می‌کنیم. مجری در ۲ را باز می‌کند و می‌توانیم در این شرایط انتخابمان را به در ۳ تغییر دهیم. پیش از هر چیز باید به این نکته توجه کنیم که اساسا در شرایطی مسئله این قسمت مطرح است که لامبورگینی پشت در ۲ نباشد؛ چراکه اگر باشد ما برنده می‌شویم و دیگر فرصتی برای تغییر انتخاب یا عدم تغییر نمی‌ماند. پس در همان مدل قبلی شرط عدم وجود لامبورگینی پشت در ۲ اضافه می‌شود. (X_2')

$$p(X_3|X_2') = \frac{1}{2}$$

$$p(Y_2|X_3X_2') = \frac{1}{2}$$

$$p(Y_2|X_2') = p(Y_2|X_1X_2') * p(X_1X_2') + p(Y_2|X_2X_2') * p(X_2X_2') + p(Y_2|X_3X_2') * \ p(X_2) = \frac{1}{2} * \frac{1}{2} + 0 + \frac{1}{2} * \frac{1}{2} = \frac{1}{2}$$

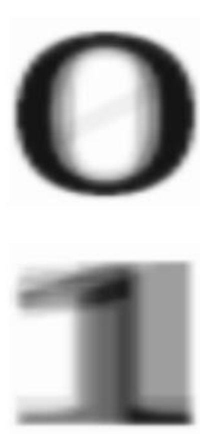
$$p(X_3|Y_2X_2') = \frac{p(Y_2|X_3X_2') * p(X_3|X_2')}{p(Y_2|X_2')} = \frac{\frac{1}{2} * \frac{1}{2}}{\frac{1}{2}} = \frac{1}{2}$$

در این حالت تغییر در سودی ندارد چراکه اگر لامبورگینی پشت در ۲ نباشد، پافشاری بر انتخاب در ۱ به سود نیم منجر می‌شود و تغییر انتخاب هم باز به سود نیم منجر می‌شود. اگر هم که لامبورگینی پشت در ۲ باشد که اصلا مسئله تغییر در مطرح نخواهد بود. نهایتا برای مقایسه بین این قسمت و قسمت‌های اول سوال توجه کنید که شانس کلی پیروزی برابر با نیم نیست و این احتمال برای وقتی است که لامبورگینی پشت در ۲ نباشد در این حالت احتمال موفقیت چه انتخاب عوض شود و چه نشود برابر است با:

$$P(X_1|X_2') + P(X_2) = \frac{1}{2} * \frac{2}{3} + \frac{1}{3} = \frac{2}{3}$$

سوال ۲

الف)



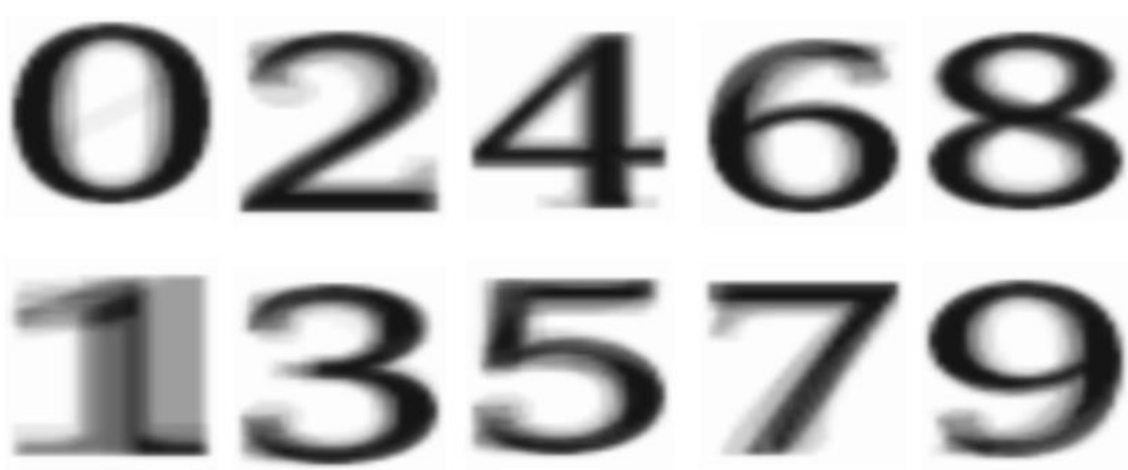
ب) صحت برای این گروه برابر است با ۱۰۰٪ و برای تمام داده‌ها به درستی کلاس تعیین می‌شود.

ج)

صحت برای این گروه برابر با ۹۲/۵٪ است. ماتریس درهم‌ریختگی این گروه هم در ادامه آورده شده است.

	۰	۱	۲	۳
۰	۹	۰	۱	۰
۱	۰	۸	۰	۲
۲	۰	۰	۱۰	۰
۳	۰	۰	۰	۱۰

د)



صحت برای این گروه برابر با ۹۲٪ درصد است. ماتریس درهم‌ریختگی این گروه هم در ادامه آورده شده است.

	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۹	۰	۱	۰	۰	۰	۰	۰	۰	۰
۱	۰	۸	۰	۲	۰	۰	۰	۰	۰	۰
۲	۰	۰	۱۰	۰	۰	۰	۰	۰	۰	۰
۳	۰	۰	۰	۹	۰	۰	۰	۱	۰	۰
۴	۰	۰	۰	۰	۱۰	۰	۰	۰	۰	۰
۵	۰	۰	۰	۰	۰	۱۰	۰	۰	۰	۰
۶	۰	۰	۰	۰	۰	۱	۹	۰	۰	۰
۷	۰	۰	۰	۰	۰	۰	۰	۱۰	۰	۰
۸	۰	۰	۰	۲	۰	۰	۰	۰	۸	۰
۹	۰	۰	۰	۱	۰	۰	۰	۰	۰	۹

ه) اولین نکته‌ای که می‌توان درنظر گرفت این است که هرچه تعداد کلاس‌ها بیشتر شود، طبیعتا کار مدل برای پیش‌بینی سخت‌تر می‌شود و صحت مدل پایین می‌آید. نکته دوم این است که برخی کلاس‌ها با یکدیگر اشتباه گرفته شده‌اند. بیشترین کلاسی که مورد اشتباه بوده است، کلاس 3 است و این احتمالا به این دلیل است که الگوی این کلاس به الگوی کلاس‌های زیادی شبیه است. در عین حال می‌بینیم که یک داده کلاس 6 به کلاس 5 و یک داده از 0 به 2 رفته است. در این موارد هم می‌توان شبیه بودن الگوی دو کلاس را درنظر داشت. در همین حال می‌بینیم کلاس 4 که الگوی نسبتا متمایزی دارد، به صورت مناسبی از بقیه جدا شده است.

سوال ۳

الف)

$$p(Y_1, \dots, Y_s \mid \lambda) = p(Y_1|\lambda) * \dots * p(Y_s|\lambda) = \prod_{i=1}^{s} \frac{\lambda_i^{k_i} e^{-\lambda_i}}{k_i!} = \prod_{i=1}^{s} \frac{(n_i x_i \lambda)^{k_i} e^{-n_i x_i \lambda}}{k_i!}$$

ب) باتوجه به اینکه در تحقیق i-ام، تعداد y_i نفر مبتلا به سرطان شدند. لذا پارامتر λ را باید به گونه‌ای تخمین زد که در مدل هر تحقیق احتمال مربوط به $k_i = y_i$ یک احتمال زیاد باشد (در حالت آرمانی، بیشترین احتمال). لذا باید تابع توام که در قسمت الف بدست آوردیم برای $k_i = y_i$ ها به بیشترین مقدار خود برسد:

$$\hat{\lambda} = \underset{\lambda}{argmax}\left(p(Y_1, \dots, Y_s|\lambda)\right) = \underset{\lambda}{argmax}\left(\prod_{i=1}^{s} \frac{(n_i x_i \lambda)^{k_i} e^{-n_i x_i \lambda}}{k_i!}\right)$$

$$k_i = y_i \to \hat{\lambda} = \underset{\lambda}{argmax}\left(\prod_{i=1}^{s} \frac{(n_i x_i \lambda)^{y_i} e^{-n_i x_i \lambda}}{y_i!}\right) = \underset{\lambda}{argmax}\left(I(\lambda)\right)$$

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\left(I(\lambda)\right) = \underset{\lambda}{\operatorname{argmax}}\left(\ln\left(I(\lambda)\right)\right)$$

$$\frac{d}{d\lambda} I(\lambda) = 0 \to \frac{d}{d\lambda} I(\lambda) = \sum_{i=1}^{s} \frac{d}{d\lambda} \ln\left(\frac{(n_i x_i \lambda)^{y_i} e^{-n_i x_i \lambda}}{y_i!}\right)$$

$$= \sum_{i=1}^{s} \frac{d}{d\lambda}\left(\ln((n_i x_i)^{y_i}) + \ln(\lambda^{y_i}) + \ln\left(e^{-n_i x_i \lambda}\right) - \ln(y_i!)\right)$$

$$= \sum_{i=1}^{s} \frac{d}{d\lambda}\left(y_i \ln(\lambda) - n_i x_i \lambda\right) = \sum_{i=1}^{s}\left(\frac{y_i}{\lambda} - n_i x_i\right) = \frac{1}{\lambda}\sum_{i=1}^{s} y_i - \sum_{i=1}^{s} n_i x_i$$

$$\to \frac{1}{\lambda}\sum_{i=1}^{s} y_i - \sum_{i=1}^{s} n_i x_i = 0 \to \frac{1}{\lambda}\sum_{i=1}^{s} y_i = \sum_{i=1}^{s} n_i x_i \to \lambda = \frac{\sum_{i=1}^{s} y_i}{\sum_{i=1}^{s} n_i x_i}$$

$$\hat{\lambda} = \frac{\sum_{i=1}^{s} y_i}{\sum_{i=1}^{s} n_i x_i}$$

ج) بله؛ طبق محاسبات زیر MLE یک تخمین‌گر unbiased است:

$$E[\hat{\lambda}] = \hat{\lambda} = \frac{\sum_{i=1}^{s} y_i}{\sum_{i=1}^{s} n_i x_i} = \frac{\sum_{i=1}^{s} \lambda n_i x_i}{\sum_{i=1}^{s} n_i x_i} = \lambda * \frac{\sum_{i=1}^{s} n_i x_i}{\sum_{i=1}^{s} n_i x_i} = \lambda$$

د)

$$E\left[\left(\hat{\lambda} - \lambda\right)^2\right] = E[\hat{\lambda}^2 - 2\hat{\lambda}\lambda + \lambda^2] = E[\hat{\lambda}^2] - 2E[\hat{\lambda}\lambda] + E[\lambda^2]$$
$$= \hat{\lambda}^2 - 2\hat{\lambda}E[\lambda] + E[\lambda^2]$$

?

ه)

$$mean\ squered\ error\ (\hat{\lambda}) = variance(\hat{\lambda}) + bias(\hat{\lambda}) = variance(\hat{\lambda}) + 0$$

?

و)

$$\hat{\lambda} = \frac{\sum_{i=1}^{s} y_i}{\sum_{i=1}^{s} n_i x_i} = \frac{2 + 5 + 14}{20 * 0.4 + 50 * 0.3 + 100 * 0.6} = \frac{21}{83}$$

ز)

$$\hat{\lambda} = argmax_{\lambda}(p(D|\theta)p(\theta)) = arg \max_{\lambda}\left(\prod_{i=1}^{s} \frac{(n_i x_i \lambda)^{k_i} e^{-n_i x_i \lambda}}{k_i!} * \Gamma(\alpha, \beta)\right)$$

$$k_i = y_i\ , x_{Gamma} = \lambda \to \hat{\lambda} = argmax_{\lambda}\left(\left(\prod_{i=1}^{s} \frac{(n_i x_i \lambda)^{y_i} e^{-n_i x_i \lambda}}{y_i!}\right) * \frac{\lambda^{\alpha-1} e^{-\beta\lambda} \beta^{\alpha}}{\Gamma(\alpha)}\right)$$
$$= argmax_{\lambda}(I(\lambda))$$

$$\frac{d}{d\lambda}I(\lambda) = 0 \rightarrow \frac{d}{d\lambda}I(\lambda) = \sum_{i=1}^{s}\frac{d}{d\lambda}\ln\left(\frac{(n_i x_i \lambda)^{y_i} e^{-n_i x_i \lambda}}{y_i!}\right) + \frac{d}{d\lambda}\ln\left(\frac{\lambda^{\alpha-1}e^{-\beta\lambda}\beta^{\alpha}}{\Gamma(\alpha)}\right)$$

$$= \sum_{i=1}^{s}\frac{d}{d\lambda}\left(\ln((n_i x_i)^{y_i}) + \ln(\lambda^{y_i}) + \ln\left(e^{-n_i x_i \lambda}\right) - \ln(y_i!)\right)$$

$$+ \frac{d}{d\lambda}\left(\ln(\lambda^{\alpha-1}) + \ln\left(e^{-\beta\lambda}\right) + \ln(\beta^{\alpha}) - \ln(\Gamma(\alpha))\right)$$

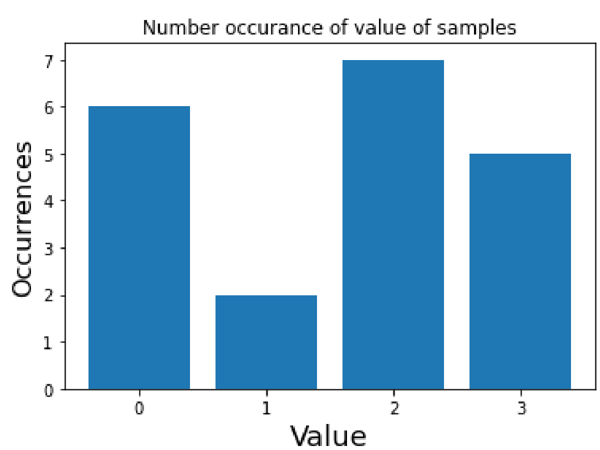
$$= \sum_{i=1}^{s}\frac{d}{d\lambda}(y_i \ln(\lambda) - n_i x_i \lambda) + \frac{\alpha-1}{\lambda} - \beta$$

$$= \sum_{i=1}^{s}\left(\frac{y_i}{\lambda} - n_i x_i\right) + \frac{\alpha-1}{\lambda} - \beta = \frac{1}{\lambda}\left(\sum_{i=1}^{s} y_i + \alpha - 1\right) - \sum_{i=1}^{s} n_i x_i - \beta$$

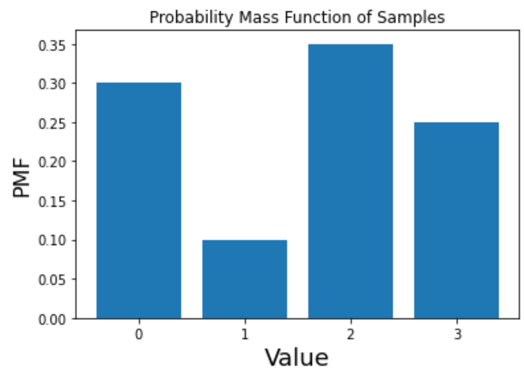
$$\rightarrow \frac{1}{\lambda}\left(\sum_{i=1}^{s} y_i + \alpha - 1\right) - \sum_{i=1}^{s} n_i x_i - \beta = 0 \rightarrow \lambda = \frac{\sum_{i=1}^{s} y_i + \alpha - 1}{\sum_{i=1}^{s} n_i x_i + \beta}$$

سوال ۴

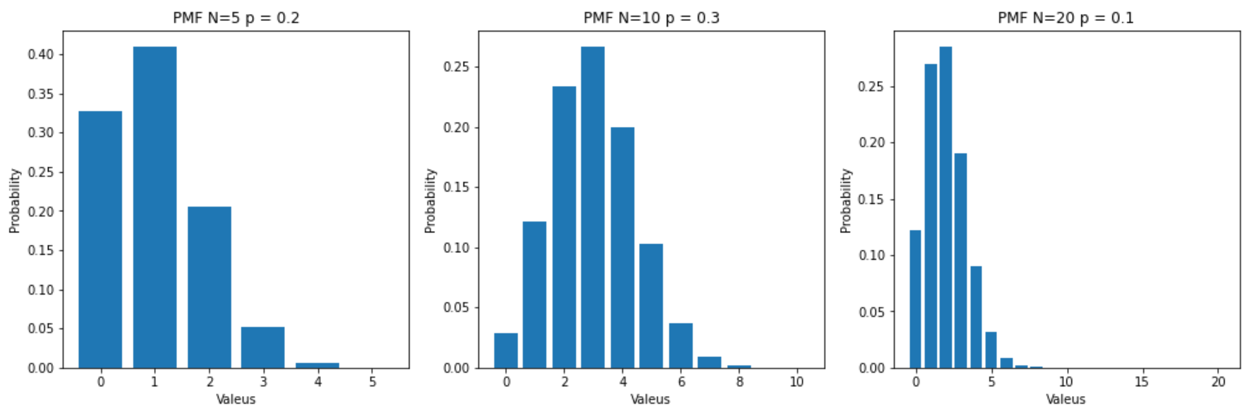
الف)



ب)

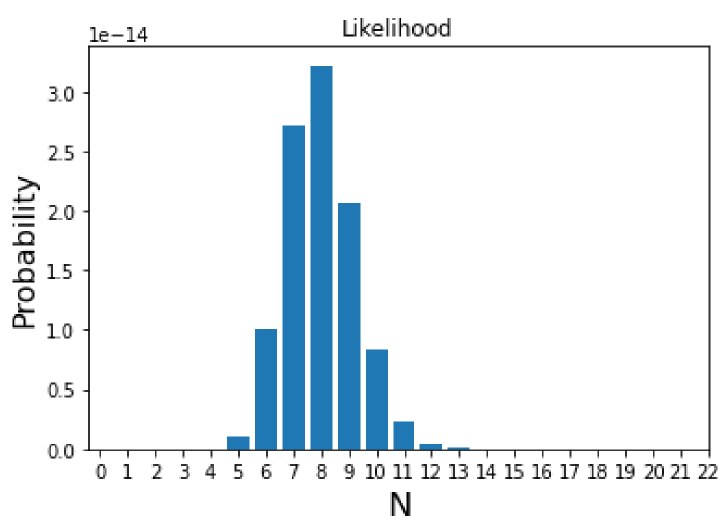


ج)

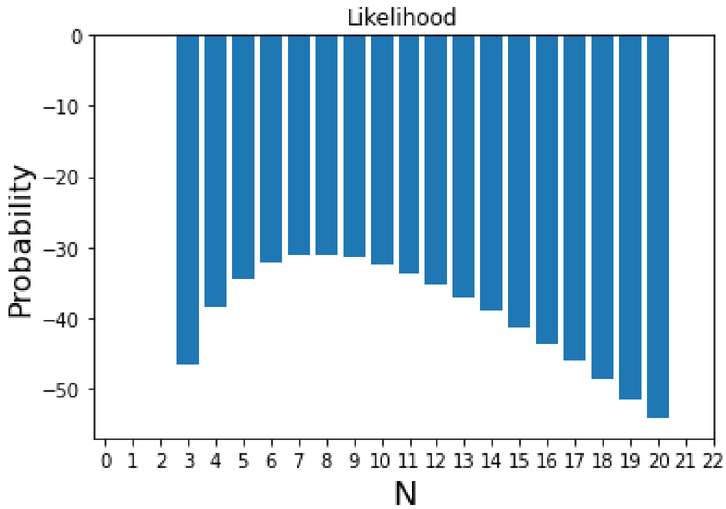


د) به نظر نمودار مربوط به $N = 5$ و $p = 0.2$ از مابقی مناســب‌تر باشــد چراکه دو نمودار دیگر، احتمال زیادی را برای اعداد بزرگ‌تر از ۳ دارند.

ه)

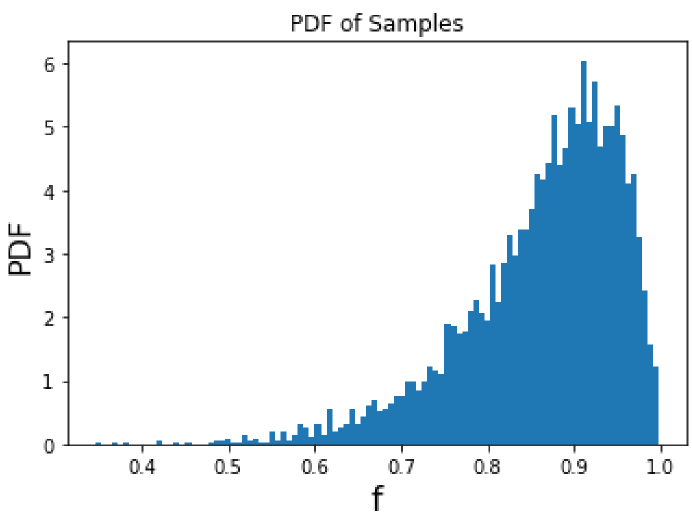


و) توجه کنید که برای این قسمت، برای N=1 و N=2 چون احتمال غیر لگاریتمی صفر است، نموداری ترسیم نشده است.



ز) به نظر می‌رسد برای N=8 می‌توان به بیشترین احتمال رسید.

ح)



ط)

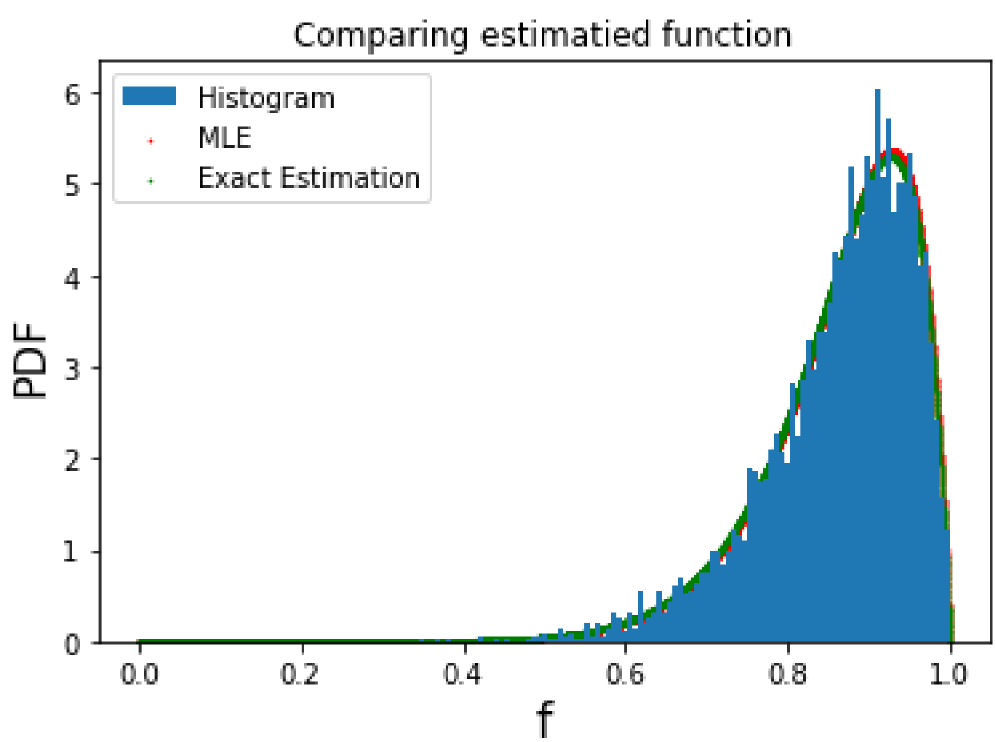
$$L(a,b) = L(D|a,b) = \prod L(x_i|a,b) = \prod \left(abx_i^{a-1}(1-x_i^a)^{b-1}\right)$$

$$lnL(a,b) = \ln\left(\prod\left(abx_i^{a-1}(1-x_i^a)^{b-1}\right)\right)$$
$$= \sum\left(\ln a + \ln b + \ln x_i^{a-1} + \ln(1-x_i^a)^{b-1}\right)$$
$$= n * \ln a + n * \ln b + (a-1)\sum \ln(x_i) + (b-1)\sum \ln(1-x_i^a)$$

ی) a=10 و b=2

ک) a=9.787 و b=2.02

ل) به نظر می‌رســـد هر دو تخمین به خوبی توانســـته‌اند روی داده‌ها فیت شـــوند و تفاوت جدی‌ای میان این دو مشاهده نمی‌شود. انتظار تفاوت جدی هم وجود نداشت چراکه واقعا دو پارامتر a و b در هر دو روش به یکدیگر بسیار نزدیک هستند.



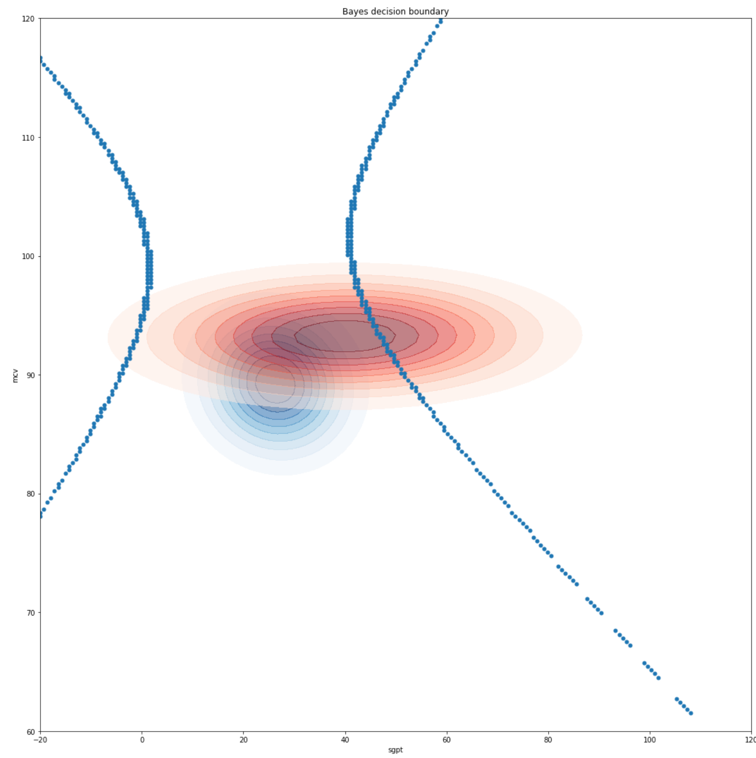
سوال ۵

الف، ب، ج)

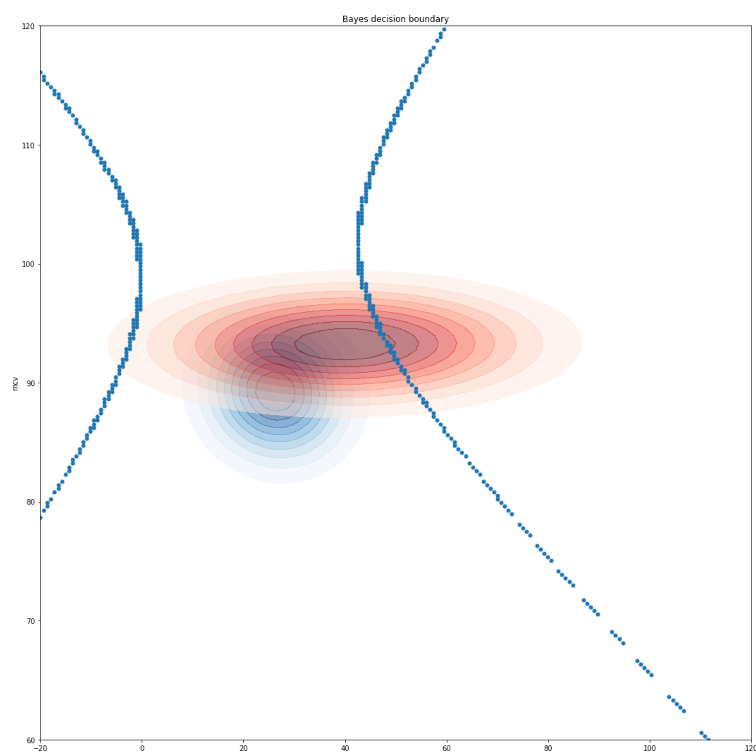
خطای BE	خطا MLE	تعداد داده
۲۱/۵٪	۲۱/۵٪	۱۰
۲۲٪	۲۷٪	۵۰
۲۱/۵٪	۲۱٪	۱۰۰

برای حالت ۱۰، هر دو روش تمامی داده‌ها را در یک کلاس قرار می‌دادند که همین کار برای کسب دقت ۷۸/۵٪ کافی بود! برای ۱۰۰ داده هر دو الگوریتم تقریبا مشابه عمل کرده‌اند اما برای ۵۰ داده الگوریتم BE بهتر بوده است. شاید این برتری به دلیل استفاده از اطلاعات پیشین بوده است.

د) با بررسی میانگین و واریانس دو کلاس به نظر می‌رسد دو ویژگی mcv و sgpt به نحو بهتری می‌تواند داده‌های دو کلاس را از یکدیگر جدا کند. با بررسی این جفت ویژگی و جفت ویژگی‌های مناسب احتمالی دیگر متوجه شدم که همین جفت ویژگی با صحت ۷۷/۵٪، از مابقی مناسب‌تر است.


Bayes decision boundary
mcv
sgpt
120
110
100
90
80
70
60
−20
0
20
40
60
80
100
120

۵)


Bayes decision boundary
mcv
sgpt
120
110
100
90
80
70
60
−20
0
20
40
60
80
100
120

سوال ۶

الف) احتمال پیشین اینکه یک پیکسل متعلق به پوست باشد حدودا برابر با ۰/۷۲ و غیر آن حدودا برابر با ۰/۲۷ است.

ب) برای کلاس پوست خواهیم داشت:

$$\mu \approx \begin{bmatrix} 129 \\ 106 \\ 96 \end{bmatrix}, \Sigma \approx \begin{bmatrix} 7762 & 6717 & 6294 \\ 6717 & 7075 & 6916 \\ 6294 & 6916 & 7145 \end{bmatrix}$$

برای کلاس غیرپوست خواهیم داشت:

$$\mu = \begin{bmatrix} 189 \\ 140 \\ 111 \end{bmatrix}, \Sigma = \begin{bmatrix} 1962 & 1953 & 1861 \\ 1953 & 2383 & 2449 \\ 1861 & 2449 & 2675 \end{bmatrix}$$

ج)

د)

مطابق انتظار پیش‌بینی انجام‌شده توسط MDC دقت کمتری دارد. این تفاوت در اولین تصویر بیشتر مشهود است. می‌دانیم میانگین کلاس پوست برخلاف کلاس غیرپوست، یک پیکسل کم‌رنگ است. لذا رنگ‌های کم‌رنگ دیوار در MDC جز پوست به حساب آمدند ولی در روش دیگر چون توزیع احتمالاتی سه‌بعدی وجود دارد و برای کلاس پوست مقادیر ماتریس کوارانس اعداد کوچکی است، باعث شده است تا بخش زیادی از دیوار به درستی جز غیرپوست حساب شود.

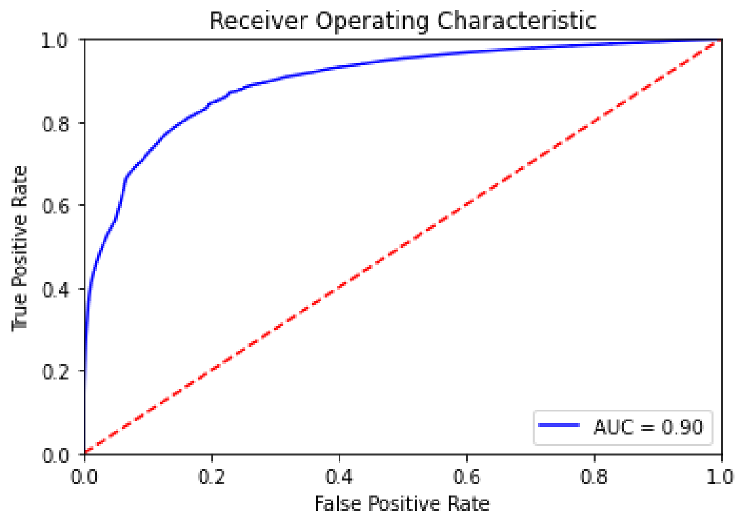
ه) صحت مدل حدودا برابر با ۶۲/۶۶٪ است و خطای مدل حدودا برابر با ۳۷/۳۴٪ خواهد بود.

و)

	skin	non-skin
skin	۴۰۲۳۹۰	۲۱۲۴۵
non-skin	۹۹۱۶۲۴	۱۲۹۷۹۱۹

ز) حدودا ۱۱/۱۵٪ است.

ح) با فرض درنظر گرفتن پوست به عنوان کلاس مثبت خواهیم داشت:



سوال ۷

الف) مدل بیز بهترین قانون است اگر بتوان آن را حساب کرد! برای اعمال آن نیاز است تا احتمال prior و likelihood را داشته باشیم ولی مثلا ممکن است احتمال prior را نداشته باشیم.

ب) بله می‌توان؛ به عنوان مثال برای روش‌های خطی، روشی تحت عنوان Bayesian Linear Regression وجود دارد که این کار را انجام می‌دهد. در مسائل رگرسیون به دنبال پارامترهایی هستیم که خطا را عبارتی مانند عبارت زیر کمینه کند:

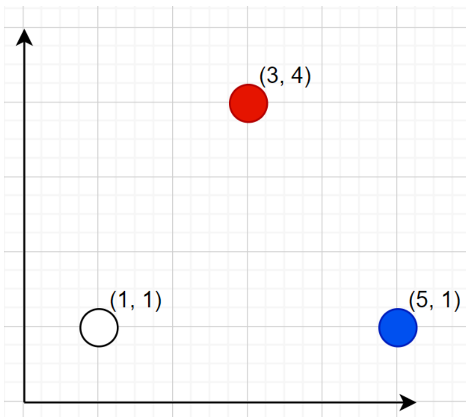
$$y = \beta^T X + \varepsilon$$

در این روش y یک مقدار ثابت نیست که بدست آید بلکه آن را باید با یک توزیع احتمالاتی بدست آورد. با داشتن پارامترها یعنی β، می‌توان توزیع زیر را پیشنهاد کرد:

$$y \sim N(B^T X, \sigma^2 I)$$

نهایتا باید β را پیدا کرد که می‌توان از رابطه زیر برای پیداکردن محتمل‌ترین آن استفاده کرد.[1]

$$P(\beta|y,x) = \frac{P(y|\beta,X) * P(\beta|X)}{P(y|X)}$$



ج) فرض کنید میانگین نمونه‌های کلاس ۱، نقطه آبی و میانگین نمونه‌های کلاس ۲، نقطه قرمز باشد و به دنبال تعیین کلاس نقطه سفید باشیم. اگر از معیار منهتن استفاده کنیم، داده تست به کلاس آبی (فاصله ۴) نزدیک‌تر از کلاس قرمز (فاصله ۵) است. اگر از معیار اقلیدوسی استفاده کنیم، داده تست به کلاس قرمز (فاصله $\sqrt{۱۳}$) نزدیک‌تر از کلاس آبی (فاصله ۴) است.

د) در اصل محاسبه احتمال‌های prior و likelihood نقش فاز آموزش را خواهد داشت. در مورد MDC هم، بدست آوردن پروتوتایپ کلاس‌های مختلف فاز آموزش محسوب می‌شود.

ه) چنانچه دانش پیشین ما، نسبت به مشاهدات فعلی دارای خطا باشد. مثلا اگر شرایط مسئله تغییر جدی داشته باشد و همچنین مشاهدات فعلی کافی باشد احتمالا MLE از MAP نتیجه بهتری داشته باشد.

---

[1] https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7