به نام خدا

پاسخنامه تمرین اول

استاد درس:

دکتر مریم امیر مزلقانی

نیم سال اول ۱۴۰۲-۱۴۰۳

راه ارتباطی:

Aut.DataMining.Fall@gmail.com

سوال ۱.

الف) $H(\text{Target class}) = H(\frac{4}{9}, \frac{5}{9}) = -\frac{4}{9}\log_2(\frac{4}{9}) - \frac{5}{9}\log_2(\frac{5}{9}) = 0.9911$

Target class

ب) $IG(TC, a_1) = H(TC) - H(TC|a_1) = 0.9911 - \frac{4}{9}H(\frac{3}{4}, \frac{1}{4}) - \frac{5}{9}H(\frac{1}{5}, \frac{4}{5}) = 0.2294$

$IG(TC, a_2) = H(TC) - H(TC|a_2) = 0.9911 - \frac{5}{9}H(\frac{2}{5}, \frac{3}{5}) - \frac{4}{9}H(\frac{2}{4}, \frac{2}{4}) = 0.0072$

ج) از آن‌جایی که مقدار آستانه در اختیار ما است باید آن‌ها به گونه‌ای انتخاب کنیم که برای ما بهتر باشد ← IG را برای آستانه‌های مختلف حساب می‌کنیم

a_2	۱	۳	۴	۵	۶	۷	۸
TC	+	−	+	− −	+	+ −	−
	۲	۳٫۵	۴٫۵	۵۵	۶٫۵	۷٫۵	

* همچنین آستانه باید جایی باشد که class تغییر کند.

۲ → $IG(TC, a_2) = H(TC) - \frac{1}{9}H(1,0) - \frac{8}{9}H(\frac{3}{8}, \frac{5}{8}) = 0.1422$

۳٫۵ → $IG(TC, a_2) = H(TC) - \frac{2}{9}H(\frac{1}{2}, \frac{1}{2}) - \frac{7}{9}H(\frac{3}{7}, \frac{4}{7}) = 0.0026$

۴٫۵ → $IG(TC, a_2) = H(TC) - \frac{3}{9}H(\frac{1}{3}, \frac{2}{3}) - \frac{6}{9}H(\frac{2}{6}, \frac{4}{6}) = 0.0278$

۵٫۵ → $IG(TC, a_2) = H(TC) - \frac{5}{9}H(\frac{2}{5}, \frac{3}{5}) - \frac{4}{9}H(\frac{2}{4}, \frac{2}{4}) = 0.0072$

۶٫۵ → $IG(TC, a_2) = H(TC) - \frac{6}{9}H(\frac{3}{6}, \frac{3}{6}) - \frac{3}{9}H(\frac{1}{3}, \frac{2}{3}) = 0.0183$

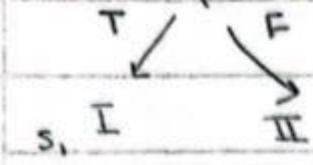
۷٫۵ → $IG(TC, a_2) = H(TC) - \frac{8}{9}H(\frac{4}{8}, \frac{4}{8}) - \frac{1}{9}H(1,0) = 0.1022$

← می‌بینیم که مقداری بهتر از ۹۰ درصد ← آستانه ۲ را انتخاب می‌کنیم

→

a_2	۲	۴
TC	+	− + − − ++ − −

ت) ابتدا a_1 را در ریشه قرار می دهیم.



$$I: H(S_1) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) = 0.8113$$

$$IG(S_1|a_2) = H(S_1) - \frac{1}{4}H(1,0) - \frac{3}{4}H\left(\frac{1}{3},\frac{2}{3}\right) = 0.3113$$

$$IG(S_1|a_3) = H(S_1) - \frac{1}{4}H(1,0) - \frac{3}{4}H\left(\frac{2}{3},\frac{1}{3}\right) = 0.1226$$

← a_2 را انتخاب می کنیم

$$II: H(S_2) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.7219$$

$$IG(S_2|a_2) = H(S_2) - \frac{3}{5}H(1,0) - \frac{2}{5}H\left(\frac{1}{2},\frac{1}{2}\right) = 0.3219$$

$$IG(S_2|a_3) = H(S_2) - 1 \times H\left(\frac{1}{5},\frac{4}{5}\right) = 0$$

← a_2 را انتخاب می کنیم

در شاخه I بیان اینکه $IG(S_1|a_2)$ بیشترین مقدار را میگیرد با آستانه ۰٫۵ می گرفتیم

$$IG(S_1|a_2) = H(S_1) - \frac{2}{4}H\left(\frac{1}{2},\frac{1}{2}\right) - \frac{2}{4}H(1,0) = 0.3113$$

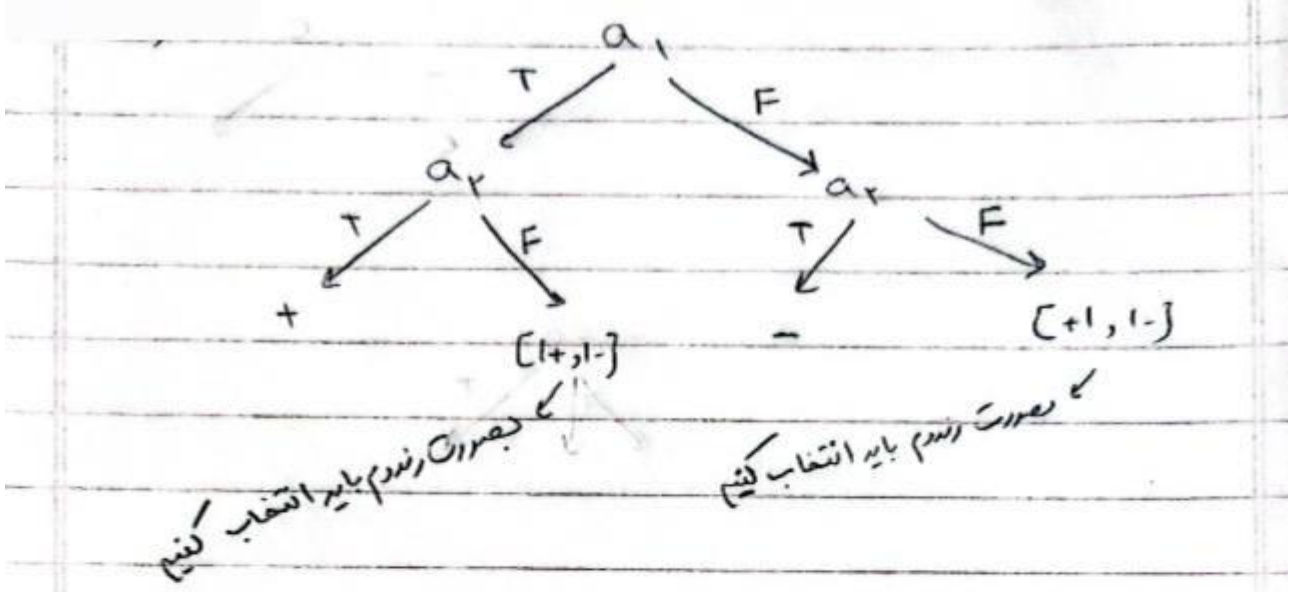
برابر با مقدار $IG(S_1|a_2)$ می شود ← فرقی ندارد کدام را انتخاب می کردیم

بیان شاخه II نیز باید آستانه را ۰٫۵ قرار می دادیم

$$IG(S_2|a_3) = H(S_2) - \frac{2}{5}H\left(\frac{1}{2},\frac{1}{2}\right) - \frac{3}{5}H(1,0) = 0.3219$$

بازهم فرقی نداشت

← درنهایت آستانه را همان ۲ در نظر می گیریم و برای S_1 و S_2 a_2 را درنظر می گیریم

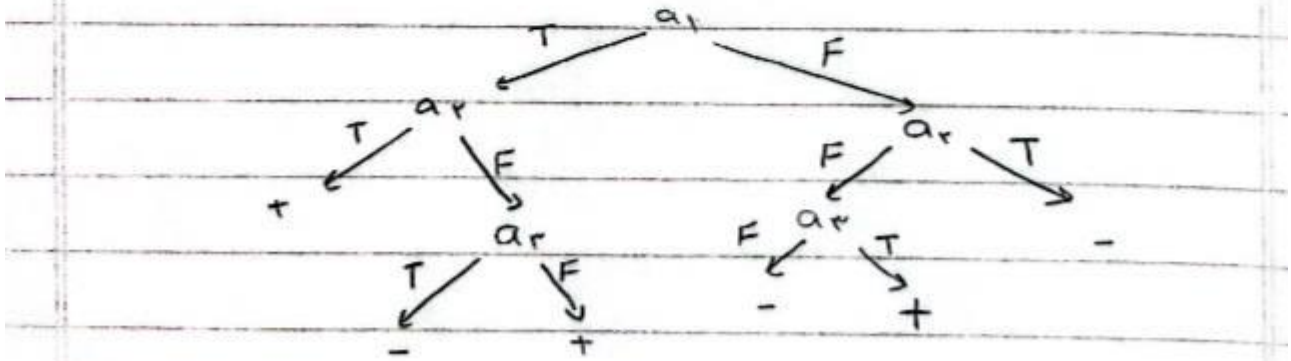


با بررسی این درخت متوجه می شویم اگر آستانه را ۰٫۵ قرار دهیم بطور کامل دسته بندی خواهد شد :

I : $IG(S_1|a_2) = 0.3113$

II : $IG(S_2|a_2) = H(S_2) - \frac{3}{5}H(\frac{1}{3},\frac{2}{3}) - \frac{2}{5}H(1,0) = 0.1710$

a2	۱	۳	۴	۵	۶	۷	۸
	+	-	+	-	+	-+	-
	T				F		



* البته اینگار ممکن باعث overfitting روی داده آموزش شود

سوال ۲.

الف)

اگر به جای مربع خطا از قدرمطلق استفاده گردد مسئله بهینه‌سازی به $\min ||X\beta - y||$ تغییر میکند که باید آن را بدست بیاوریم و قدرمطلق در مقایسه با توان ۲ به مقادیر کوچک تفاضل$X\beta - y$ اهمیت بیشتری داده و به مقادیر بزرگ $X\beta - y$ اهمیت کمتری میدهد. لذا در مقابل داده‌های پرت مقاوم‌تر است و همچنین به سمت صفر و تنک کردن مقادیر$X\beta - y$ میرویم.

ب)

در روش gradient descent از گرادیان و مشتق‌گیری استفاده میشود، اما با در نظر گذفتن قدرمطلق دیگر تابع مشتق‌پذیر نخواهد بود لذا نمیتوان از روش gradient descentاستفاده کرد. اما با توجه به محدب بودن تابع میتوان از subgradient استفاده کرد و عینا روشهای حل قبلی را با subgradient جلو برد.

ج)

به شکل زیر تشکیل میدهیم.Xبرای هر کلاس داده‌ها یک ماتریس

$$X = \begin{bmatrix} 1 & x_{A11} & x_{A21} & x_{A31} & x_{A41} & x_{A51} \\ 1 & x_{A12} & x_{A22} & x_{A32} & x_{A42} & x_{A52} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{A1K} & x_{A2K} & x_{A3K} & x_{A4K} & x_{A5K} \end{bmatrix}$$

ماتریس y را نیز همان داده جدید y در نظر میگیریم. و با حل مسئله بهینه‌سازی min خطا را بدست میاوریم. این کار را برای هر کلاس تکرار میکنیم.(با X متناسب با همان کلاس) سپس هرکدام که داری خطای مینیمم کمتری باشد کلاس داده جدید خواهد بود.

(البته میتوان از رگرسیون شماره کلاس بر حسب داده‌ها هم استفاده کرد که کمتر کلی است)

سوال ۳.

$$x = \begin{bmatrix} 1 & 42 \\ 1 & 74 \\ 1 & 48 \\ 1 & 35 \\ 1 & 56 \\ 1 & 26 \\ 1 & 60 \end{bmatrix}$$

$$y = \begin{bmatrix} 98 \\ 130 \\ 120 \\ 88 \\ 182 \\ 80 \\ 135 \end{bmatrix}$$

$$x^T x\beta = x^T y$$

$$x^T x = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 42 & 74 & 48 & 35 & 56 & 26 & 60 \end{bmatrix} \times \begin{bmatrix} 1 & 42 \\ 1 & 74 \\ 1 & 48 \\ 1 & 35 \\ 1 & 56 \\ 1 & 26 \\ 1 & 60 \end{bmatrix} = \begin{bmatrix} 7 & 341 \\ 341 & 18181 \end{bmatrix}$$

$$x^T y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 42 & 74 & 48 & 35 & 56 & 26 & 60 \end{bmatrix} \times \begin{bmatrix} 98 \\ 130 \\ 120 \\ 88 \\ 182 \\ 80 \\ 135 \end{bmatrix} = \begin{bmatrix} 833 \\ 42948 \end{bmatrix}$$

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 7 & 341 \\ 341 & 18181 \end{bmatrix}^{-1} \times \begin{bmatrix} 833 \\ 42948 \end{bmatrix} = \begin{bmatrix} 44.34 \\ 1.532 \end{bmatrix}$$

$$y = 1.532x + 44.34$$

تخمین فشار خون برای سن ۴۰ :

$$y = 1.532 \times 40 + 44.34 = 105.62$$

سوال ۴.

۱.

دقت آموزش بسیار بالا (۱۰۰٪) است در حالی که دقت آزمون پایینتر است (۵۰٪). این نشان‌دهنده‌ی احتمال زیاد overfitting است. در واقع، مدل به خوبی به داده‌های آموزشی هماهنگ شده است (دقت بالا)، اما نمی‌تواند با داده‌های جدید (آزمون) به خوبی عمل کند، که نشان‌دهنده‌ی overfitting است. بنابراین، در احتمال overfitting بیشتر است، زیرا دقت آموزش بسیار بالا و دقت آزمون بسیار پایین است، که نشان‌دهنده‌ی عدم تعمیم‌پذیری خوب مدل به داده‌های جدید است.

۲.

دقت آموزش و دقت آزمون به نسبت هم نزدیک‌تر هستند (۸۰٪ در مقابل ۷۰٪)، که نشان‌دهنده‌ی یک تطابق بهتر بین عملکرد مدل در داده‌های آموزش و آزمون است. این حالت ممکن است نشانگر یک مدل بهتر و کمتر overfit شده باشد.

سوال ۵.

منظم سازی برای جلوگیری از برازش بیش از حد داده ها، به ویژه زمانی که اختلاف زیادی بین عملکرد مجموعه آموزش و مجموعه تست وجود دارد. با منظم سازی، تعداد ویژگیهای مورد استفاده در تمرین ثابت نگه داشته میشود، اما مقدار ضرایب (w) کاهش مییابد .

Lasso Regression

این یک تکنیک منظمسازی است که در انتخاب ویژگی با استفاده از روش انقباض استفاده می شود که به آن روش رگرسیون جریمه شده نیز گفته می شود Lasso .مخفف Least Absolute Shrinkage and Selection Operatorاست که هم برای منظم سازی و هم برای انتخاب مدل استفاده می شود. این روش عبارت جریمه را به تابع هزینه اضافه می کند. که این عبارت جریمه مجموع مطلق ضرایب است که باعث کاهش مقدار ضرایب به منظور کاهش ضرر میشود. این روش تمایل دارد که ضرایب را به صفر مطلق میل دهد .

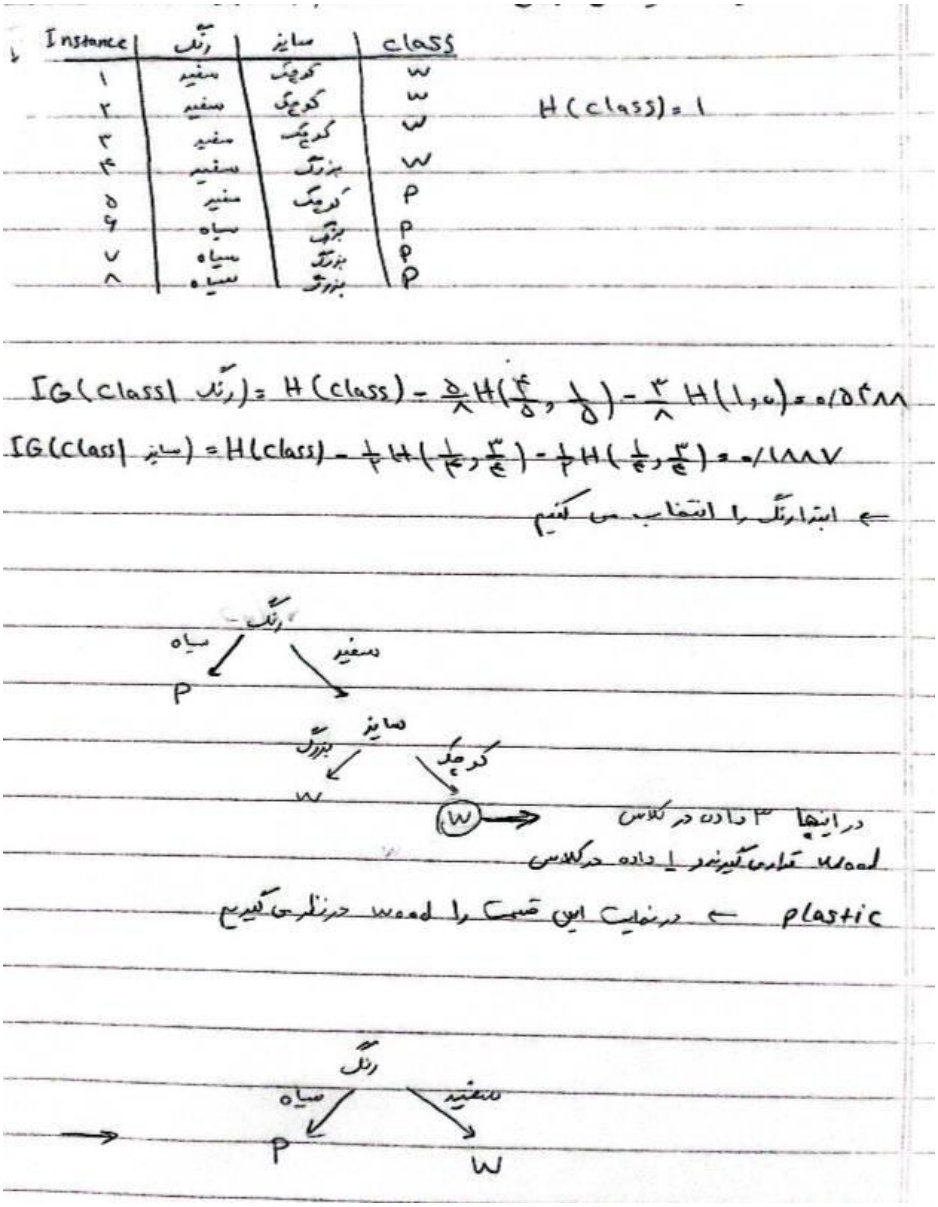
$$L_{lasso} = argmin_{\hat{\beta}} \left(\|Y - \beta * X\|^2 + \lambda * \|\beta\|_1 \right)$$

Ridge Regression

تفاوت این روش با روش قبلی در این است که مقدار جریمه‌ای که به تابع هزینه اضافه میشود برابر با مجذور ضرایب است. بر خلاف Lassoاین روش هیچگاه ضرایب را به سمت صفر مطلق سوق نمیدهد .

$$L_{ridge} = argmin_{\hat{\beta}} \left(\|Y - \beta * X\|^2 + \lambda * \|\beta\|_2^2 \right)$$

سوال ۶.



سوال ۷.

الف)

برای هر کدام از ویژگی‌های درون جدول مقدار Gini|index را محاسبه میکنیم.

Gini Index for Heavy :

$$Gini(Heavy=NO) = 1-\left(\left(\frac{2}{5}\right)^2+\left(\frac{3}{5}\right)^2\right) = \frac{12}{25}$$
$$Gini(Heavy=Yes) = 1-\left(\left(\frac{1}{4}\right)^2+\left(\frac{3}{4}\right)^2\right) = \frac{4}{9}$$

$$Gini(Heavy) = \frac{5}{8}\times\frac{12}{25}+\frac{3}{8}\times\frac{4}{9} = \frac{3}{10}+\frac{1}{6} = .1447$$

Gini Index for Spotted :

$$Gini(Spotted=NO) = 1-\left(\left(\frac{2}{5}\right)^2+\left(\frac{3}{5}\right)^2\right) = \frac{12}{25}$$
$$Gini(Spotted=Yes) = 1-\left(\left(\frac{1}{4}\right)^2+\left(\frac{2}{4}\right)^2\right) = \frac{4}{9}$$

$$Gini(Spotted) = \frac{5}{8}\times\frac{12}{25}+\frac{3}{8}\times\frac{4}{9} = .1447$$

Gini Index for Smooth :

$$Gini(Smooth=NO) = 1-\left(\left(\frac{2}{6}\right)^2+\left(\frac{2}{4}\right)^2\right) = \frac{1}{2}$$
$$Gini(Smooth=Yes) = 1-\left(\left(\frac{1}{4}\right)^2+\left(\frac{3}{4}\right)^2\right) = \frac{3}{8}$$

$$Gini(Smooth) = \frac{1}{2}\times\frac{1}{2}+\frac{1}{2}\times\frac{3}{8} = \frac{1}{4}+\frac{3}{16} = \frac{7}{16} = .4375$$

لذا Smooth انتخاب میگردد چرا که Gini آن کمتر است.

ب)

$$I(parent) = -\frac{3}{8}\log\frac{3}{8} - \frac{5}{8}\log\frac{5}{8} = 0.4238 + 0.3476 = 0.9544$$

$$I(Smooth=NO) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$$

$$I(Smooth=Yes) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.3113 + 0.5 = 0.8113$$

$$Gain(Smooth) = 0.9544 - \left(\frac{1}{2}\times 1 + \frac{1}{2}\times 0.8113\right) = 0.9544 - 0.9056 = 0.0488$$

ج)

Heavy	Spotted	Poisonous
NO	NO	NO
NO	Yes	NO
NO	Yes	Yes
Yes	NO	Yes

Gini for Heavy:

$Gini(Heavy=NO) = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right) = \frac{3}{8}$

$Gini(Heavy=Yes) = 0$

$Gini(Heavy) = \frac{4}{8}\times\frac{3}{8} + \frac{4}{8}\times 0 = \frac{3}{8} = 0.375$

Gini for spotted:

$Gini(Spotted=NO) = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right) = \frac{1}{2}$

$Gini(Spotted=Yes) = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right) = \frac{1}{2}$

$Gini(Spotted) = \frac{1}{4}\times\frac{1}{2} + \frac{3}{4}\times\frac{1}{2} = 0.5$

Heavy انتخاب

Heavy	Spotted	Poisonous
Yes	NO	NO
Yes	NO	Yes
NO	Yes	Yes
NO	NO	Yes

$Gini(Heavy=NO) = 1 - 1 = 0$

$Gini(Heavy=Yes) = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right) = \frac{1}{2}$

$Gini(Heavy) = \frac{1}{4}\times\frac{1}{2} + \frac{1}{4}\times 0 = 0.25$

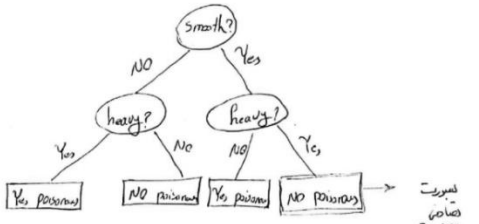
$Gini(Spotted=NO) = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right) = \frac{3}{8}$

$Gini(Spotted=Yes) = 0$

$Gini(Spotted) = \frac{4}{8}\times\frac{3}{8} + \frac{4}{8}\times 0 = 0.375$

Heavy انتخاب

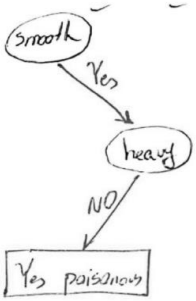
لذا درادامه بار دوم تقسیم را داریم:



د)

داده‌ی نمونه‌ی (a) عبارت است از Smooth=Yes, Spotted=No, Heavy=No

مثال در درخت تصمیم جلو می‌رویم:



با توجه به درخت تصمیم این نوع قارچ سمی خواهد بود