

Q1)

a) $1/2$

b) $1/2 * (0.35) + 1/2 * (0.65) = 0.5$

c) $P(x|y) P(y) = P(y|x) P(x)$

$$P(x|y) * 1/2 = 0.65 * 1/2$$

$$\Rightarrow P(x|y) = 0.65$$

d)

$$P(\text{not}(X) \text{ and } \text{not}(Y)) = 0.37$$

$$P(X) = 0.44$$

$$P(Y) = 0.52$$

$$\Rightarrow P(X \text{ or } Y) = 0.63$$

$$P(X \text{ or } Y) = P(X) + P(Y) - P(X \text{ and } Y)$$

$$0.63 = 0.44 + 0.52 - P(X \text{ and } Y)$$

$$P(X \text{ and } Y) = 0.44 + 0.52 - 0.63 = 0.33$$

$P(Y|X)$ where Y = better result on exams, X = sleep before midnight

$$P(Y|X) = P(XY) / P(X)$$

$$= 0.33 / 0.44$$

$$= 0.75$$

which means 75% of students who sleep before midnight have GPD more than 16

e)

$$1 - 0.57 = 0.43$$

f)

$$P(\text{female} | \text{smoking}) = P(\text{female and smoking}) / P(\text{smoking})$$

$$= 0.024 / 0.146 + 0.024 = 0.024 / 0.17 = 0.1411$$

g)

$$P(\text{pos} | \text{used})(100 - 1.4) \% = 0.986$$

$$P(\text{pos} | \text{not_used}) = 0.098$$

$$P(\text{neg} | \text{used}) = 0.014$$

h)

$$P(\text{used}) = 0.02$$

if by population you mean all the athletes:

$$p(\text{pos}) = P(\text{used}) * P(\text{pos} | \text{used}) + P(\text{not_used}) * P(\text{pos} | \text{not_used})$$

$$= 0.986 * 0.02 + 0.098 * 0.98$$

$$= 0.09604 + 0.01972$$

$$= 0.11576$$

I)

$$\begin{aligned} P(\text{used} \mid \text{pos}) \\ P(\text{used} \mid \text{pos}) * P(\text{pos}) = P(\text{pos} \mid \text{used}) * P(\text{used}) \Rightarrow \\ P(\text{used} \mid \text{pos}) * 0.11576 = 0.986 * 0.02 \\ P(\text{used} \mid \text{pos}) = 0.17035 \end{aligned}$$

j)

$$\begin{aligned} P(\text{used}) &= 1/2 \\ P(\text{used} \mid \text{pos}) * P(\text{pos}) &= P(\text{pos} \mid \text{used}) * P(\text{used}) \end{aligned}$$

$$\begin{aligned} P(\text{pos}) &= P(\text{used}) * P(\text{pos} \mid \text{used}) + P(\text{not_used}) * P(\text{pos} \mid \text{not_used}) \\ P(\text{pos}) &= 0.5 * 0.986 + 0.5 * 0.098 \\ P(\text{pos}) &= 0.542 \end{aligned}$$

$$\begin{aligned} P(\text{used} \mid \text{pos}) * 0.542 &= 0.986 * 0.5 \\ P(\text{used} \mid \text{pos}) &= 0.9095 \end{aligned}$$

K)

Q2)

a) features L and b are selected

b)

١٩٣ - ١٧٧

$$P(x|w_i) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (x - \mu_i)^T [\Sigma]^{-1} (x - \mu_i) \right\}$$

$$\mu_1 = [42.4, 1.8], \mu_2 = [74, 41.8], \mu_3 = [5.4, 51.2]$$

$$P(x|w_1) = \frac{1}{2\pi} e^{-\frac{1}{2} \{(x_1 - 42.4)^2 + (x_2 - 1.8)^2\}}$$

$$P(m|w_1) = \frac{1}{2\pi} e^{-\frac{1}{2} \{(m_1 - 74)^2 + (m_2 - 41.8)^2\}}$$

$$P(m|w_2) = \frac{1}{2\pi} e^{-\frac{1}{2} \{(m_1 - 5.4)^2 + (m_2 - 51.2)^2\}}$$

$$\textcircled{1} P(w_1) * P(m|w_1) \underset{w_2}{\gtrless} P(w_2) P(m|w_2)$$

١٩٣ - ١٧٧

جامعة

١٣٩٦/٧/٧ | 2017/9/29
الخميس ١٢٩ | Fri.

$$(m_1 - 74)^2 + (m_2 - 41.8)^2 \underset{w_1}{\gtrless} (m_1 - 42.4)^2 + (m_2 - 1.8)^2$$

$$(-31.6) \times (2m_1 - 116.4) \underset{w_2}{\gtrless} (40) \times (2m_2 - 43.6)$$

$$\frac{-31.6}{40} (2m_1 - 116.4) \underset{w_1, w_2}{\gtrless} (2m_2 - 43.6)$$

M	T	W	T	F	S	S
1	2	3				
4	5	6	7	8	9	10
						١٧

① $\text{IV. } \begin{matrix} w_1 \\ -1,58w_1 + 135,556 \end{matrix} \begin{matrix} w_2 \\ w_2 \end{matrix} \rightarrow \begin{matrix} w_2 + .79w_1 \\ \sum w_1 \end{matrix} \begin{matrix} w_2 \\ w_1 \end{matrix} \begin{matrix} 67,778 \\ ? \end{matrix}$

② $\text{VI and VII. } \begin{matrix} w_3 \\ (w_1 - 42,4)^2 + (w_2 - 1,8)^2 \end{matrix} \begin{matrix} w_1 \\ w_1 \end{matrix} \begin{matrix} (-37) \times (2w_1 - 47,8) \\ 2w_1 - 47,8 \end{matrix} \begin{matrix} w_3 \\ w_1 \end{matrix} \begin{matrix} (-49,4) \times (2w_2 - 53) \\ 1,33(2w_2 - 53) \end{matrix} \begin{matrix} w_1 \\ w_1 \end{matrix} \begin{matrix} w_1 + 11,345 \\ w_3 \end{matrix} \begin{matrix} 1,33w_2 \\ \rightarrow w_1 - 1,33w_2 \end{matrix} \begin{matrix} -11,345 \\ w_3 \end{matrix}$

③ $\text{VIII and IX. } \begin{matrix} w_3 \\ (w_1 - 74)^2 + (w_2 - 41,8)^2 \end{matrix} \begin{matrix} w_2 \\ w_2 \end{matrix} \begin{matrix} -68,6 \times (2w_1 - 79,4) \\ 7,29(2w_1 - 79,4) \end{matrix} \begin{matrix} w_3 \\ w_2 \end{matrix} \begin{matrix} -9,4 \times (2w_2 - 93) \\ 2w_2 - 93 \end{matrix} \begin{matrix} 242,913 \\ w_3 \end{matrix}$

c)

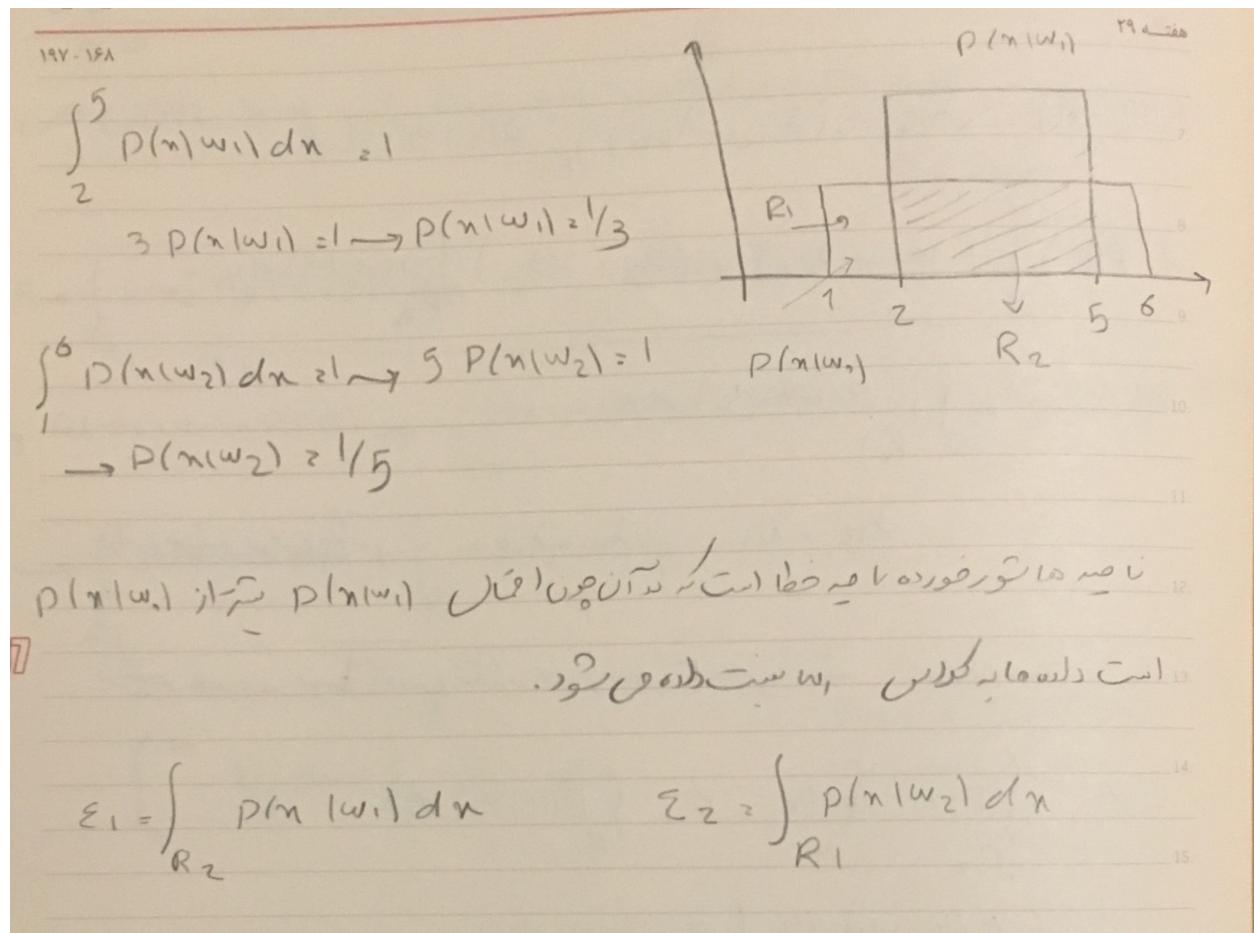
	w2_w1	w1_w3	w2_w3	class
84, 62	60.5820	12.88499	307.447	2
90, 63	39.322	53.464999	377.187	2
8, 3	-58.458	15.355	-187.593	1
57, 3	-19.74800	64.355	169.6170	1
63, 49	30.9920	9.1749	167.356999	2

d)

در این دسته‌بند فقط از دو ویژگی استفاده شده است که این محدودیت در انتخاب ویژگی‌ها می‌تواند دسته‌بند را محدود کند و در دقت آن تاثیر بگذارد. همچنین اگر داده‌های یک کلاس در مجموعه داده آموزش وجود نداشته باشد و یا تعداد آن‌ها کم باشد، احتمال prior آن‌ها کم می‌شود.

Q3)

A)



B)

18
17

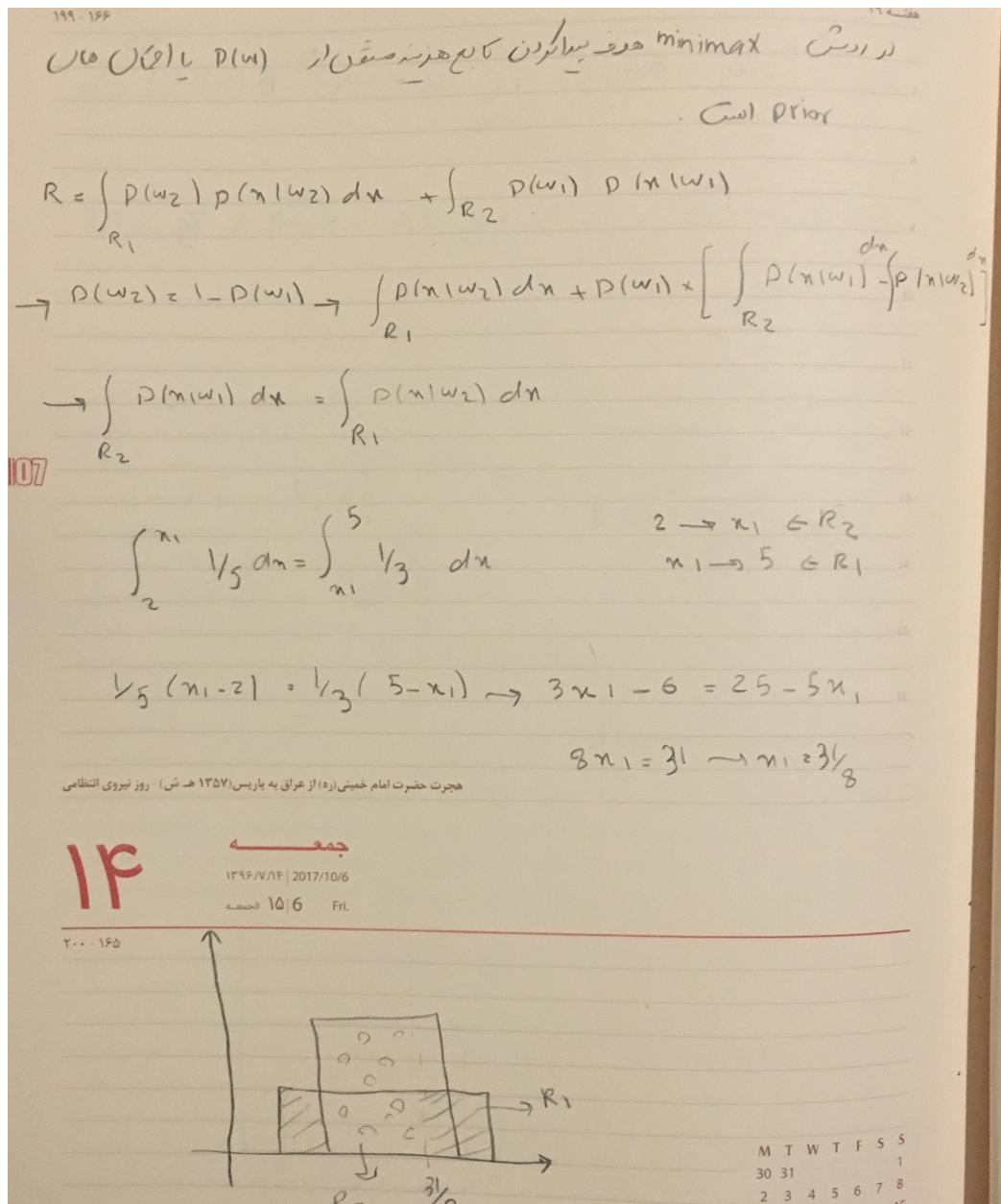
$$\varepsilon_1 = \int_{R_2} p(m|w_1) dm = \frac{1}{100} \rightarrow \int_{R_2} dm = \frac{3}{100}$$

فرضیه ε_2 کویند \min ناپذیر است

$$\varepsilon_2 = \min_{R_1} \int_{R_1} p(m|w_2) dm \rightarrow \varepsilon_2 = \min_{R_1} \int_{R_1} 1/5 dm$$

$$\varepsilon_2 = 1/5 \times 1/3 - \frac{3}{100} = 1/5 \times 2.97 = 0.594$$

C)

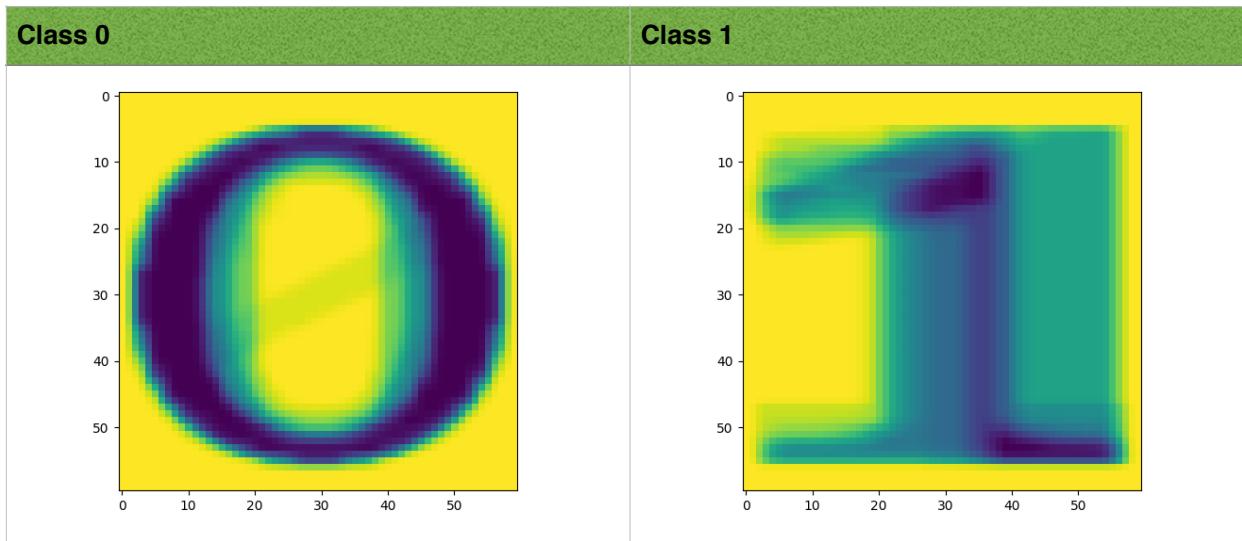


D)

$$\begin{aligned}
 \mathcal{E}_4 &= P(w_1)^s P(w_2)^{1-s} \int_{-\infty}^{\infty} P(m|w_1), P(n|w_2) dn \\
 &= \sqrt{0.4} \times \sqrt{0.6} \times \int_2^5 \frac{1}{3} \times \frac{1}{5} dn = \sqrt{0.24} \times \frac{1}{5} = 0.1
 \end{aligned}$$

Q4)

A)

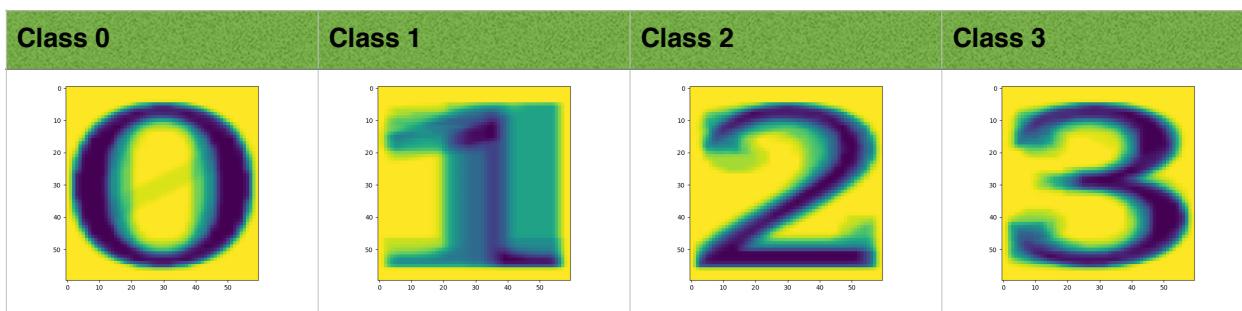


B)

confusion matrix	error
$\begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$	0

در این مجموعه داده، کلاس تمام داده‌ها درست تشخیص داده شده است و خطای دسته‌بند صفر است.

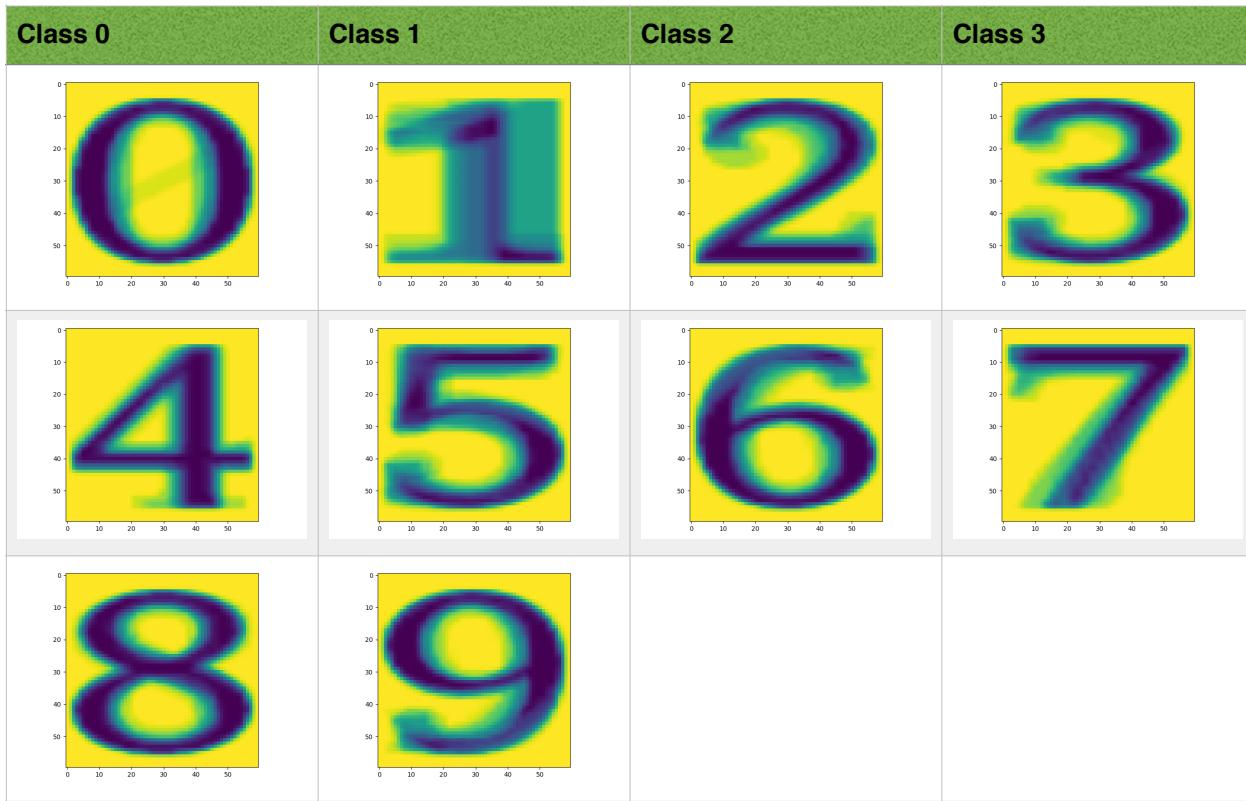
C)



confusion matrix	error
$\begin{bmatrix} 10 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 10 \end{bmatrix}$	0

در این مجموعه داده نیز کلاس تمام داده‌ها درست تشخیص داده شده است و خطای دسته‌بند صفر است.

D)



در این مجموعه داده، با افزایش تعداد کلاس‌ها کلاس‌ها تمام داده‌ها به جز یک داده درست تشخیص داده شده‌است. در اینجا فقط در یک مورد کلاس ۸ به ۳ اشتباه نسبت داده شده‌است. در این مورد نوشته به کلاس ۸ نزدیک‌تر بوده که این به دلیل این است که در این دسته بند فقط فاصله‌ی برداری مورد استفاده قرار می‌گیرد.

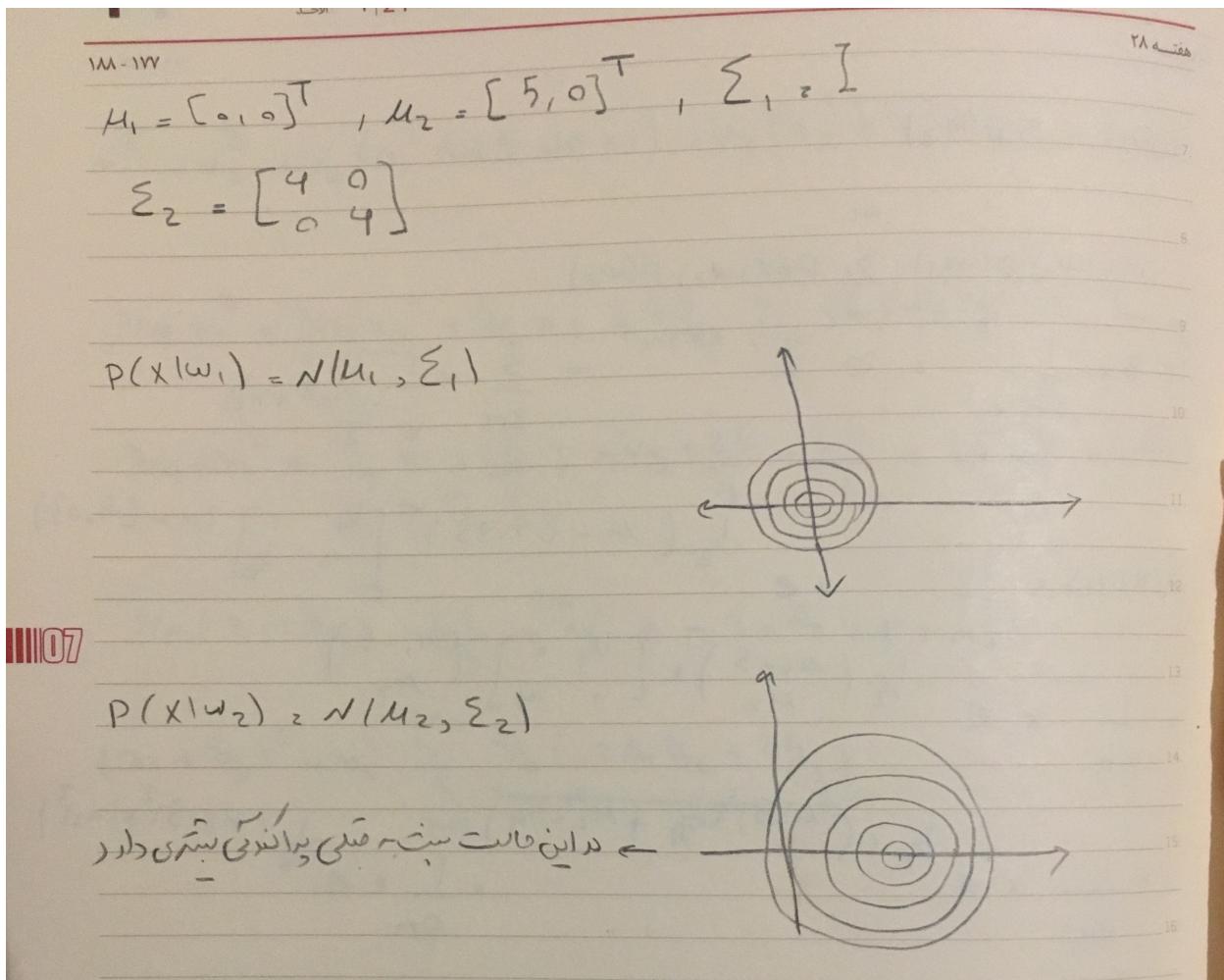
confusion matrix	error
<pre>[[10 0 0 0 0 0 0 0 0 0] [0 10 0 0 0 0 0 0 0 0] [0 0 10 0 0 0 0 0 0 0] [0 0 0 10 0 0 0 0 0 0] [0 0 0 0 10 0 0 0 0 0] [0 0 0 0 0 10 0 0 0 0] [0 0 0 0 0 0 10 0 0 0] [0 0 0 0 0 0 0 10 0 0] [0 0 0 1 0 0 0 0 9 0] [0 0 0 0 0 0 0 0 10]]</pre>	0.01

E)

در این قسمت الگوریتم MDC در دسته‌بندی عکس‌ها با خطای خیلی پایین و نزدیک به صفر عمل کرده است. در این روش با جابجایی نوشته در صفحه و در مجموعه داده‌های بد خط خطای الگوریتم بالا می‌رود. می‌توان گفت یکی از ضعف‌های آن عدم استخراج ویژگی و تصمیم‌گیری با مقایسه بردارها است. همچنین مقایسه بردارها با ابعاد بالا زمانبر است.

Q5)

a)



b)

TA - 2020
1A9 - 1V8

$$\begin{aligned}
 & P(w_1) = 3 P(w_2) \\
 & P(x|w_1) P(w_1) \underset{w_1}{\geq} P(x|w_2) P(w_2) \\
 & \hookrightarrow 3 \times \frac{1}{2\pi} \times e^{-\frac{1}{2} \times (x)^T \mu} = \frac{3}{2\pi} \times \frac{1}{e^{-\frac{(x_1^2 + x_2^2)}{2}}} \\
 & P(x|w_2) = \frac{1}{2\pi} \times e^{-\frac{1}{2} (\mu - [5, 0]^T)^T \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{4} \end{bmatrix} (\mu - [5, 0])} \\
 & = \frac{1}{8\pi} \times e^{-\frac{1}{2} \left(\frac{\mu_1 - 5}{\mu_2} \right)^2 \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} \mu_1 - 5 \\ \mu_2 \end{bmatrix}} \\
 & = \frac{1}{8\pi} \times e^{-\frac{1}{2} \times \left(\frac{(\mu_1 - 5)^2}{\mu_2} \times \frac{1}{4} + \mu_2^2 \times \frac{1}{4} \right)} = \frac{1}{8\pi} \times e^{-\frac{1}{8} ((x_1 - 5)^2 + (x_2)^2)} \\
 & = \frac{3}{2\pi} \times e^{-\frac{1}{2} (x_1^2 + x_2^2)} = 12 \times e^{-\frac{1}{8} (x_1^2 + x_2^2 - \frac{1}{4} (\mu_1 - 5)^2 - \frac{1}{4} \mu_2^2)} \\
 & \frac{1}{8\pi} \times e^{-\frac{1}{8} ((x_1 - 5)^2 + x_2^2)} \geq 12 \times e^{-\frac{1}{4} (x_2^2)}
 \end{aligned}$$

c)

$\omega_1 = \omega_2$

$$\omega_1^2 + \omega_2^2 - \frac{1}{4} (\omega_1^2 + 25 - 10\omega_1) - \frac{1}{4} (\omega_2^2)$$

$$\frac{3}{4}\omega_1^2 + \underbrace{\frac{3}{4}\omega_2^2}_{\omega_1^2} + 5/2\omega_1 - \frac{25}{4} \gtrless \frac{-2 \times \ln 1/2}{\omega_1}$$

$$\frac{3}{4}(\omega_1^2 + \frac{10}{3}\omega_1 + \frac{25}{9}) - \frac{1}{3} \times \frac{25}{4} - \frac{25}{4} + \omega_2^2 \gtrless \frac{-2 \ln 1/2}{\omega_1}$$

07

$$\frac{3}{4}(\omega_1 + 5/3)^2 - \frac{25}{3} + \omega_2^2 \gtrless \frac{-2 \times \ln 1/2}{\omega_1}$$

$$(\omega_1 + 5/3)^2 + \omega_2^2 \gtrless \frac{\frac{4}{3}(-2 \ln 1/2 + 25/3)}{R^2}$$

لکه دلیل ها به صورت زیر مذکورند

آنکه ω_1, ω_2 ناصل بودند و $\omega_1 = \omega_2$ است

d)

در این حالت چون خطای تشخیص فرد راستگو به دروغگو هزینه بیشتری دارد، انتظار می‌رود ناحیه تشخیص کلاس اول (دروغگو) بیشتر به سمت چپ و پایین حرکت کند و ناحیه کمی از کلاس اول به عنوان کلاس دوم تشخیص داده شود. اگر مرز ناحیه دایره‌ای شکل باشد مرکز دایره بیشتر به سمت چپ و پایین حرکت می‌کند و شعاع آن کاهش می‌یابد. و اگر ناحیه با خط راست جدا شود این خط x_1 به سمت چپ و منفی محورها حرکت می‌کند.

e)

Handwritten notes for problem e) showing the calculation of energy levels and entropy.

Given $E_1 \varepsilon_{\mu} (S=1/2) = \sqrt{3} \times e^{-\delta(1/2)} = \sqrt{3} \times e^{-2.72}$

$\delta(1/2) = 1/4 + \frac{25}{5/2} + 1/2 \ln \left(\frac{5/2}{4}\right)^2$

$\delta(1/2) = 5/4 + 1/2 \ln \frac{25}{16} = 5/4 + \ln 5/4 = 2.72$

f)

$$\begin{aligned}
 F) \quad \mathbb{E}_M &= P(w_1) \frac{s}{P(w_2)} \int_{-\infty}^{+\infty} P(X|w_1) \cdot P(X|w_2) dx \\
 &= 3 P(w_2) \frac{s}{P(w_2)} e^{-\delta(s)} \\
 \delta(s) &= s \frac{(1-s)}{2} \times [5, 0]^T \times \left[\frac{1}{2} \times \begin{vmatrix} 4-3s & 0 \\ 0 & 4-3s \end{vmatrix} \right]^{-1} \times [5, 0] \\
 &+ \frac{1}{2} \ln \left| \frac{\begin{vmatrix} 4-3s & 0 \\ 0 & 4-3s \end{vmatrix}}{(16)^{1-s}} \right| = \\
 \delta(s) &= s(1-s) \times \frac{2.5 \times 2}{(4-3s)} + \frac{1}{2} \ln \frac{(4-3s)^2}{(16)^{1-s}} \\
 \mathbb{E}_M &= 3 P(w_2) \times e^{-\delta(s)} \\
 E) \quad \mathbb{E}_M(s=1/2) &= \sqrt{3} \times e^{-\delta(1/2)} = \sqrt{3} \times e^{-2.72} \\
 \delta(1/2) &= 1/4 \times \frac{2.5}{5/2} + 1/2 \ln \frac{(5/2)^2}{4} \\
 \delta(1/2) &= 5/2 + 1/2 \ln \frac{25}{16} = 5/2 + \ln 5/4 = 2.72
 \end{aligned}$$

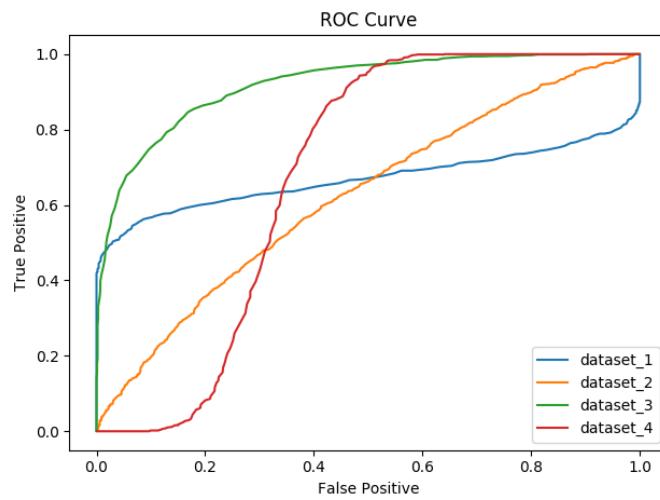
Q6)

A)

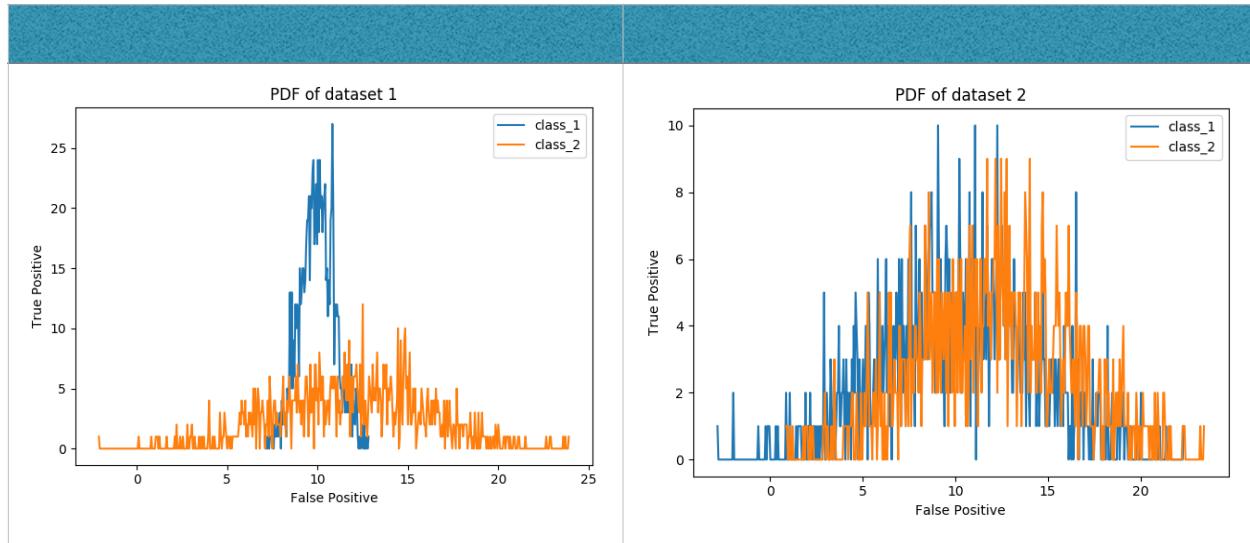
discriminability measure:

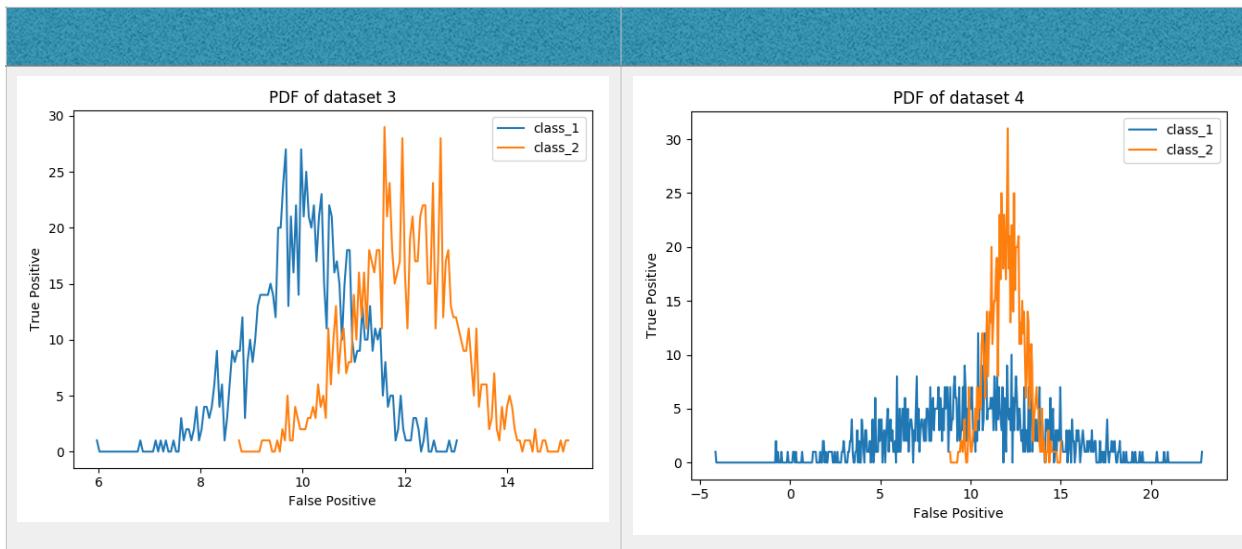
dataset_1	dataset_2	dataset_3	dataset_4
0.4454876	0.3127626	1.3950806	0.4622313

B)



C)





d)

این معیار در واقع فاصله‌ی بین میانگین دو توزیع و نسبت آن به پراکندگی آن دو معیار را محاسبه می‌کند. در واقع این متريک به ازاي فاصله‌ی زياد بین میانگين دو توزيع و پراکندگي داده‌هاي کم مقدار بيشتری می‌گيرد. پس اين متريک هر چه دو توزيع از هم فاصله داشته باشند و هم پوشانی آنها کم تر باشد مقدار بيشتری می‌گيرد. طبق نتایج هم میبینم که در حالتهایی که فاصله توزیع‌های بیشتر و همپوشانی آنها کمتر است هم مقدار مساحت زیر نمودار ROC و هم مقدار discriminability آنها بیشتر است.

Q7)

A)

Prior_false	Prior_true
0.2219	0.7780

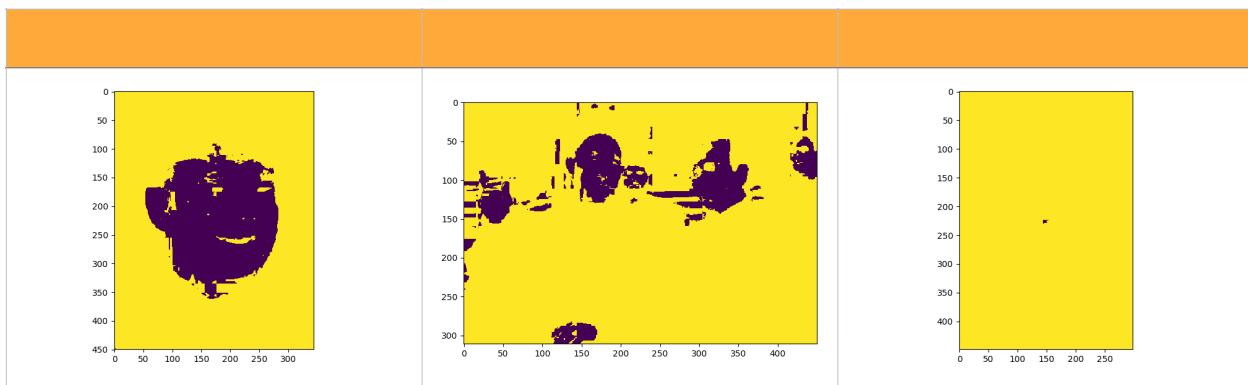
B)

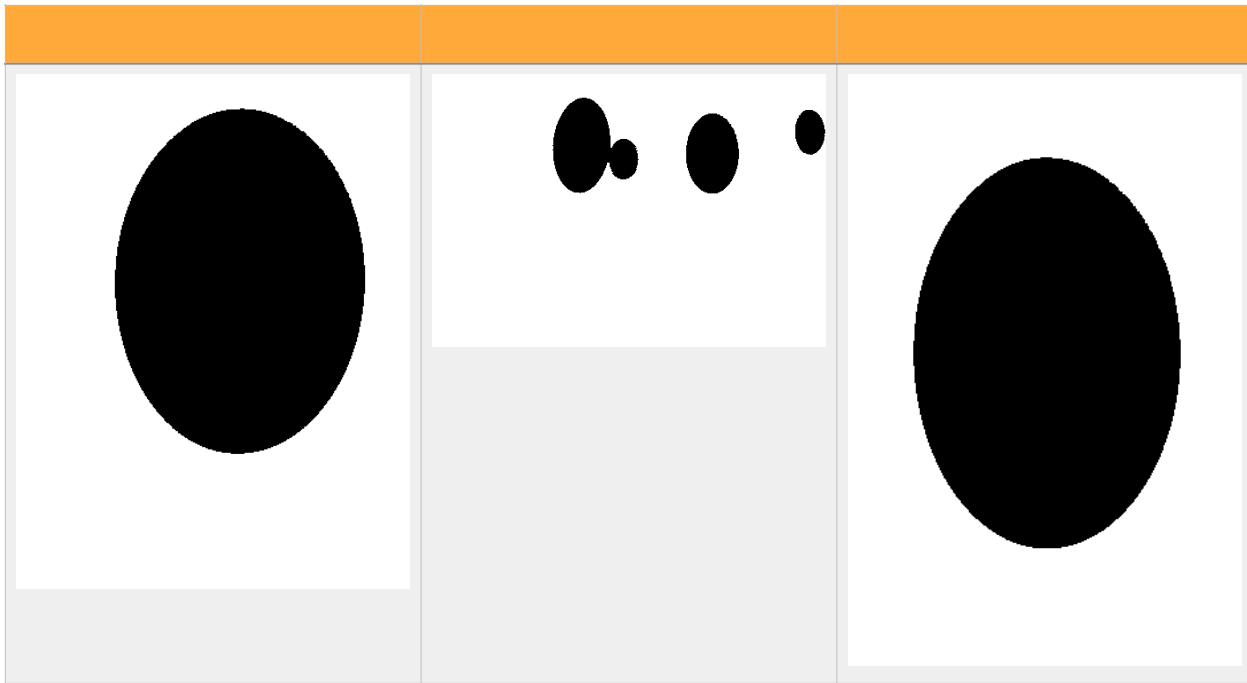
class	cov	mean
TRUE	[[5579.94668384 5204.81783809 4512.04066769] [5204.81783809 5452.54203556 5042.57303635] [4512.04066769 5042.57303635 5370.3082715]]	[100.97083102072469, 94.53559124002433, 91.13386508319815]
FALSE	[[3166.13940873 2621.49270677 2363.9468919] [2621.49270677 2488.74111368 2329.58480857] [2363.9468919 2329.58480857 2390.05702394]]	[155.40323621558971, 118.02923035376072, 99.62844027592578]

C and D)

طبق نتایج میبینیم که در عکس اول تقریباً تا حد خوبی چهره شخص تشخیص داده شده است. در عکس دوم دست فرد نیز به عنوان چهره تشخیص داده شده است. چون توزیع رنگ پوست دست و صورت یکسان است، دسته‌بند دست فرد را نیز به عنوان صورت تشخیص می‌دهد. دسته‌بند با یادگیری توزیع رنگها چهره فرد را تشخیص می‌دهد که این باعث شده به دلیل تعداد کم افراد سیاه پوست در مجموعه داده آموزش احتمال کمی برای آن در توزیع چهره‌ها به نسبت توزیع دیگر در نظر گرفته شود و به این دلیل چهره فرد سوم تشخیص داده نشده است.

Confusion matrix	Test Error
[[41898 68755] [12259 306539]]	81014/429451 = 0.188645503212241

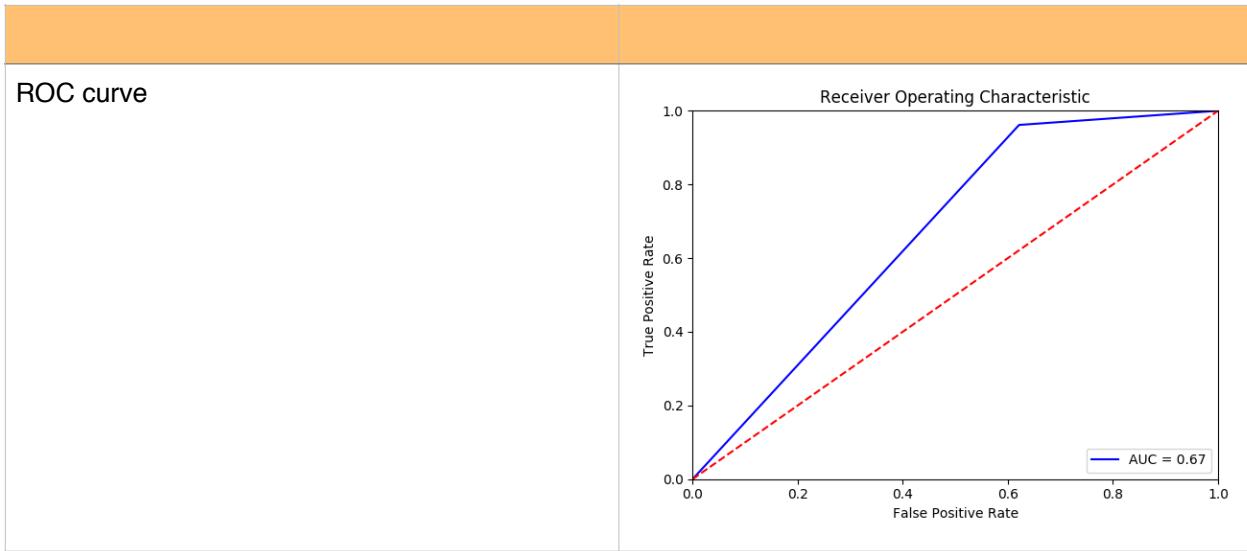




E)

bayes error: 0.188645503212241

F)



g)

این دسته‌بند چون با یادگیری توزیع رنگ‌ها عمل می‌کند، در مجموعه داده‌ای مثل داده سوم که فرد سیاه پوست است و تعداد افراد سیاه پوست در مجموعه داده آموزش آن کم است، نمی‌تواند چهره فرد را تشخیص بدهد. همچنین با دیدن هر توزیع رنگی مانند چهره فرد، آن را چهره تشخیص می‌دهد که این امر باعث شده دست فرد را هم چهره تشخیص دهد.

Q8)

A) yes. goal of regression is to find coefficient B in $y = BX + e$. By using bayesian decision rule we can draw y from a probability density function like:

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

model parameters can also come from a distribution, that means our goal is to find the most probable parameters for given input and output:

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

<https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7>

B) we know A and B are independent in universe space and restricting the A and B to smaller universe doesn't change their dependency.

**** A and B are independent that means whether B is given or not doesn't change the probability of A. Similarly, A and B independent given C means that when C occurs, whether we are further given B or not wouldn't change the probability of A [$P(A|B,C) = P(A|C)$] then we have:

$$\begin{aligned} P(AB|C) &= P(ABC)/P(C) \\ &= P(A|BC) * P(BC) / P(C) \\ &= P(A|BC) * P(B|C) * P(C) / P(C) \\ &= P(A|BC) * P(B|C) -^{***} \\ &P(A|C) * P(B|C) \end{aligned}$$

C)

Bayesian techniques are optimal with respect to a specific metric, namely expected utility. No-Free-Lunch was proved for a very specific cost function, namely misclassification error rate. For cost functions other than misclassification error rate it is possible to have *a priori* distinctions between learning algorithms. Wolpert wrote: "if the error function induces a geometrical structure over Y then we can have *a priori* distinctions between learning algorithms." Although misclassification error rate can be represented as a specific utility function which does not impose such a geometric structure, the broader class of utility functions can impose such a structure and are therefore outside of the realm in which No-Free-Lunch holds.

Thus, there is an optimal off training set distribution over classes even given the uniform prior, and that is the uniform posterior. For decision making it is just as important to express uncertainty as it is to express what is known. Even the uniform posterior contains important information for making decisions. This also means that there is an optimal classifier for off

training set examples even over the space of all functions, namely any classifier that returns the uniform distribution (as Bayes does if a Multinomial uniform prior is used). However, a Bayesian algorithm will also be optimal with respect to expected utility for any prior over functions.

Reference:

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.142.7564&rep=rep1&type=pdf>

D) different classifiers may pay attention differently to each feature which results in different decisions. for example consider two centers $c_1 = [0,1]$ and $c_2=[1,0]$ and point $p = [2,1]$ manhattan and cosine distance of this point from centers are as below:

manhattan:

$c_1: 2$

$c_2: 2$

which means C_2 and C_1 are at the same distance from P .

cosine:

$c_1 = 1 - 1/\sqrt{5}$

$c_2 = 1 - 2/\sqrt{5}$

which means C_2 is more nearer to P than C_1

E) Yes, result of this classifier for classes with fixed distribution is always the same. If posterior and prior probabilities are the same then the classifier would make the same decision.

F) In Bayes classifiers training phase is just measuring the conditional probability of each class, and in MDC training phase is finding the center of each class.