دانشگاه صنعتی خواجه نصیرالدین طوسی
دانشکده مهندسی کامپیوتر

یادگیری ماشین - دکتر ناصرشریف

سوال ۱

بخش a

شاخص جینی [1] برای تمام داده های این جدول به صورت زیر بدست می آید:

$$1 - (\frac{10}{20})^2 - (\frac{10}{20})^2 = \frac{1}{2} \quad (۱)$$

بخش b

در صورتی که داده ها را بر اساس Customer ID دسته بندی کنیم، هر ردیف از جدول جدا می شود. در این صورت هر کدام از داده ها متعلق به یکی از دسته ها خواهند بود. بنابراین برای تمام این دسته ها به صورت جداگانه خواهیم داشت:

$$1 - (\frac{1}{1})^2 - (\frac{0}{1})^2 = 0 \quad (۲)$$

بنابراین شاخص جینی کل داده ها بر اساس شماره مشتری نیز صفر خواهد بود.

بخش c

شاخص جینی برای داده های مربوط به افراد مذکر (یا مونث) به صورت زیر بدست می آید:

$$1 - (\frac{4}{10})^2 - (\frac{6}{10})^2 = 0.48 \quad (۳)$$

پس شاخص جینی برای تمام داده ها بر اساس جنسیت به صورت زیر خواهد بود:

$$\frac{10}{20} \times 0.48 + \frac{10}{20} \times 0.48 = 0.48 \quad (۴)$$

[1] Gini Index

بخش d

شاخص جینی برای ماشین های Sports برابر صفر است. چون تمام این ماشین ها در دسته ی C_0 قرار میگیرند.
شاخص جینی برای ماشین های Family برابرست با:

$$1 - (\frac{1}{4})^2 - (\frac{3}{4})^2 = 0.375 \quad (۵)$$

و برای ماشین های Luxury هم از رابطه زیر بدست می آید:

$$1 - (\frac{1}{8})^2 - (\frac{7}{8})^2 = 0.219 \quad (۶)$$

حالا برای بدست آوردن شاخص جینی کل داده ها بر اساس نوع ماشین داریم:

$$\frac{8}{20} \times 0 + \frac{4}{20} \times 0.375 + \frac{8}{20} \times 0.219 = 0.162 \quad (۷)$$

بخش e

برای سایز Small داریم:

$$1 - (\frac{3}{5})^2 - (\frac{2}{5})^2 = 0.48 \quad (۸)$$

برای سایز Medium داریم:

$$1 - (\frac{3}{7})^2 - (\frac{4}{7})^2 = 0.49 \quad (۹)$$

برای سایز Large داریم:

$$1 - (\frac{2}{4})^2 - (\frac{2}{4})^2 = 0.5 \quad (۱۰)$$

برای سایز Extra Large داریم:

$$1 - (\frac{2}{4})^2 - (\frac{2}{4})^2 = 0.5 \quad (۱۱)$$

حالا برای بدست آوردن شاخص جینی کل داده ها بر اساس سایز پیراهن داریم:

$$\frac{5}{20} \times 0.48 + \frac{7}{20} \times 0.49 + \frac{4}{20} \times 0.5 + \frac{4}{20} \times 0.5 = 0.49 \quad (۱۲)$$

بخش f

بهترین شاخص نوع ماشین است زیرا کمترین مقدار جینی را دارد.

بخش g

چون از آنجا که هر مشتری تازه ای شماره منحصر بفرد و تازه ای دریافت خواهد کرد، دسته بندی براساس این خصیصه باعث می شود درخت تصمیم، قدرت پیش بینی نداشته باشد.

سوال ۲

بخش a

آنتروپی تمام داده های جدول به صورت زیر بدست می آید:

$$-\frac{4}{9} \times \log_2 \frac{4}{9} - \frac{5}{9} \times \log_2 \frac{5}{9} = 0.9911 \quad (۱۳)$$

بخش b

آنتروپی a_1 به صورت زیر بدست می آید:

$$\frac{4}{9}[-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}] + \frac{5}{9}[-\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}] = 0.7616 \quad (۱۴)$$

بنابراین مقدار Information Gain برای a_1 برابر با $0.9911 - 0.7616 = 0.2294$ است.
آنتروپی a_2 به صورت زیر بدست می آید:

$$\frac{5}{9}[-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}] + \frac{4}{9}[-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}] = 0.9839 \quad (۱۵)$$

بنابراین مقدار Information Gain برای a_2 برابر با $0.9911 - 0.9839 = 0.0072$ است.

بخش c

a_3	کلاس	محل تقسیم	آنتروپی	Info Gain
1.0	+	2.0	0.8484	0.1427
3.0	-	3.5	0.9885	0.0026
4.0	+	4.5	0.9183	0.0728
5.0	-	5.5	0.9839	0.0072
5.0	-	5.5	.09839	0.0072
6.0	+	6.5	0.9728	0.0183
7.0	+	7.5	0.8889	0.1022
7.0	-	7.5	0.8889	0.1022

بهترین نقطه برای تقسیم داده ها، 2.0 است.

بخش d

بهترین گزینه a_1 است.

بخش e

برای a_1 مقدار Error Rate برابر $\frac{2}{9}$ است. این عدد برای دسته ی a_2 برابر $\frac{4}{9}$ می باشد. بنابرین a_1 برای تقسیم کردن گزینه بهتری است.

بخش f

برای a_1 مقدار شاخص جینی برابر است با:

$$\frac{4}{9}[1-(\frac{3}{4})^2-(\frac{1}{4})^2]+\frac{5}{9}[1-(\frac{1}{5})^2-(\frac{4}{5})^2]=0.3444 \quad (۱۶)$$

این مقدار برای a_2 برابرست با:

$$\frac{5}{9}[1-(\frac{2}{5})^2-(\frac{3}{5})^2]+\frac{4}{9}[1-(\frac{2}{4})^2-(\frac{2}{4})^2]=0.4889 \quad (۱۷)$$

مجددا a_1 به علت کوچکتر بودن برای تقسیم گزینه بهتری است.

سوال ۳

اگر $Y=\{y_1,y_2,...,y_c\}$ نشانگر c دسته مختلف باشد و $X=\{x_1,x_2,...,x_k\}$ نشانگر k مقدار برای ویژگی X باشد، پیش از آنکه داده ها را برحسب X تقسیم کنیم، آنتروپی برابر است با:

$$E(Y)=-\sum_{j=1}^{c}P(y_j)\log_2 P(y_j)=\sum_{j=1}^{c}\sum_{i=1}^{k}P(x_i,y_j)\log_2 P(y_j) \quad (۱۸)$$

در معادله ۱۸ ما از این حقیقت که بر اساس قوانین احتمال $P(y_j)=\sum_{i=1}^{k}P(x_i,y_j)$ است استفاده کردیم.
بعد از تقسیم، آنتروپی برای هر گره فرزند X که x_i است، برابر است با:

$$E(Y|x_i)=-\sum_{j=1}^{c}P(y_j|x_i)\log_2 P(y_j|x_i) \quad (۱۹)$$

که در آن $P(y_j|x_i)$ کسری از نمونه های دسته x_i است که به کلاس y_j تعلق دارد. آنتروپی پس از تقسیم داده ها به کمک جمع وزن دار آنتروپی ها بدست می آید:

$$\begin{aligned}E(Y|X)&=-\sum_{i=1}^{k}P(x_i)E(Y|x_i)\\&=-\sum_{i=1}^{k}\sum_{j=1}^{c}P(x_i)P(y_j|x_i)\log_2 P(y_j|x_i)\\&=-\sum_{i=1}^{k}\sum_{j=1}^{c}P(x_i,y_j)\log_2 P(y_j|x_i)\end{aligned} \quad (۲۰)$$

در این معادله نیز ما از قانون آشنای احتمالات که می گوید $P(x_i, y_j) = P(y_j|x_i) \times P(x_i)$ استفاده میکنیم.
برای رسیدن به پاسخ این سوال باید نشان دهیم که $E(Y|X) \leq E(Y)$ برای این کار اختلاف بین آنتروپی های قبل و بعد از تقسیم شدن را حساب میکنیم. برای این امر از معادلات ۱۸ و ۲۰ استفاده میکنیم.

$$
\begin{aligned}
& E(Y|X) - Y(E) \\
&= -\sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i, y_j) \log_2 P(y_j|x_i) + \sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i, y_j) \log_2 P(y_j) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i, y_j) \log_2 \frac{P(y_j}{P(y_j|x_i)} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i, y_j) \log_2 \frac{P(x_i)P(y_j)}{P(x_i, y_j)}
\end{aligned}
\tag{۲۱}
$$

برای اینکه اثبات کنیم معادله ۲۱ غیر مثبت است، از خاصیت زیر در توابع لگاریتم استفاده می نماییم:

$$
\sum_{k=1}^{d} a_k \log(z_k) \leq \log(\sum_{k=1}^{d} a_k z_k) \tag{۲۲}
$$

فرمول بالا به شرطی صادق است که $\sum_{k=1}^{d} a_k = 1$. با استفاده از این قاعده می توان نوشت:

$$
\begin{aligned}
E(Y|X) - Y(E) &\leq \log_2[\sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i, y_j) \log_2 \frac{P(x_i)P(y_j)}{P(x_i, y_j)}] \\
&= \log_2[\sum_{i=1}^{k} P(x_i) \sum_{j=1}^{c} P(y_j)] \\
&= \log_2(1) \\
&= 0
\end{aligned}
\tag{۲۳}
$$

از آنجا که $E(Y|X) - Y(E) \leq 0$ ، میتوان نتیجه گرفت که آنتروپی هرگز پس از تقسیم براساس یک خصیصه، افزایش نمی یابد.

سوال ۴

بخش a

برای کل داده ها، مقدار Classification Error از معادله زیر بدست می آید:

$$
E_{all} = 1 - max(\frac{50}{100}, \frac{50}{100}) = \frac{50}{100} \tag{۲۴}
$$

بعد از تقسیم بر اساس خصیصه A خواهیم داشت:

	A=T	A=F
+	25	25
-	0	50

بنابراین جدول، مقدار Gain در Error Rate به روش زیر بدست می آید:

$$E_{A=T} = 1 - max(\frac{25}{25}, \frac{0}{25}) = \frac{0}{25} = 0 \quad (۲۵)$$

$$E_{A=F} = 1 - max(\frac{25}{75}, \frac{50}{75}) = \frac{25}{75} \quad (۲۶)$$

$$Gain_A = E_{all} - \frac{25}{100}E_{A=T} - \frac{75}{100}E_{A=f} = \frac{25}{100} \quad (۲۷)$$

بعد از تقسیم بر اساس خصیصه B خواهیم داشت:

	B=T	B=F
+	30	20
-	20	30

بنابراین جدول، مقدار Gain در Error Rate به روش زیر بدست می آید:

$$E_{B=T} = 1 - max(\frac{30}{50}, \frac{20}{50}) = \frac{20}{50} \quad (۲۸)$$

$$E_{B=F} = 1 - max(\frac{20}{50}, \frac{30}{50}) = \frac{20}{50} \quad (۲۹)$$

$$Gain_B = E_{all} - \frac{50}{100}E_{B=T} - \frac{50}{100}E_{B=f} = \frac{10}{100} \quad (۳۰)$$

بعد از تقسیم بر اساس خصیصه C خواهیم داشت:

	C=T	C=F
+	25	25
-	25	25

بنابراین جدول، مقدار Gain در Error Rate به روش زیر بدست می آید:

$$E_{C=T} = 1 - max(\frac{25}{50}, \frac{25}{50}) = \frac{25}{50} \quad (۳۱)$$

$$E_{C=F} = 1 - max(\frac{25}{50}, \frac{25}{50}) = \frac{25}{50} \quad (۳۲)$$

$$Gain_C = E_{all} - \frac{50}{100}E_{C=T} - \frac{50}{100}E_{C=f} = \frac{0}{100} = 0 \quad (۳۳)$$

چون مقدار gain برای خصیصه A از همه بیشتر بود، این خصیصه برای تقسیم کردن داده ها باید مورد استفاده قرار گیرد.

بخش b

برای دسته ی $A = T$ دیگر نیازی به تقسیم کردن بیشتر نداریم. اما برای دسته ی $A = F$ می توان تقسیم بندی کرد و به جدول زیر دست یافت:

B	C	برچسب کلاس	
		+	−
T	T	0	20
F	T	0	5
T	F	25	0
F	F	0	25

بنابراین خطای دسته بندی گره فرزند $A = F$ برابر خواهد بود با:

$$E_{all} = \frac{25}{75} \quad (۳۴)$$

حالا به کمک جدولی که بدست آورده ایم میتوانیم gain ها را برای دسته های دیگر نیز بدست آوریم.
بعد از تقسیم بر اساس خصیصه B خواهیم داشت:

	B=T	B=F
+	25	0
−	20	30

بنابراین جدول، مقدار Gain در Error Rate به روش زیر بدست می آید:

$$E_{B=T} = 1 - max(\frac{25}{45}, \frac{20}{45}) = \frac{20}{45} \quad (۳۵)$$

$$E_{B=F} = 1 - max(\frac{0}{30}, \frac{30}{30}) = \frac{0}{30} = 0 \quad (۳۶)$$

$$Gain_B = E_{all} - \frac{45}{75}E_{B=T} - \frac{30}{75}E_{B=f} = \frac{5}{100} \quad (۳۷)$$

بعد از تقسیم بر اساس خصیصه C خواهیم داشت:

	C=T	C=F
+	0	25
-	25	25

بنابراین جدول، مقدار Gain در Error Rate به روش زیر بدست می آید:

$$E_{C=T} = 1 - max(\frac{0}{25}, \frac{25}{25}) = \frac{0}{25} = 0 \quad (۳۸)$$

$$E_{C=F} = 1 - max(\frac{25}{50}, \frac{25}{50}) = \frac{25}{50} \quad (۳۹)$$

$$Gain_C = E_{all} - \frac{25}{75}E_{C=T} - \frac{50}{75}E_{C=f} = \frac{0}{100} = 0 \quad (۴۰)$$

در این شرایط جدید، دسته بندی بعدی باید براساس خصیصه B باشد.

بخش c

۲۰ داده اشتباه دسته بندی خواهند شد.

بخش d

برای گره $C = T$ میزان Error rate قبل از تقسیم شدن برابر با $E_{original} = \frac{25}{50}$ است.
بعد از تقسیم بر اساس خصیصه A خواهیم داشت:

	A=T	A=F
+	25	0
-	0	25

بنابراین جدول، مقدار Gain در Error Rate به روش زیر بدست می آید:

$$E_{A=T} = 1 - max(\frac{25}{25}, \frac{0}{25}) = \frac{0}{25} = 0 \quad (۴۱)$$

$$E_{A=F} = 1 - max(\frac{0}{25}, \frac{25}{25}) = \frac{0}{25} = 0 \quad (۴۲)$$

$$Gain_A = E_{all} - \frac{25}{50}E_{A=T} - \frac{25}{50}E_{A=f} = \frac{25}{50} \quad (۴۳)$$

بعد از تقسیم بر اساس خصیصه B خواهیم داشت:

	B=T	B=F
+	5	20
-	20	5

بنابراین جدول، مقدار Gain در Error Rate به روش زیر بدست می آید:

$$E_{B=T} = 1 - max(\frac{5}{25}, \frac{20}{25}) = \frac{5}{25} \quad (۴۴)$$

$$E_{B=F} = 1 - max(\frac{20}{25}, \frac{5}{25}) = \frac{5}{25} \quad (۴۵)$$

$$Gain_B = E_{all} - \frac{25}{50}E_{B=T} - \frac{25}{50}E_{B=f} = \frac{15}{50} \quad (۴۶)$$

بنابراین، باید A به عنوان خصیصه ای که تقسیم را براساس آن انجام می دهیم انتخاب شود.
برای گره $C = F$ میزان Error rate قبل از تقسیم شدن برابر با $E_{original} = \frac{25}{50}$ است.
بعد از تقسیم بر اساس خصیصه A خواهیم داشت:

	A=T	A=F
+	0	25
-	0	25

بنابراین جدول، مقدار Gain در Error Rate به روش زیر بدست می آید:

$$E_{A=T} = 0 \quad (۴۷)$$

$$E_{A=F} = 1 - max(\frac{25}{50}, \frac{25}{50}) = \frac{25}{50} \quad (۴۸)$$

$$Gain_A = E_{all} - \frac{0}{50}E_{A=T} - \frac{50}{50}E_{A=f} = \frac{0}{50} = 0 \quad (۴۹)$$

بعد از تقسیم بر اساس خصیصه B خواهیم داشت:

	B=T	B=F
+	25	0
-	0	25

بنابراین جدول، مقدار Gain در Error Rate به روش زیر بدست می آید:

$$E_{B=T} = 1 - max(\frac{25}{25}, \frac{0}{25}) = \frac{0}{25} = 0 \quad (۵۰)$$

$$E_{B=F} = 1 - max(\frac{0}{25}, \frac{25}{25}) = \frac{0}{25} = 0 \quad (۵۱)$$

$$Gain_B = E_{all} - \frac{25}{50}E_{B=T} - \frac{25}{50}E_{B=f} = \frac{25}{50} \quad (۵۲)$$

بنابراین، باید B به عنوان خصیصه ای که تقسیم را براساس آن انجام می دهیم انتخاب شود.

بخش e

از این تمرین می توان نتیجه گرفت که روش حریصانه لزوما بهترین درخت را به ما نمی دهد.

سوال ۵

بخش a

اگر داده های Train را وارد درخت تصمیم داده شده کنیم، متوجه میشویم که این درخت ۵ مورد را به طور اشتباه دسته بندی می کند. بنابراین خطای این درخت به طور خوشبینانه برابر با $\frac{5}{10}$ خواهد بود.

بخش b

برای بدست آوردن خطای بدبیانه از معادله زیر استفاده می کنیم:

$$\frac{5 + (4 \times 0.5)}{10} = \frac{7}{10} \quad (۵۳)$$

بخش c

اگر داده های Test را وارد درخت تصمیم داده شده کنیم، متوجه میشویم که این درخت ۱ مورد را به طور اشتباه دسته بندی می کند. بنابراین خطای این درخت به طور خوشبینانه برابر با $\frac{1}{5}$ خواهد بود.

سوال ۶

بخش a

با فرض اینکه داده های Train و Test از هر دسته به میزان کافی داشته باشند، انتظار می رود که خطای test نزدیک ۵۰ درصد باشد. زیرا هر کدام از داده های مجموعه test با احتمالی مساوی سایر داده ها میتواند مثبت یا منفی باشد، همانطور که در داده های train نیز این موضوع صدق می کند.

بخش c

در یک درخت کامل Training Error باید ۱۰۰ درصد باشد. در حالیکه برای هر نمونه bootstrap نزدیک به ۵۰ درصد است. به کمک این اطلاعات و فرمول مربوط به متد bootstrap 632. دقت تخمین به شکل زیر بدست می آید:

$$\frac{1}{b}\sum_{i=1}^{b}[0.632 \times 0.5 + 0.368 \times 1] = 0.684 \quad (۵۴)$$

بخش d

روش اول تخمین بهتری از خطا به ما می دهد. روش دوم دیدگاهی امیدوارانه تر به دقت تخمین دارد.