

Question 1)

Gradient tree boosting: در مدل روش gradient boosting یک تکنیک در یادگیری ماشین است. کاربری regression و Classification استفاده می شود. باعث تولید مجموعه ای از decision tree ها می شود. مدل ایجاد شده مرحله به مرحله (stage-by-stage) اضافه می شود (مانند دیگر روش های boosting). در ساختن این مدل، معمولاً از روش CART استفاده می شود.

Random forest: (یا Random decision trees) که در زمان training، مجموعه ای از decision tree ها می سازد و در خروجی، class را که برابر با mode کلاس های ساقه است، برگه می نهد. در Classification، mean prediction درخت های ساقه شده تا این مرحله را برگه می نهد. در Regression، تفاوت: تفاوت این دو مدل ساخت decision trees، ۲ عامل است:

① چگونگی ساقه شدن درخت ها: در Random forest، هر درخت به صورت مستقل از دیگر درخت ها ساقه می شود ولی در روش Gradient boosting، در هر مرحله، فقط یک درخت را می تواند بسازد و تا ساخت درخت تمام نشده، نمی تواند درخت بعدی را بسازد، پس درخت ها به هم وابستگی دارند. پس این نوع ساختن فرم Stage-wise forward دارد. همین باعث بهبود صحایب مدل های موجود قبلی می شود. ✓

② ترکیب نتایج: در Random forest، نتایج در انتهای پروسه با هم ترکیب می شوند (به وسیله میانی شدن و یا "majority rules"). ولی در روش gradient boosting، در طول پروسه، نتایج با هم ترکیب می شوند.

در tuning مناسبی انجام شود، روش gradient boosting می تواند عملکرد بهتری از روش Random forest داشته باشد اما اگر Noise زیادی داشته باشد، روش gradient boosting خیلی کار ساز نمی باشد چون می تواند باعث overfitting شود. همین tuning این روش، سخت تر از روش gradient boosting است.

هم کلام از این روش ها در جاهای متفاوتی استفاده می شوند. برای مثال Random forest برای multi-class object detection کاربرد دارد چون در این دو کاربرد، Noise زیادی مواجه می شود.

bioinformatics (D)

ولی روش Gradient boosting برای جاهایی که unbalanced-data داریم مانند real-time risk assessment کاربرد بیشتری دارد.

برای حل این سوال، باید Entropy را در هر level و root درخت تعیین حساب کنیم. و از آنجایی که در درخت تعیین از بالا به پایین، Entropy باید کاهش پیدا کند، (تا بتوانیم به label های نهایی برسیم) پس آنکه برای سنجش یک level خاص از پایین درخت، Entropy افزایش پیدا کرده یعنی در طبقه بندی سبدها، یک مشکل وجود داشته است.

(۱) level 1 :

$$\begin{matrix} \text{orange} & \text{green} \\ 10 & + 7 \end{matrix} \Rightarrow E(r_1) = - \sum_{i=1}^N p_i \log p_i = - \frac{10}{17} \log \frac{10}{17} - \frac{7}{17} \log \frac{7}{17} = \boxed{0.9192}$$

(۲) level 2 :

$$\begin{matrix} \text{orange} & \text{green} \\ 1 & + 0 \end{matrix} \Rightarrow E(r_1) = - \frac{1}{1} \log \frac{1}{1} - 0 = \boxed{0}$$

$$\begin{matrix} \text{orange} & \text{green} \\ 2 & + 5 \end{matrix} \Rightarrow E(r_2) = - \frac{2}{7} \log \frac{2}{7} - \frac{5}{7} \log \frac{5}{7} = \boxed{0.913}$$

(۳) level 3 :

$$\begin{matrix} \text{orange} & \text{green} \\ 1 & + 4 \end{matrix} \Rightarrow E(r_1) = - \frac{1}{5} \log \frac{1}{5} - \frac{4}{5} \log \frac{4}{5} = \boxed{0.7173}$$

$$\begin{matrix} \text{orange} & \text{green} \\ 1 & + 3 \end{matrix} \Rightarrow E(r_2) = - \frac{1}{4} \log \frac{1}{4} - \frac{3}{4} \log \frac{3}{4} = \boxed{0.9166} *$$

(۴) level 4 :

$$\begin{matrix} \text{orange} & \text{green} \\ 1 & + 2 \end{matrix} \Rightarrow E(r_1) = - \frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3} = \boxed{0.9175} * \Rightarrow \text{آنتروپی تغییر یافته ها و level 4 ها به دلیل اینکه به label رسیده ایم، برابر صفر است.}$$

همان طور که در بالا می بینیم، آنتروپی سبدها در level 2 برابر 0.913 است. ولی آنتروپی سبدها در level 3 برابر 0.9166 شده است یعنی به سمت پایین درخت رفته ایم ولی آنتروپی سبدها در level 4 برابر 0.9175 شده است که در صورت سوال بیان شده است. از طرفی این اتفاق در سبدهای level 4 درخت هم مشاهده شده که آنتروپی اش برابر 0.9175 شده است و {0.913 < 0.9166} و {0.913 < 0.9175} Entropy(level 2 & 3) > Entropy(level 2) < Entropy(level 4)

Question 3)

$TP \Rightarrow$ True Positive
 $TN \Rightarrow$ True Negative
 $FP \Rightarrow$ False Positive
 $FN \Rightarrow$ False Negative

$$\Rightarrow accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

مثال) فرض کنیم در یک سیستم spam filtering نتایج زیر را به دست آورده ایم:

(Original)	Classified Positive	Classified Negative
Positive class	10 \Rightarrow TP	15 \Rightarrow FN
Negative class	25 \Rightarrow FP	100 \Rightarrow TN

$$\Rightarrow total\ cases = TP + TN + FP + FN = 150$$

$$\Rightarrow accuracy = \frac{10 + 100}{10 + 100 + 25 + 15} = 87.3\%$$

معنی اینست که کلاس (Classifier) بسیار خوبی داریم زیرا ۱۳۵ از ۱۵۰ ایمیل های Spam را شناسایی کرده است. که درست فکر کرده ایم. ولی آنده همین metric را به این کلاس classifier می شناسیم که استفاده کنیم به نام accuracy

فرض ورودی ها را به عنوان "No Spam" می کند (یعنی همه ورودی ها فقط در یک کلاس خاص قرار می دهند) خواهم داشت:

(original)	Classified Positive	Classified Negative
Positive class	0 \Rightarrow TP	25 \Rightarrow FN
Negative class	0 \Rightarrow FP	150 \Rightarrow TN

$$\Rightarrow total\ cases = 150 \rightarrow \text{ما تبه بالا}$$

$$\Rightarrow accuracy = \frac{0 + 150}{0 + 0 + 150 + 25} = 86.7\%$$

در اینجا از یک Classifier بسیار بدتر از بالا $accuracy$ بهتر از قبل شد!! استفاده کردیم ولی

این اتفاق، accuracy Paradox می گویند. در واقع ما می بینیم $TP < FP$ باشد، آن گاه با تغییر classification rule به گونه ای که کلاس "Negative" را به عنوان خروجی بدهد، $accuracy$ به صورت گمراه کننده ای بالا می رود!! و نیز برعکس هم صادق است. یعنی اگر $TN < FN$ شد و classification rule را به گونه ای تغییر دهیم که کلاس "Positive" را به عنوان خروجی بدهد، باز هم $accuracy$ (به صورت گمراه کننده ای) افزایش می یابد!! راه حل اینست که از سایر metric ها استفاده کنیم مانند:

$$recall = \frac{TP}{TP + FN} \quad (1)$$

$$precision = \frac{TP}{TP + FP} \quad (2)$$

(توجه) Precision و recall یک trade off وجود دارد و ما باید با توجه به کاربرد خود، تعادل بین این دو را انتخاب کنیم.

$$F_1\text{-score} = \frac{(1 + \beta^2) \times precision \times recall}{(\beta^2 \times precision + recall)} \quad (3)$$

$$F1\text{-score} = \frac{2 \times precision \times recall}{precision + recall} \quad (4)$$

Confusion Matrix

ص ۴

Question 4)

۲۵ classifiers

$e = 0,۳۵$

\Rightarrow به این که یک record را misclassify اشتباه تشخیص بدهیم یا درست طبقه بندی کنیم و در واقع باید حداقل نصف classifiers ها اشتباه کنند.

$P(wrong) = \binom{N}{1} e^1 (1-e)^{N-1} =$

۲۵ بار classifier

در توضیحات بالا خواص را

کمی به جایی از $i=1$ تا $i=۲۵$ می توانستیم از $i=1$ تا $i=۲۵$ را حساب کنیم.

$P(wrong) = \sum_{i=1}^{۲۵} \binom{۲۵}{i} e^i (1-e)^{۲۵-i} = 0,۵۴۰۴۴$

چون ما به دن این مقدار به صورت دستی طولانی می شود. این جواب را با کدن در پایتون در فایل "question 4-theory.py" در فایل "supporting Materials" قرار داده ام.

① $E(\text{Edible}) = H(\omega, A) = -\frac{6}{16} \log \frac{6}{16} - \frac{9}{16} \log \frac{9}{16} = 0,9403 = \text{class}$

Question 5)

Edible:

	Poisonous	Edible	
① Habitat :			
woods	3	2	$\Rightarrow E(\text{Habitat} = \text{woods}) = -\frac{3}{5} \log \frac{3}{5} - \frac{2}{5} \log \frac{2}{5} = 0,971 \rightarrow$
Grasses	0	4	$\Rightarrow E(\text{Habitat} = \text{Grasses}) = -\frac{0}{5} \log \frac{0}{5} - \frac{4}{5} \log \frac{4}{5} = 0$
Leaves	2	3	$\Rightarrow E(\text{Habitat} = \text{Leaves}) = E(3, 2) = 0,971$

یا اینکه می توانیم بر حسب \log حساب کنیم و بعد جواب نهایی را $(\log 2) \div$ کنیم.

$$\Rightarrow E(\text{Edible} | \text{Habitat}) = P(\text{Habitat} = \text{woods}) E(\text{Edible} | \text{Habitat} = \text{woods}) + P(\text{Habitat} = \text{Grasses}) E(\text{Edible} | \text{Habitat} = \text{Grasses}) + P(\text{Habitat} = \text{Leaves}) E(\text{Edible} | \text{Habitat} = \text{Leaves}) = \frac{5}{16} (0,971) + \frac{4}{16} (0) + \frac{5}{16} (0,971) = 0,4944$$

$$\Rightarrow \text{Information Gain} = E(\text{Edible}) - E(\text{Edible} | \text{Habitat}) = 0,9403 - 0,4944 = 0,4459$$

Edible:

	Poisonous	Edible	$= E(2, 2)$
② Cap color:			
Red	2	2	$\Rightarrow E(\text{cap color} = \text{Red}) = -\frac{2}{4} \log \frac{2}{4} - \frac{2}{4} \log \frac{2}{4} = 0,9999$
Green	2	4	$\Rightarrow E(\text{cap color} = \text{Green}) = -\frac{2}{6} \log \frac{2}{6} - \frac{4}{6} \log \frac{4}{6} = 0,9112$
white	1	3	$\Rightarrow E(\text{cap color} = \text{white}) = -\frac{1}{4} \log \frac{1}{4} - \frac{3}{4} \log \frac{3}{4} = 0,9112$

$E(\text{cap color} = \text{Red})$ $E(\text{dible} | \text{cap color} = \text{Red})$

$$\Rightarrow E(\text{dible}, \text{cap color}) = P(\text{Red}) E(\text{cap color} = \text{Red}) + P(\text{Green}) E(\text{cap color} = \text{Green}) + P(\text{white}) E(\text{cap color} = \text{white})$$

$E(\text{dible} | \text{cap color}) = I(\text{cap color}) = \text{cap color}$ $\sqrt{\text{class}} = \text{Edible}$ $\sqrt{\text{class}}$

$$= \frac{4}{16} (0,9999) + \frac{4}{16} (0,9112) + \frac{4}{16} (0,9112) = 0,9109$$

$$\Rightarrow \text{Information Gain} = E(\text{Edible}) - E(\text{dible}, \text{cap color}) = 0,9403 - 0,9109 = 0,0294$$

Edible:

	Poisonous	Edible	
③ Cap shape:			
flat	4	3	$\Rightarrow E(\text{cap shape} = \text{flat}) = -\frac{4}{7} \log \frac{4}{7} - \frac{3}{7} \log \frac{3}{7} = 0,9102$
convex	1	4	$\Rightarrow E(\text{cap shape} = \text{convex}) = -\frac{1}{5} \log \frac{1}{5} - \frac{4}{5} \log \frac{4}{5} = 0,7219$

$$\Rightarrow E(\text{Edible}, \text{cap shape}) = P(\text{flat}) E(\text{Edible}, \text{cap shape} = \text{flat}) + P(\text{convex}) E(\text{Edible}, \text{cap shape} = \text{convex}) =$$

$$\frac{4}{7} (0,9102) + \frac{1}{7} (0,7219) = 0,8814$$

$$\Rightarrow \text{Information Gain} = 0,9403 - 0,8814 = 0,0589$$

۲

Edible:

⊙ Odor:

	Poisonous	Edible
None	۲	۴
Foul	۳	۳

$$\Rightarrow E(\text{odor} = \text{None}) = -\frac{2}{7} \log_2 \frac{2}{7} - \frac{4}{7} \log_2 \frac{4}{7} = 0.1113$$

$$\Rightarrow E(\text{odor} = \text{Foul}) = -\frac{3}{7} \log_2 \frac{3}{7} - \frac{3}{7} \log_2 \frac{3}{7} = 0.9999$$

$\text{میان} = E(2, 3) = 1$

$$\Rightarrow E(\text{Edible, odor}) = P(\text{None}) E(\text{Edible, odor} = \text{None}) + P(\text{Foul}) E(\text{Edible, odor} = \text{Foul}) =$$

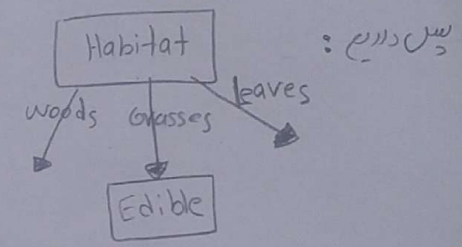
$$\frac{1}{14} (0.1113) + \frac{4}{14} (0.9999) = 0.1921$$

$$\Rightarrow \text{Information Gain} = 0.9503 - 0.1921 = 0.7582$$

پس باقی‌مانده Information Gain در سطح Habitat از دست می‌رود، پس Habitat به عنوان root درخت تصمیم انتخاب می‌شود.

= برای Woods و Leaves با تقسیم بندی هر یکی را باید کنیم. برای این منظور جدول را به دو قسمت

Habitat	...	Habitat	...
Leaves		woods	
...		...	
Leaves		woods	



و برای هر جدول، class یکی تعیین کرده و مثلاً به ترتیب برای جدول ۱ و جدول ۲ برابر با woods و Leaves می‌شود. و بعد باید Information Gain را برای هر کدام از جدول‌ها و ماتریس قسمت قبل حساب کنیم. پس:

⊙ Class = woods = Class

$$\Rightarrow E(\text{Edible}) = H(3, 2) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.971$$

$\text{"woods" به نسبت نقیض} E(\text{Edible})$

⊙ Cap color:

	Poisonous	Edible
(۲) Red	۲	۰
(۱) Green	۱	۱
(۱) white	۰	۱

$$\Rightarrow E(\text{cap color} = \text{Red}) = 0$$

$$\Rightarrow E(\text{cap color} = \text{Green}) = 1$$

$$\Rightarrow E(\text{cap color} = \text{white}) = 0$$

و cap color
= برای "woods"
باقی‌مانده است.

$$\Rightarrow E(\text{Edible, cap color}) = P(\text{Red}) E(\text{Edible, cap color} = \text{Red}) + P(\text{Green}) E(\text{Edible, cap color} = \text{Green}) + P(\text{white}) E(\text{Edible, cap color} = \text{white}) = \frac{2}{5} (0) + \frac{1}{5} (1) + \frac{1}{5} (0) = 0.14$$

$\text{Information Gain} = E(\text{Edible, cap color} | \text{woods}) = 1(\text{cap color}) = \text{ماتریس "woods" فقط}$

$$\Rightarrow \text{Information Gain} = 0.971 - 0.14 = 0.831$$

cap shape:

		Edible:	
		Poisonous	Edible
(Lr) flat		2	0
(Lr) convex		0	2

$\Rightarrow E(\text{cap shape} = \text{flat}) = 0$
 $\Rightarrow E(\text{cap shape} = \text{convex}) = 0$

$\Rightarrow E(\text{Edible}, \text{cap shape}) = P(\text{flat}) E(\text{Edible}, \text{cap shape} = \text{flat}) + P(\text{convex}) E(\text{Edible}, \text{cap shape} = \text{convex})$
 $= \frac{2}{4}(0) + \frac{2}{4}(0) = 0$
 $\Rightarrow \text{Information Gain} = 0.971 - 0 = 0.971$

odor:

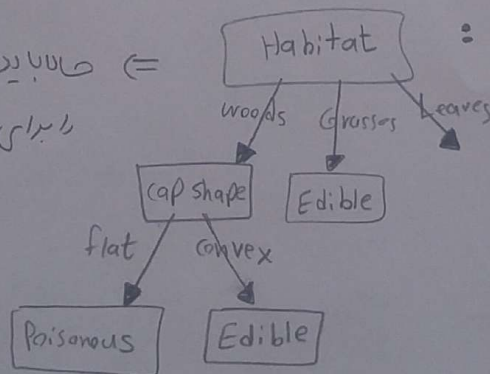
		Edible:	
		Poisonous	Edible
(Lr) None		2	1
(Lr) foul		1	1

$\Rightarrow E(\text{odor} = \text{None}) = -\frac{2}{4} \log_2 \frac{2}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.9183$
 $\Rightarrow E(\text{odor} = \text{foul}) = 1$

$\Rightarrow E(\text{Edible}, \text{odor}) = P(\text{None}) E(\text{Edible}, \text{odor} = \text{None}) + P(\text{foul}) E(\text{Edible}, \text{odor} = \text{foul}) =$
 $\frac{3}{4}(0.9183) + \frac{2}{4}(1) = 0.941$

$\Rightarrow \text{Information Gain} = 0.971 - 0.941 = 0.030$

سبب با معیار Information Gain برای انتخاب "woods" به عنوان "cap shape" و "Leaves" به عنوان "Habitat" انتخاب کنیم. (سبب درج)



① class = Leaves = class

$\Rightarrow E(\text{Edible}) = E(2, 2) = -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 0.971$
 Leaves (Lr)

cap color:

		Edible:	
		Poisonous	Edible
(Lr) Red		—	—
(Lr) Green		1	2
(Lr) white		1	1

$\Rightarrow E(\text{cap color} = \text{Green}) = E(1, 2) = 0.9183$
 $\Rightarrow E(\text{cap color} = \text{white}) = 1$

10

$\Rightarrow E(\text{Edible}, \text{cap color}) = P(\text{Red})E(\text{Edible}, \text{cap color} = \text{Red}) + P(\text{Green})E(\text{Edible}, \text{cap color} = \text{Green}) + P(\text{White})E(\text{Edible}, \text{cap color} = \text{White})$
 $= \frac{3}{5}(0,9113) + \frac{2}{5}(1) = 0,951$

$\Rightarrow \text{Information Gain} = 0,951 - 0,951 = 0,000$

Edible:

Cap shape:	Poisonous	Edible
(i) flat	1	1
(ii) Convex	1	2

$\Rightarrow E(\text{cap shape} = \text{flat}) = E(1,1) = 1$
 $\Rightarrow E(\text{cap shape} = \text{convex}) = E(1,2) = E(2,1) = 0,9113$

$\Rightarrow E(\text{Edible}, \text{cap shape}) = P(\text{flat})E(\text{Edible}, \text{cap shape} = \text{flat}) + P(\text{convex})E(\text{Edible}, \text{cap shape} = \text{convex})$
 $= \frac{3}{5}(1) + \frac{2}{5}(0,9113) = 0,951$

$\Rightarrow \text{Information Gain} = 0,951 - 0,951 = 0,000$

Edible:

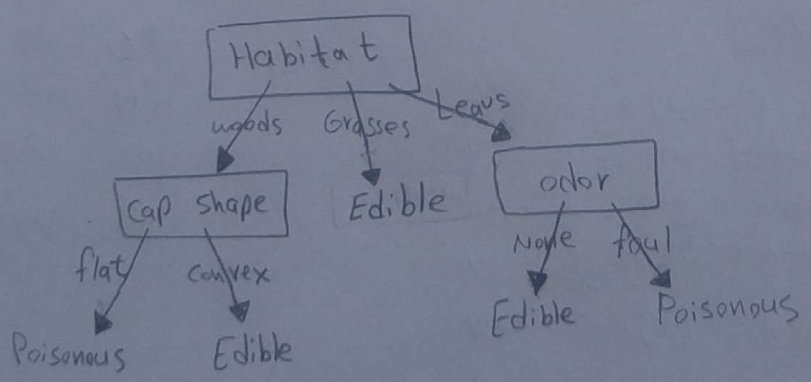
odor:	Poisonous	Edible
(i) None	0	2
(ii) foul	2	0

$\Rightarrow E(\text{odor} = \text{None}) = 0$
 $\Rightarrow E(\text{odor} = \text{foul}) = 0$

$\Rightarrow E(\text{Edible}, \text{odor}) = P(\text{None})E(\text{Edible}, \text{odor} = \text{None}) + P(\text{foul})E(\text{Edible}, \text{odor} = \text{foul})$
 $= \frac{3}{5}(0) + \frac{2}{5}(0) = 0$

$\Rightarrow \text{Information Gain} = 0,951 - 0 = 0,951$

پس با مقایسه Information Gain های بدست آمده برای "Leaves" در سطح "odor" به عنوان root = root
 باید انتخاب شود. پس درخت نهایی به صورت زیر در فواصل آمده است:



9

Question 6)

Bayes rule

chain rule

independent
a, b, c

$$1) P(y|a,b,c) = \frac{P(a,b,c|y) P(y)}{P(a,b,c)} = \frac{P(a|y) P(b|y) P(c|y) P(y)}{P(a) P(b) P(c)}$$

2)

P(each class)

طبخ بدول $Y=ok$ فواهد بدول و افقات بدول
 $\begin{cases} A = \text{false} \text{ برای} \\ B = \text{True} \\ C = \text{false} \end{cases}$

$$P(c_j) : \begin{cases} P(Y=ok) = \frac{\omega}{\lambda} = 0,44\omega \\ P(Y=bad) = \frac{\nu}{\lambda} = 0,33\nu \end{cases}$$

P(each feature | each class)

$$P(x|c_j) : \begin{cases} P(A=false|Y=ok) = \frac{\xi}{\omega} = 0,1 \\ P(A=true|Y=ok) = \frac{1}{\omega} = 0,1 \\ P(A=false|Y=bad) = \frac{1}{\nu} = 0,33 \\ P(A=true|Y=bad) = \frac{\gamma}{\nu} = 0,44 \\ \hline P(B=false|Y=ok) = \frac{\mu}{\omega} = 0,4 \\ P(B=true|Y=ok) = \frac{\gamma}{\omega} = 0,1 \\ P(B=false|Y=bad) = \frac{1}{\nu} = 0,33 \\ P(B=true|Y=bad) = \frac{\gamma}{\nu} = 0,44 \end{cases}$$

$$\begin{cases} P(C=false|Y=ok) = \frac{\xi}{\omega} = 0,1 \\ P(C=true|Y=ok) = \frac{1}{\omega} = 0,1 \\ P(C=false|Y=bad) = \frac{\gamma}{\nu} = 0,44 \\ P(C=true|Y=bad) = \frac{1}{\nu} = 0,33 \end{cases}$$

$$X = \{A = \text{false}, B = \text{true}, C = \text{false}\}$$

$$\Rightarrow P(x|c_j): \begin{cases} P(x|Y=ok) = P(A=\text{false}|Y=ok) \times P(B=\text{true}|Y=ok) \times P(C=\text{false}|Y=ok) = 0,1 \times 0,1 \times 0,1 = 0,001 \\ P(x|Y=bad) = P(A=\text{false}|Y=bad) \times P(B=\text{true}|Y=bad) \times P(C=\text{false}|Y=bad) = 0,1 \times 0,1 \times 0,1 = 0,001 \end{cases}$$

$$\Rightarrow P(x|c_j) \times P(c_j): \begin{cases} P(x|Y=ok) P(Y=ok) = 0,001 \times 0,42 = 0,00042 \\ P(x|Y=bad) P(Y=bad) = 0,001 \times 0,37 = 0,00037 \end{cases}$$

$$\Rightarrow P(x|Y=bad) P(Y=bad) < P(x|Y=ok) P(Y=ok) \quad \text{در نتیجه "Y=ok" به عنوان label انتخاب می شود}$$

\Rightarrow Posterior Probability :

$$P(Y|X=\{A=\text{false}, B=\text{true}, C=\text{false}\}) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$P(X) = P(A=\text{false}) \times P(B=\text{true}) \times P(C=\text{false}) = \frac{1}{8} \times \frac{1}{2} \times \frac{1}{2} = 0,03125$$

Naive - Bayes

$\Rightarrow A, B, C \rightarrow \text{independent}$

$$\Rightarrow P(Y=ok|X) = \frac{P(X|Y=ok) P(Y=ok)}{P(X)} = \frac{0,00042}{0,03125} = 0,01344$$

$$\Rightarrow P(Y=bad|X) = \frac{P(X|Y=bad) P(Y=bad)}{P(X)} = \frac{0,00037}{0,03125} = 0,01184$$

$\Rightarrow P(Y=bad|X) < P(Y=ok|X)$ پس چون $P(Y=ok|X) > P(Y=bad|X)$ ، احتمال اینکه "Y=ok" باشد برای آن بیشتر است. label "Y=ok" به عنوان label انتخاب می شود.