به‌نام خدا

تمرین سری دوم

مریم رضوانی
شماره دانشجویی: ۹۹۲۱۱۶۰۰۱۹

1) به سؤالات زیر پاسخ دهید.

الف)f1-score و Accuracy یک معیارارزیابی برای مدل‌های categorical است.برای هر دو معیار، هر چه مقدار بالاتر باشد، یک مدل بهتر می‌تواند مشاهدات را به کلاس‌ها طبقه‌بندی کند.
*f1-score*به این‌صورت عمل می‌کند که *میانگین هارمونیک دقت* (*precision*) *و* (*recall*) *است*. دقت نسبت تعداد پیش بینی های مثبت صحیح به تعداد کل پیش بینی های مثبت است.
f1-score = 2 * recall * precision / recall + precision
recall نسبت تعداد پیش بینی های مثبت صحیح به تعداد کل مثبت های واقعی است.این معیار نشان می دهد که چه درصد از داده های مثبت را مدل به درستی شناسایی کرده است. **این معیار زمانی مناسب است که داده ها توزیع نامتوازنی از کلاس ها داشته باشند و هزینه ی خطای منفی و مثبت متفاوت باشد.**
اما *accuracy نسبت تعداد پیش بینی‌های درست به تعداد کل پیش بینی ها است.*
Accuracy = TN+TP / TN+TP+FN+FP
این معیار نشان می دهد که چه درصد از داده ها را مدل به درستی دسته بندی کرده است. این معیار زمانی مناسب است که داده ها توزیع متوازنی از کلاس ها داشته باشند و هزینه‌ی خطای منفی و مثبت یکسان باشد.

به عنوان یک قاعده کلی: ما اغلب زمانی از Accuracy استفاده می‌کنیم که کلاس‌ها متعادل هستند و هیچ نقطه ضعف عمده‌ای برای پیش‌بینی منفی‌های کاذب وجود ندارد. **ما اغلب از زمانی‌f1-score استفاده می‌کنیم که کلاس‌ها نامتعادل هستند و پیش‌بینی منفی‌های کاذب جنبه منفی جدی دارد.**
به عنوان مثال، اگر از یک مدل رگرسیون لجستیک برای پیش‌بینی اینکه آیا فردی سرطان دارد یا نه استفاده کنیم، منفی‌های کاذب واقعاً بد هستند (مثلاً پیش‌بینی اینکه کسی سرطان ندارد در حالی که واقعاً سرطان دارد) بنابراین‌f1-score مدل‌هایی را که دارای منفی کاذب بیش از حد هستند جریمه می‌کند. بیش از Accuracy .

ب) بین sensitivity و specificity در یک confusion matrix این است که:

- sensitivity: تقسیم پیش‌بینی‌های درست رده مثبت(TP) بر تعداد کل نمونه‌های مثبت موجود در مجموعه داده‌ها (TP+FN) به‌دست می‌آید. برای مثال برای تشخیص بیماران قلبی در یک مدل؛ نسبت بیماران قلبی که درست تشخیص داده شده‌اند(TP)به مجموع تعداد بیمارانی که درست تشخیص داده شده‌اند(TP) و بیمارانی که درست تشخیص داده نشده اند(FN).

همچنین به عنوان نرخ مثبت واقعی(TPR)شناخته می‌شوند.

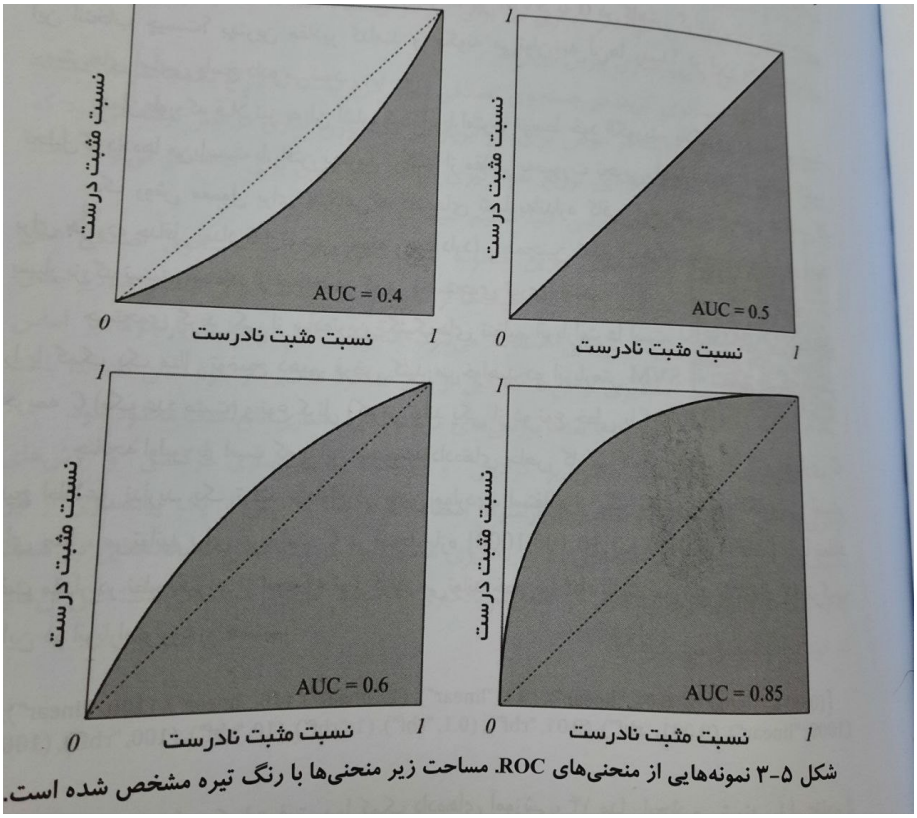
- Specificity: پیش‌بینی درست تعداد منفی های رده منفی(TN) بر تعداد کل منفی های واقعی (TN+FP) است. برای نمونه درآزمایش افراد بیماران قلبی؛ تعداد افرادی که مدل توانسته سالم بودنشان(بیماری ندارند)را تشخیص دهد(TN). به نسبت مجموع افرادی که درست تشخیص دادیم (بیمار نیستند)سالم اند(TN) و افرادی که مدل نتوانسته سالم بودنشان را تشخیص دهد (به اشتباه بیمار تشخیص داده) است(FP).

همچنین به عنوان نرخ منفی واقعی(TNR) شناخته می‌شوند.

- نرخ مثبت کاذب(FPR):یعنی چه تعداد از افراد سالم توسط آزمایش به اشتباه بیمار شناسایی شده اند.(Specificity-1)
- نرخ منفی کاذب(FNR):یعنی چه تعداد از افراد بیمار توسط آزمایش به اشتباه سالم شناسایی شده اند.(Sensitivity -1)

ج) برای ارزیابی یک مدل طبقه بندی، می توانیم از منحنی مشخصه عملکرد گیرنده (ROC) و سطح زیر منحنی (AUC) استفاده کنیم. این دو معیار به ما نشان می دهند که چگونه مدل ما در تشخیص کلاس های مثبت و منفی عمل می کند.منحنی‌های ROC ترکیبی از نسبت مثبت درست (TPR) که برابر با Recallاست.و نسبت مثبت نادرست (نسبتی از نمونه‌های منفی که به درستی پیش‌بینی نشده‌اند.)را برای خلق تصویری خلاصه از عملکرد رده‌بندی استفاده می‌کنند.
نسبت مثبت درست(TPR) برابر است با:TPR=TP/TP+FNو نسبت مثبت نادرست (FPR)برابر است با: FPR = FP/FP+TN منحنی‌های ROC تنها برای طبقه بندی هایی استفاده می‌شوند که همراه با پیش‌بینی خود امتیازاتی (با یک احتمال) را برمی گردانند. برای مثال طبقه بندی هایی نظیر درخت تصمیم و شبکه‌های عصبی (و مدل های تلفیقی از درختان تصمیم) را می‌توان با استفاده از منحنی‌های ROC ارزیابی کرد.
برای رسم یک منحنی ROC ابتدا باید بازه خروجی مدل گسسته سازی شود.چنانچه بازه مزبور برای یک مدل برابر با [0,1] باشد آنگاه می‌توان این بازه را بصورت [0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9] گسسته سازی کرد.سپس از هر مقدار گسسته به عنوان آستانه پیش‌بینی استفاده می‌شود و برچسب‌های نمونه‌های موجود در داده‌ها با استفاده از مدل واین آستانه پیش‌بینی می‌شوند. برای مثال اگر می‌خواهید TPR و FPRرا برای حدآستانه ای برابر با 0.7محاسبه کنید؛مدل را روی هرنمونه اعمال کرده و خروجی مدل را دریافت کنید.چنانچه مقدار خروجی مدل عددی بزرگ‌تر یا مساوی 0.7بود شما رده مثبت را پیش‌بینی می‌کنید و در غیر اینصورت پیش‌بینی شما رده منفی است. در شکل زیر به سادگی می‌توان دید که با تنظیم آستانه با مقدار صفر؛ تمام پیش‌بینی‌ها مثبت خواهند بود. و بنابراین مقادیر TPRو FPRنیز برابر با یک خواهد بود.(شکل بالا سمت راست). از طرف دیگر چنانچه مقدار آستانه برابر با یک باشد هیچ یک از پیش‌بینی‌ها مثبت نخواهد بود و مقادیر TPRو FPRنیز برابر با صفر خواهد بود.(شکل پایین سمت چپ). رده بندی که

مساحت زیر منحنی ROCآن بیشتر باشد رده بندی بهتر است. این مساحت با AUC در شکل نشان داده شده است. یک رده بند با مقدار AUC بزرگ‌تر از 0.5 بهتر از یک رده بند تصادفی است. چنانچه مقدار AUCکمتر از 0.5 باشد موردی در مدل شما اشتباه است. مقدار AUCیک رده‌بند کامل برابر با یک است. معمولاً اگر مدل شما رفتاری مناسب داشته باشد با انتخاب مقداری برای آستانه که باعث نزدیک شدن TPRبه یک می‌شود وFPRرا نزدیک به صفر نگه میدارد. به رده بندی خوبی دست پیدا می‌کنید.



شکل ۵-۳ نمونه‌هایی از منحنی‌های ROC. مساحت زیر منحنی‌ها با رنگ تیره مشخص شده است.

د)یک مجموعه اعتبار سنجی جداگانه زمانی مفید است که شما بخواهید پارامترهای مدل خود را بهینه کنید و عملکرد مدل را بر روی داده های جدید ارزیابی کنید. مجموعه اعتبار سنجی به شما این امکان را می‌دهد که مدل خود را بر روی داده هایی که در آموزش دیده نشده اند، امتحان کنید و پارامترهایی را انتخاب کنید که بهترین نتیجه را بدهند. مجموعه آزمایشی فقط برای تخمین خطای نهایی مدل بر روی داده های جدید استفاده می‌شود و نباید در فرآیند بهینه سازی مدل تاثیر گذار باشد. اگر شما فقط از دو مجموعه آزمایشی و آموزشی استفاده کنید، ممکن است خطای مدل را در آزمایش بالاتر یا پایین تر از حقیقت برآورد کنید و عملکرد مدل را به درستی نشان ندهید.

2)

Year: Month: Day: Page : ( )

سوال 2)

$$E_{out} = E(h(x) - y)^2$$

$$E_{out} = \int (h(x) - y)^2 \underbrace{p(x,y)}\, dx\, dy$$

تابع چگالی احتمال مشترک x و y

$$\rightarrow E_{out} = \int (h(x) - E[y|x] + E[y|x] - y)^2 p(x,y)\, dx\, dy$$

با استفاده از ماکتور یعنی $a^2 + b^2 + 2ab$

$$E_{out} = \int (h(x) - E[y|x])^2 p(x,y)\, dx\, dy + \int (E[y|x] - y)^2 p(x,y)\, dx\, dy$$

$$+ 2 \int (h(x) - E[y|x])(E[y|x] - y) p(x,y)\, dx\, dy$$

باتوجه به اینکه جمله دوم برابر با صفر است: زیرا:

$$p(x,y)\, dx\, dy = \int (h(x) - E[y|x])(E[y|x] - y) p(x,y)\, dx\, dy =$$

$$= \int (h(x) - E[y|x]) p(y|x) p(x)\, dy\, x = \int (h(x) - E[y|x]) p([y|x])$$

$$\int p(x)\, dx = 0$$

بنابراین خطای خارج از نمونه را می‌توان به صورت زیر نوشت:

$$E_{out} = \int (h(x) - E[y|x])^2 p(x,y)\, dx\, dy + \int (E[y|x] - y)^2 p(x,y)\, dx\, dy$$

برای کمینه کردن این عبارت باید جمله اول را کمینه کنیم این جمله همواره مثبت است و زمانی صفر می‌شود که

$h(x) = E[y|x]$ باشد بنابراین اگر $h^*(x) = E[y|x]$ باشد $h(x) = E[y|x]$

خطای خارج از نمونه کمینه می‌شود. جمله دوم نیز یا نگر خطای ناشی از تغییرات تصادفی y و y است و به مدل ما

بستگی ندارد.

Senobar

Page : ( )

SUBJECT:

Year: Month: Day:

قسمت بود: برای اثبات $\varepsilon(x)$ مقدار صفر داده، داریم:

$$h^*(x) = E[y|x]$$

$$E[h^*(x)|x] = E[E[y|x]|x] = E[y|x] = h^*(x)$$

که دران $E(x)$ نشان دهنده مقدار مورد انتظار است. داریم:

$$y = h^*(x) + \varepsilon(x)$$

$$E[y|x] = E[h^*(x) + \varepsilon(x)|x] = E[h^*(x)|x] + E[\varepsilon(x)|x]$$

$$= h^*(x) + E[\varepsilon(x)|x] = h^*(x) + E[\varepsilon(x)|x]$$

حاصل $h^*(x)$ از دوطرف معادله میتوانیم بنویسیم:

$$E[\varepsilon(x)|x] = E[\varepsilon(x)|x] = 0$$

بدین معنا که $\varepsilon(x)$ برابر با صفر است

(3

Page : ( )

SUBJECT:

Year: Month: Day:

سوال 3)

• چی معیاری برای مقایسه دو مدل پیشنهاد می کنید؟ (MSE)

• آن را برای این مسئله محاسبه کنید.

(Mean Squerd Error میانگین مجذور مربعات خطا)

مقدار واقعی	مقدار هدف پیش بینی شده
$(Y_1) = 10$	$(\hat{Y}_1) = 12$
$(Y_2) = 20$	$(\hat{Y}_2) = 18$
$(Y_3) = 15$	$(\hat{Y}_3) = 10$
$(Y_4) = 25$	$(\hat{Y}_4) = 28$
$(Y_5) = 30$	$(\hat{Y}_5) = 32$

$$MSE = \frac{1}{2} \sum_{i=1,\dots N} (f(x_i) - y_i)^2$$

$$MSE = \frac{1}{2} \sum_{i=1}^{5} ((\hat{Y}_i^{\circ}) - Y_i)^2 =$$

$$(\hat{Y}_1 - Y_1)^2 = (12-10)^2 = 4 \quad , \quad (\hat{Y}_2 - Y_2)^2 = (18-20)^2 = 4$$

$$(\hat{Y}_3 - Y_3) = (10-15)^2 = 25, \quad (\hat{Y}_4 - Y_4)^2 = (28-25)^2 = 9$$

$$(\hat{Y}_5 - Y_5)^2 = (32-30)^2 = 4$$

$$MSE = \frac{1}{2} \sum_{i=1}^{5} (4+4+25+9+4) = \frac{46}{2} = 23$$

Senobar

4)مدل ۲ بهتر است؛ زیرا حساسیت (sensitivity) بالاتری نسبت به بقیه مدل ها برخوردار است.بدلیل اینکه حساسیت نشان می‌دهد چند درصد از تومورهای بدخیم را مدل به درستی شناسایی کرده است و این برای ما خیلی مهم است زیرا هزینه‌ی FN (false negative)خیلی بالا است.به این معنا که فرض کنید شخصی تومور بدخیم دارد ولی مدل ما آن را سالم پیش‌بینی کرده است. در اینجور مسائل بدنبال دقت یا حساسیت(sensitivity) بالا هستیم.

Accuracy = TN+TP / (TN+TP+FN+FP)

sensitivity = TP/(TP+FN)

specificity = TN/(TN+FP)

با قراردادن مقادیر بالا برای هریک از مدل های زیر بدست می آوریم:

کل داده‌ها را ۱۰۰۰نفر در نظر می‌گیریم.

مدل1:

actual		
Predicted	TP=20	FP=0
	FN=80	TN=900

مدل2:

actual		
Predicted	TP=70	FP=110
	FN=30	TN=790

مدل3:

actual		
Predicted	TP=40	FP=27
	FN=60	TN=873

5) Confusion Matrix:

برای class1 داریم:

		actual		
		class1	class2	class3
Predicted	class1	TP=7	FP=0	FP=4
	class2	FN=0	TN=6	TN=1
	class3	FN=1	TN=0	TN=1

Class1: **TP=7,** **FP=0+4=4,** **FN=0+1=1,** **TN=6+1+1+0=8**

sensitivity=TP/TP+FN=7/(7+1)=7/8
FNR=1-Sensitivity=1-7/8=1/8=12.5%

Accuracy = TN+TP / TN+TP+FN+FP =(7+8)/(7+8+1+4)=15/20
Error Rate=*1-Accuracy*=1-15/20=1/4=0/25=25%

Precision=tp/tp+fp=7/(7+4)=7/11
recall=tp/tp+fn=7/8
F1-Score= 2 * recall * precision / recall + precision
F1-Score=(2*7/8*7/11) / (7/8+7/11)=0/73=73%

برای کلاس ۲ و ۳ هم به همین صورت داریم:

Class2: **TP=6,** **FP=0+1=1,** **FN=0+0=0,** **TN=4+7+1+1=13**
Class3: **TP=1,** **FP=0+1=1,** **FN=4+1=5,** **TN=6+7+0+0=13**

سوال 7) مجموعه داده‌ی آزمون (ورودی‌ها: A, B, C، خروجی: کلاس) همراه با نتایج آزمون دسته‌بندی (خروجی پیش‌بینی شده) داده شده است. دو نقطه روی منحنی ROC برای تمام آستانه‌های 15٪، 18٪ پیدا کرده و رسم کنید.

A	B	C	class	predicted
10	2	A	1	0.97
20	1	B	0	0.61
30	3	A	1	0.77
40	2	B	1	0.91
15	1	B	0	0.12

TP: تعداد داده‌ای که خروجی واقعی آنها 1 است. خروجی پیش‌بینی‌ی آستانه FP: تعداد داده‌ای که خروجی واقعی آنها صفر است و خروجی پیش‌بینی آنها بزرگتر یا مساوی آستانه است.

FN: به معنای تعداد داده‌ای که خروجی واقعی آنها 1 است و خروجی پیش‌بینی شده کوچکتر از آستانه است.

TN: به معنای تعداد داده‌ای که خروجی واقعی آنها 0 است و خروجی پیش‌بینی شده کوچکتر از آستانه است

$$FPR = \frac{FP}{FP+TN} \qquad TPR = \frac{TP}{TP+FN}$$

برای آستانه 15٪ داریم:

$TP=3$, $FN=0$, $FP=1$

$TN=1$ $\quad TPR = \frac{3}{3+0} = 1 \quad FPR = \frac{1}{1+1} = 0.5$

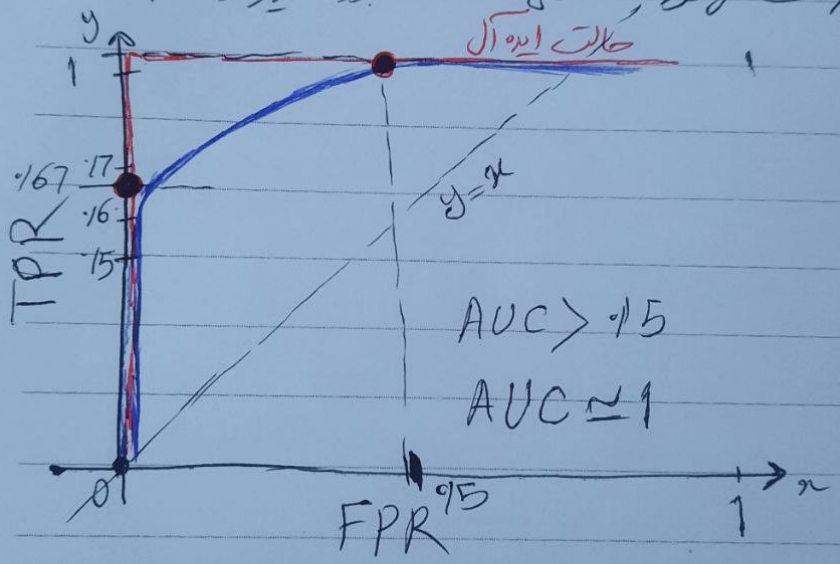
برای آستانه 18٪ داریم: $TP=2$, $FN=1$, $FP=0$, $TN=2$

$$TPR = \frac{2}{2+1} = 0.67 \qquad FPR = \frac{0}{0+2} = 0$$

حالا با رسم نقطه TPR, FPR برای هر آستانه شکل منحنی ROC را بدست می‌آوریم

Subject:
Year : Month : Day : ( )

منحنی ROC شامل نقطه (۰و۰) که به معنای همگان سالم است و نقطه (۱و۱) که به معنا
همگان بیمار است رسم شود. منحنی ROC بصورت زیر فرض شود:



(با انتخاب مقداری برای آستانه که باعث نزدیکشدن TPR به یک می‌شود و FPR را نزدیک
به صفر نگه می‌دارد به رده بندی خوبی دست پیدا می کنید)

8)

الف)

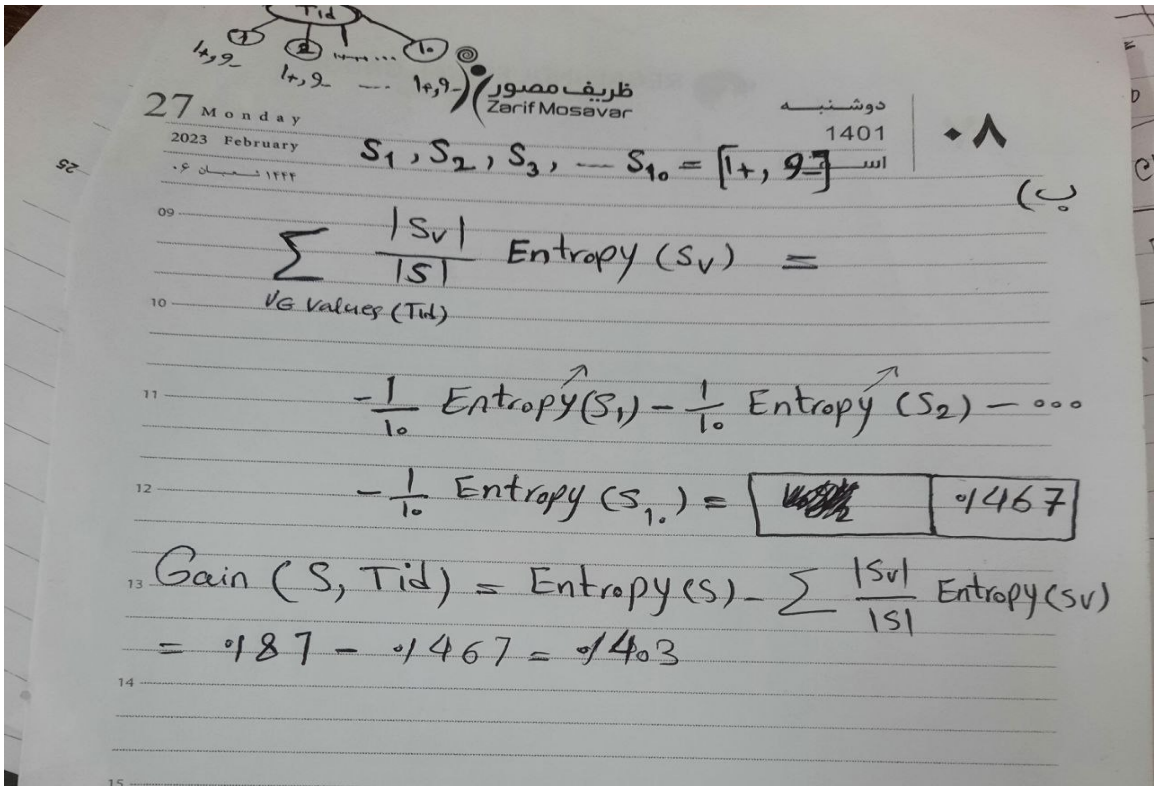
Entropy([3+,7-])=-3/10 $\log_2$ (3/10)+(7/10$\log_2$ 7/10)=0.87

ب)

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

values(Tid)=(1,2,3,4,5,6,7,8,9,10)

S=([3+,7-])



Values (Tid) = 1, 2, 3, 4, 5, 6, 7, 8, 9 $\quad S = [3+, 7-]$

$S_1 \Leftarrow [1+, 0]$, $S_2 \Leftarrow [1-, 0]$, $S_3 = [1-, 0]$, $S_4 = [1-, 0]$, $S_5 = [1+, 0]$ $S_6 = [1-, 0]$

$S_7 \Leftarrow [1-, 0]$, $S_8 \Leftarrow [0, 1+]$, $S_9 = [1-, 0]$, $S_{10} = [0, 1+]$

$\text{Entropy}(S_v) = -\frac{1}{10} \log_2^{\frac{1}{10}} = .1332$ $\quad \sum_{v=1}^{10} \frac{|S_v|}{|S|} \text{Entropy}(S_v) = 10 \times .0332 = .1332$

$.187 - \sum \frac{|S_v|}{|S|} \text{Entropy}(S_v) = .187 - .1332 = .1538$

ج)

Values (Marital Status) = Single, Married , Divorced

$S = [3+, 7-]$

$S_{\text{Single}} \Leftarrow [2-, 2+]$, $S_{\text{married}} \Leftarrow [4-, 0]$

$S_{\text{Divorced}} \Leftarrow [1-, 1+]$

$\text{Gain}(S, \text{Marital Status}) = \text{Entropy}(S) - \sum \frac{|S_v|}{|S|} \text{Entropy}(S_v)$

$= .187 - \frac{4}{10} \text{Entropy}(S_{\text{single}}) - \frac{4}{10} \text{Entropy}(S_{\text{married}})$

$- \frac{2}{14} \text{Entropy}(S_{\text{Divorced}}) =$

روز حمایت از بیماران نادر

Values (Marital Status) = Single, Married, Divorced

$S = [3+, 7-]$

$S_{Single} \Leftarrow [2-, 2+]$, $S_{married} \Leftarrow [4-, 0]$

$S_{Divorced} \Leftarrow [1-, 1+]$

$Gain(S, Marital\ Status) = Entropy(S) - \sum \frac{|S_v|}{|S|} Entropy(S_v)$

$= .187 - \frac{4}{10} Entropy(S_{single}) - \frac{4}{10} Entropy(S_{married}) - \frac{2}{14} Entropy(S_{Divorced}) =$

روز حمایت از بیماران نادر

$Entropy(S_{Single}) = -\frac{2}{10} \log_2^{\frac{2}{10}} - \frac{2}{10} \log_2^{\frac{2}{10}}$

$.146 + .146 = .1924$

$Entropy(S_{married}) = -\frac{4}{10} \log_2^{\frac{4}{10}} + 0 = .1528$

$Entropy(S_{Divorced}) = -\frac{1}{10} \log_2^{\frac{1}{10}} + \frac{-1}{10} \log_2^{\frac{1}{10}} = .1664$

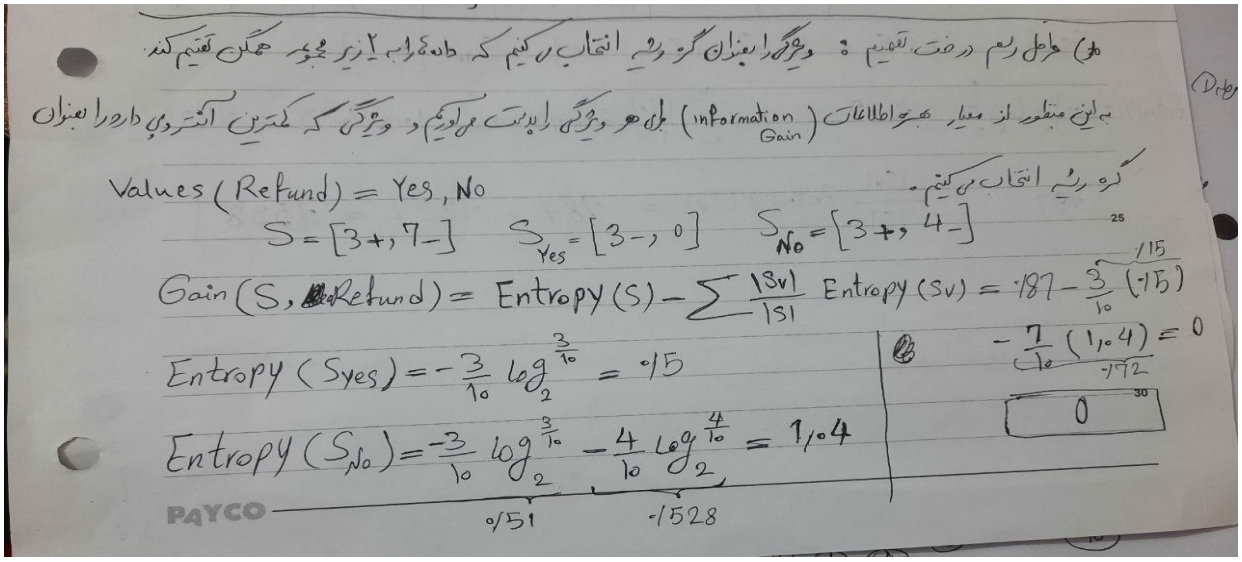
$Gain(S, Martial\ Statabs) = .187 - (\frac{4}{10}) .1924 - (\frac{4}{10}) .1528 - (\frac{2}{10}) .1664 =$

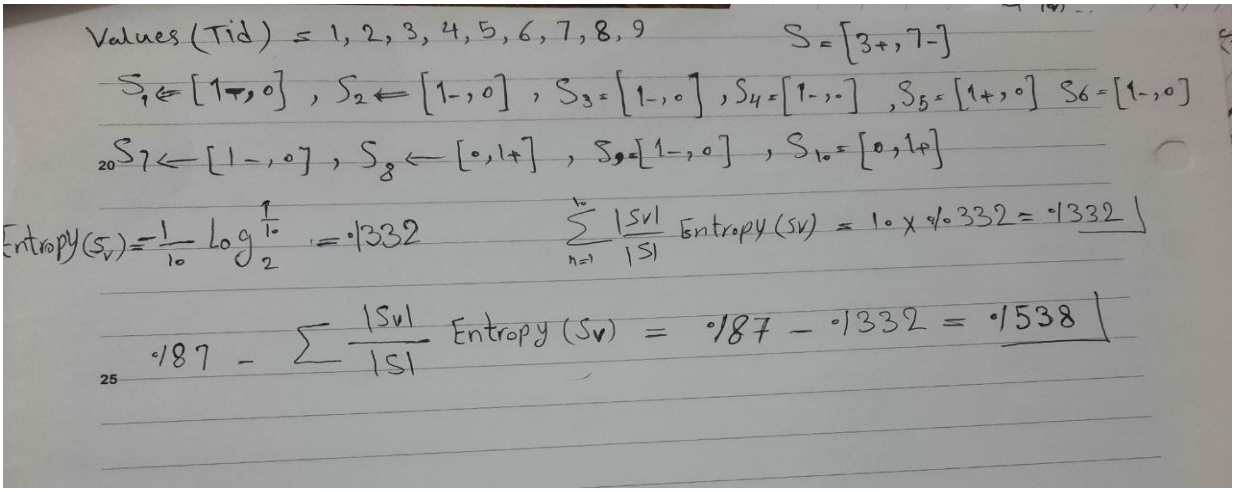
.1369 .1332 .1211

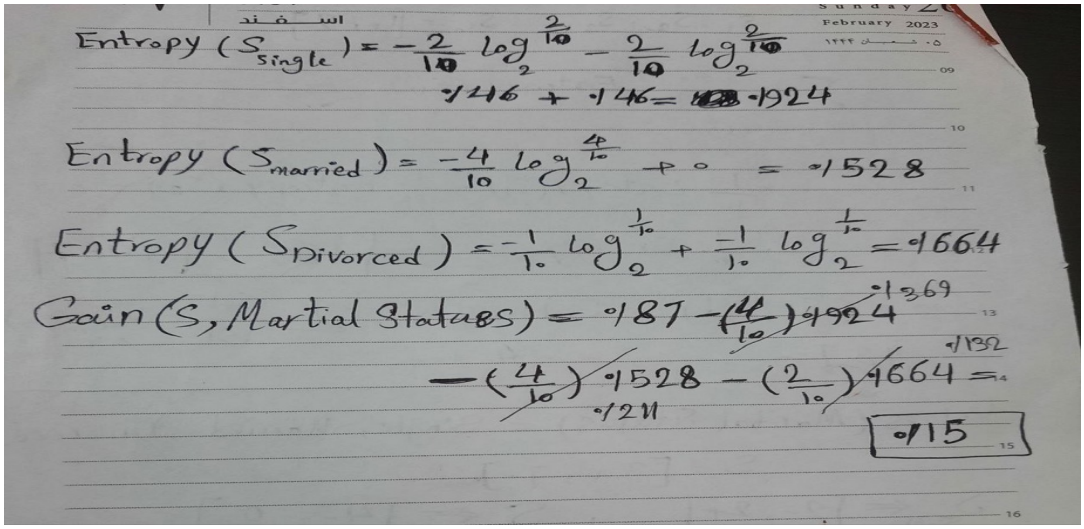
.115

د)

هرچه داده‌ها یکنواخت تر باید به دسته های بیشتر تقسیم کنیم.در اینجا ویژگی آیدی ۱۰ فرزند دارد که در هر فرزند ۱ نمونه قرار می‌گیرد. به همین دلیل برای شرط تست ویژگی مناسب نمی باشد.



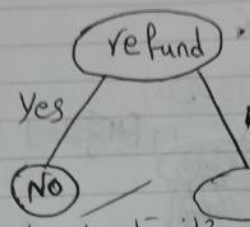




۵

فروردین
پنجشنبه
25 March 2021
۱۱ شعبان ۱۴۴۲

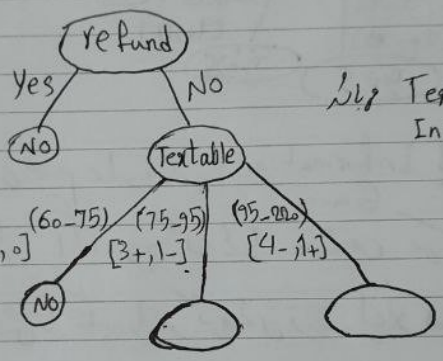
گره Refund با کمترین آنتروپی به عنوان گره ریشه انتخاب می‌کنیم:



مرحله ① این گره نشان دهنده اولین سوال برای تقسیم کسری است که می‌گوید آیا خریدار دارد یا نه

سپس از گره ریشه ۲ شاخه به سمت پایین رسم می‌کنیم و به هر شاخه مقادیر ممکن refund را می‌دهیم. (دقیقاً مقادیر No یا Yes)

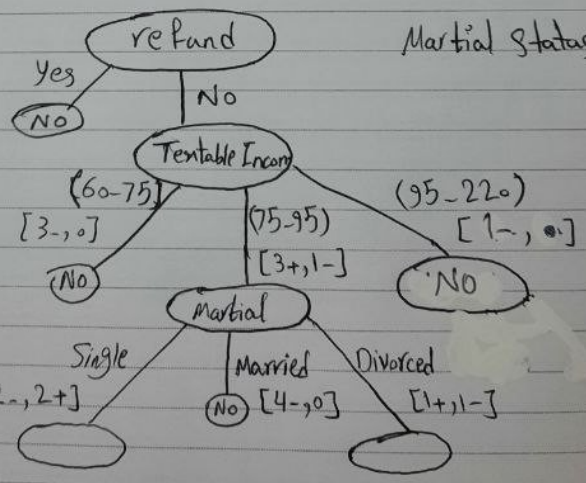
مرحله ②



دیگر بعدی برابر Taxable Income

سپس به ازای هر شاخه برگ می‌کنیم که تعداد مواردی که با آن شاخه مطابقت دارند برچسب کلاس (دارند یا خیر) اگر بله، آن شاخه را به یک گره برگ تبدیل می‌کنیم و برچسب مشترک را روی آن می‌نویسیم. وگرنه، آن شاخه را به یک گره تقسیم دیگر تبدیل می‌کنیم و یک ویژگی دیگر برای آن انتخاب می‌کنیم و همین مراحل را برای هر گره تکرار می‌کنیم.

مرحله ③

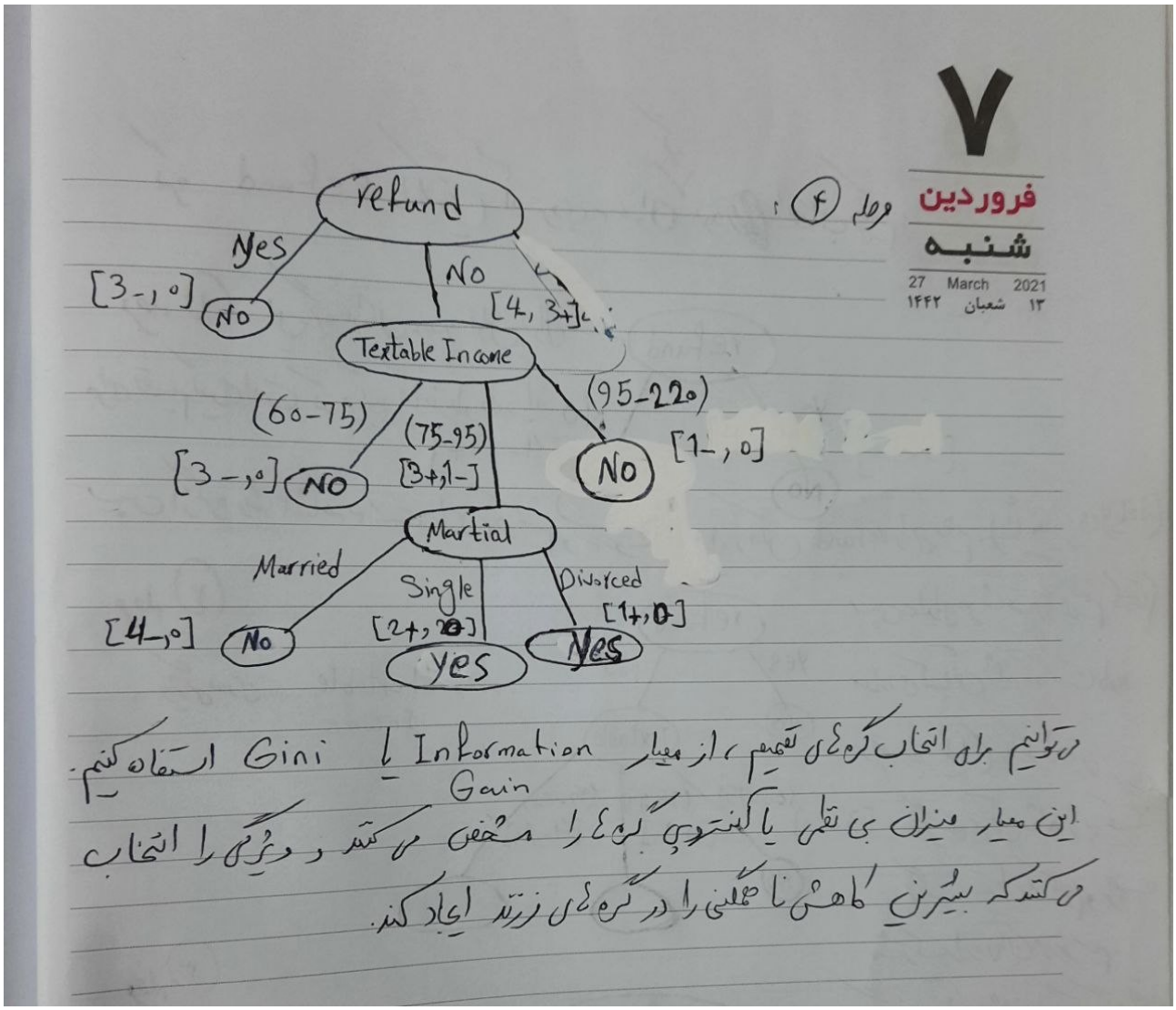


ویژگی Martial Status

۶

فروردین
جمعه
26 March 2021
۱۲ شعبان ۱۴۴۲

۵- ولادت حضرت علی اکبر علیه السلام (۳۳ هـ ق) و روز جوان
۶- ولادت زرتشت



ه) کلاس مربوط به نمونه تست {NO,Married,82k} طبق درخت تصمیم رسم شده در قسمت «ر» دارای لیبل NOمی‌باشد.