به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس داده کاوی

تمرین سوم

نام و نام خانوادگی : حسین سیفی

شماره دانشجویی : ۸۱۰۱۰۰۳۸۶

اردی‌بهشت ۱۴۰۱

سوال ۱

الف

ب

نادرست است. اگر یکی از ویژگی های دیتاست D1 به صورت یک شناسه واحد عمل کند، در صورت استفاده از شروطی که پاسخ دودویی دارند، نمی‌توان درخت تصمیمی با ارتفاع ۱ ایجاد کرد. در صورت استفاده از ویژگی ذکر شده، ارتفاع درخت به اندازه تعداد تاپل‌های دیتاست خواهد بود و در صورت استفاده از دیگر ویژگی‌ها می‌توان این ارتفاع را کاهش داد. بنابراین گزاره فوق نادرست است و گزاره درست به شکل زیر می‌باشد:

«اگر یکی از ویژگی های دیتاست D1 به صورت یک شناسه واحد عمل کند، آنگاه ارتفاع درخت تصمیم ساخته شده از روی D1 که از Gain Information به عنوان معیار انتخاب ویژگی استفاده می کند، حداقل برابر با یک خواهد بود.»

پ

ت

نادرست است. تعداد برگ‌ها در بیشترین حالت برابر تعداد تاپل‌های دیتاست است بدین صورت که هر برگ کلاس یک تاپل را مشخص می‌کند و در نتیجه به تعداد تاپل‌ها در درخت تصمیم برگ داریم. در حالات دیگر حداقل یک برگ وجود دارد که بیش از یک تاپل را طبقه بندی می‌کند بنابراین تعداد برگ‌ها کمتر از تعداد تاپل‌ها می‌شود. در نتیجه گزاره فوق نادرست است و گزاره درست به شکل زیر است:

«تعداد برگ های درخت تصمیم ساخته شده از روی دیتاست از تعداد تاپل های آن دیتاست کمتر یا مساوی است.»

ث

درست است. در آزمایش بیماری‌های صعب العلاجی که زمان شروع درمان در آن‌ها مهم است ترجیح بر این است که تا حد ممکن بیماران با نتیجه مثبت تشخیص داده شوند و اگر به تعداد کمی از نتایج منفی، مثبت اعلام شوند اهمیت کمتری نسبت به حالت مخالف دارد چرا که با انجام آزمایشات بیشتر می‌توانند از آزمایشات تکمیلی استفاده کنند. بنابراین اهمیت تشخیص درست بیماران با نتیجه مثبت بسیار بیشتر از بیماران با نتیجه منفی است و در نتیجه Sensitivity از Specificity اهمیت بیشتری دارد.

ج

نادرست است. الگوریتم Bagging به دلیل استفاده از طبقه بندهای متفاوت و بررسی نتایج آن‌ها و انتخاب کلاسی که اکثریت مدل‌ها پیش بینی کرده‌اند به منظور جلوگیری از تاثیر گذاشتن مدل‌های Underfit شده و Overfit شده ایجاد شده است. بنابراین گزاره فوق نادرست است و گزاره درست به شکل زیر است:

«در الگوریتم bagging ، حتی اگر دسته بندهای پایه دچار underfitting باشند، آنگاه ترکیب آنها (یعنی دسته بندbagged) دچار underfitting نخواهد شد.»

چ

نادرست است. در روش K-Nearest-Neighbour در صورتی که مقدار K خیلی بزرگ باشد، مرز بین کلاس‌ها Smoothتر شده و در نتیجه احتمال خطا بالاتر می‌رود و با اطمینان کمتری می‌توان کلاس مورد نظر برای نمونه جدید را تخمین زد. بنابراین گزاره فوق نادرست است و گزاره درست به شکل زیر است:

« در روش K-Nearest Neighbor ،اگر مقدار K خیلی بزرگ باشد، با اطمینان کمتری می توان کالس یک شی داده شده را تخمین زد.»

ح

درست است. در صورتی که بخواهیم از از دسته بند Naïve Bayes برای یک دیتاست با m کلاس و v ویژگی استفاده کنیم، با توجه به عبارت P(c|d)=argmax[p(c)*p(d|c)] نیاز به محاسبه m برای کلاس‌ها و محاسبه احتمال تمامی ویژگی‌ها برای هر کلاس داریم که برابر m*v می‌باشد. بنابراین تعداد پارامترهایی که باید محاسبه شوند برابر m*v+m می‌باشد.

سوال ۲

الف

در اولین مرحله مقدار Foil-Gain را برای هر یک از ویژگی‌ها و تمام مقادیر ممکن آن‌ها و هر دو کلاس موجود محاسبه می‌کنیم. مقادیر P،N'،P' و N از جدول موجود در سوال ۲ استخراج شده‌اند.

Attribute	Value	P'	N'	P	N	Foil-gain	
						T	F
A1	w	1	0	7	9	1.192645078	#NUM!
A1	x	3	4	7	9	-0.08924203	0.090880306
A1	y	2	2	7	9	0.385290156	-0.33985000
A1	z	2	2	7	9	0.385290156	-0.33985000
A2	w	0	3	7	9	#NUM!	2.490224996
A2	x	2	3	7	9	-0.25856603	0.279328213
A2	y	3	2	7	9	1.367038451	-0.98370619
A2	z	2	1	7	9	1.215365154	-0.75488750
A3	w	2	1	7	9	1.215365154	-0.75488750
A3	x	5	4	7	9	1.723240857	-1.35940001
A3	y	0	2	7	9	#NUM!	1.660149997
A3	z	0	2	7	9	#NUM!	1.660149997
A1	w	1	0	7	9	1.192645078	#NUM!

با توجه به جدول فوق بیشترین مقدار Foil-Gain را مقدار w برای ویژگی A2 دارد که تنها تاپل‌های کلاس F را پوشش می‌دهد بنابراین «IF A2==w THEN class=F» می‌تواند به تنهایی یک قانون باشد. پس قانون ذکر شده را به مجموعه قوانین اضافه می‌کنیم و تاپل‌های پوشش داده شده توسط آن را از مجموعه داده حذف می‌نماییم(تاپل‌های ۳،۷ و ۸). سپس جدول فوق را به شکل زیر به روز رسانی می‌کنیم:

Attribute	Value	P'	N'	P	N	Foil-Gain	
						T	F
A1	w	1	0	7	6	0.893084796	#NUM!
A1	x	3	2	7	6	0.468357606	-0.41290175
A1	y	2	2	7	6	-0.21383040	0.230954435
A1	z	2	1	7	6	0.616244591	-0.46948528
A2	w	0	0	7	6	#DIV/0!	#DIV/0!
A2	x	2	3	7	6	-0.85768659	1.13553487
A2	y	3	2	7	6	0.468357606	-0.41290175
A2	z	2	1	7	6	0.616244591	-0.46948528
A3	w	2	0	7	6	1.786169592	#NUM!
A3	x	5	4	7	6	0.225439448	-0.21779113
A3	y	0	0	7	6	#DIV/0!	#DIV/0!
A3	z	0	2	7	6	#NUM!	2.230954435

با توجه به جدول فوق بیشترین مقدار Foil-Gain را مقدار z برای ویژگی A3 دارد که تنها تاپل‌های کلاس F را پوشش می‌دهد بنابراین « IF A3==z THEN class=F» می‌تواند به تنهایی یک قانون باشد. پس قانون ذکر شده را به مجموعه قوانین اضافه می‌کنیم و تاپل‌های پوشش داده شده توسط آن را از مجموعه داده حذف می‌نماییم(تاپل‌های ۵ و ۱۳). سپس جدول فوق را به شکل زیر به روز رسانی می‌کنیم:

Attribute	Value	P'	N'	P	N	Foil-Gain	
						T	F
A1	w	1	0	7	4	0.652076697	#NUM!
A1	x	3	1	7	4	0.711117592	-0.54056838
A1	y	2	1	7	4	0.134228392	-0.12553088
A1	z	2	1	7	4	0.134228392	-0.12553088
A2	w	0	0	7	4	#DIV/0!	#DIV/0!
A2	x	2	2	7	4	-0.69584660	0.918863237
A2	y	3	2	7	4	-0.25466669	0.275007047
A2	z	2	0	7	4	1.304153393	#NUM!
A3	w	2	0	7	4	1.304153393	#NUM!
A3	x	5	4	7	4	-0.97960105	1.158026469
A3	y	0	0	7	4	#DIV/0!	#DIV/0!
A3	z	0	0	7	4	#DIV/0!	#DIV/0!

با توجه به جدول فوق بیشترین مقدار Foil-Gain را مقدار z برای ویژگی A2 دارد که تنها تاپل‌های کلاس T را پوشش می‌دهد بنابراین « IF A2==z THEN class=T» می‌تواند به تنهایی یک قانون باشد. پس قانون ذکر شده را به مجموعه قوانین اضافه می‌کنیم و تاپل‌های پوشش داده شده توسط آن را از مجموعه داده حذف می‌نماییم(تاپل‌های ۲و ۴). سپس جدول فوق را به شکل زیر به روز رسانی می‌کنیم:

Attribute	Value	P'	N'	P	N	Foil-Gain	
						T	F
A1	w	1	0	5	4	0.847996907	#NUM!
A1	x	3	1	5	4	1.298878222	-0.83007499
A1	y	1	1	5	4	-0.15200309	0.169925001
A1	z	1	1	5	4	-0.15200309	0.169925001
A2	w	0	0	5	4	#DIV/0!	#DIV/0!
A2	x	2	2	5	4	-0.30400618	0.339850003
A2	y	3	2	5	4	0.333093937	-0.30400618
A2	z	0	0	5	4	#DIV/0!	#DIV/0!
A3	w	2	0	5	4	1.695993813	#NUM!
A3	x	3	4	5	4	-1.12318654	1.450280318
A3	y	0	0	5	4	#DIV/0!	#DIV/0!
A3	z	0	0	5	4	#DIV/0!	#DIV/0!

با توجه به جدول فوق بیشترین مقدار Foil-Gain را مقدار w برای ویژگی A3 دارد که تنها تاپل‌های کلاس T را پوشش می‌دهد بنابراین «IF A3==w THEN class=T» می‌تواند به تنهایی یک قانون باشد. پس قانون ذکر شده را به مجموعه قوانین اضافه می‌کنیم و تاپل‌های پوشش داده شده توسط آن را از مجموعه داده حذف می‌نماییم(تاپل‌های ۱۱ و ۱۴). سپس جدول فوق را به شکل زیر به روز رسانی می‌کنیم:

Attribute	Value	P'	N'	P	N	Foil-Gain	
						T	F
A1	w	1	0	3	4	1.222392421	#NUM!
A1	x	3	1	3	4	2.422064766	-1.19264507
A1	y	0	1	3	4	#NUM!	0.807354922
A1	z	0	1	3	4	#NUM!	0.807354922
A2	w	0	0	3	4	#DIV/0!	#DIV/0!

A2	x	0	2	3	4	#NUM!	1.614709844
A2	y	3	2	3	4	1.456280482	-1.02914634
A2	z	0	0	3	4	#DIV/0!	#DIV/0!
A3	w	0	0	3	4	#DIV/0!	#DIV/0!
A3	x	3	4	3	4	0	0
A3	y	0	0	3	4	#DIV/0!	#DIV/0!
A3	z	0	0	3	4	#DIV/0!	#DIV/0!

با توجه به جدول فوق بیشترین مقدار Foil-Gain را مقدار x برای ویژگی A1 و کلاس T دارد بنابراین باید این قانون را به کمک تمامی مقادیر ممکن برای ویژگی‌های A2 و A3 گسترش دهیم تا به پوششی مناسب از تاپل‌ها برسیم. جدول قانون گسترش یافته به شرح زیر می‌باشد:

Attribute	Value	P’	N’	P	N	Foil-Gain
A2	w	0	0	3	1	#DIV/0!
A2	x	0	0	3	1	#DIV/0!
A2	y	3	1	3	1	0
A2	z	0	0	3	1	#DIV/0!
A3	w	0	0	3	1	#DIV/0!
A3	x	3	1	3	1	0
A3	y	0	0	3	1	#DIV/0!
A3	z	0	0	3	1	#DIV/0!

همانطور که در جدول فوق دیده می‌شود تنها سطرهایی که مقدار دارند به ازای ویژگی A2 با مقدار y و ویژگی A3 با مقدار x هستند و مقادیری برابر با صفر دارند که در نتیجه نسبت به حالت قبل هیچ پیشرفتی در تفکیک تاپل‌ها وجود ندارد همچنین با گسترش این قانون نیز تفکیک بهتری نخواهیم داشت بنابراین قانون مرحله قبل (IF A1==x THEN class=True) را به مجموعه قوانین اضافه می‌کنیم و تاپل‌های مربوط به آن را از را از مجموعه داده حذف می‌نماییم(تاپل‌های ۱،۶،۹ و ۱۲). سپس جدول فوق را به شکل زیر به روز رسانی می‌کنیم:

Attribute	Value	P’	N’	P	N	Foil-Gain	
						T	F
A1	w	1	0	0	3	#NUM!	#NUM!
A1	x	0	0	0	3	#DIV/0!	#DIV/0!
A1	y	0	1	0	3	#NUM!	0
A1	z	0	1	0	3	#NUM!	0
A2	w	0	0	0	3	#DIV/0!	#DIV/0!
A2	x	0	2	0	3	#NUM!	0
A2	y	0	1	0	3	#NUM!	0
A2	z	0	0	0	3	#DIV/0!	#DIV/0!
A3	w	0	0	0	3	#DIV/0!	#DIV/0!
A3	x	0	3	0	3	#NUM!	0
A3	y	0	0	0	3	#DIV/0!	#DIV/0!
A3	z	0	0	0	3	#DIV/0!	#DIV/0!

با توجه به اینکه تعداد تاپل‌های کلاس F به پایان رسیده‌اند، مقدار Foil-Gain تمامی ویژگی‌ها و مقادیر باقیمانده برابر صفر خواهد شد بنابراین از بین سطرهای جدول فوق، ویژگی و مقداری را انتخاب می‌کنیم که بیشترین پوشش را روی تاپل‌های باقیمانده داشته باشد که مقدار x برای ویژگی A3 خصوصیت فوق را دارد و به مجموعه قوانین اضافه می‌شود. پس از حذف تاپل‌های پوشش داده شده توسط آن(تاپل‌های ۱۰،۱۵ و ۱۶) دیتاست دارای سطر دیگری نخواهد بود و الگوریتم به پایان می‌رسد. جدول قوانین به وجود آمده به ترتیب اولویت به شرح زیر خواهند بود:

Priority	Rule
1	IF A2==w THEN class=F
2	IF A3==z THEN class=F
3	IF A2==z THEN class=T

4	IF A3==w THEN class=T
5	IF A1==x THEN class=T
6	IF A3==x THEN class=F

ب

پ

ت

	Rule Induction(الف)	Gain Ratio(ب)
X1	F	
X2	T	

روش بخش الف تاپل X1 را به کمک قانون ۱ با F و تاپل X2 را به کمک قانون ۴ با T برچسب گذاری می‌کند.

سوال ۳

الف

فرمول اولیه Accuracy به شکل زیر موجود است:

$$Accuracy = \frac{tp + tn}{N} = \frac{tp}{N} + \frac{tn}{N}$$

پس از تفکیک کسر اولیه، صورت و مخرج بخش اول کسر را در tp+fn ضرب می‌کنیم و صورت و مخرج کسر دوم را در tn+fp ضرب می‌کنیم:

$$Accurcay = \frac{tp}{N} * \frac{tp + fn}{tp + fn} + \frac{tn}{N} * \frac{tn + fp}{tn + fp}$$

می‌توان کسر فوق را به شکل زیر بازنویسی کرد:

$$Accurcay = \frac{tp}{tp + fn} * \frac{tp + fn}{N} + \frac{tn}{tn + fp} * \frac{tn + fp}{N}$$

در عبارت فوق، مقادیر Sensitivity=$\frac{tp}{tp+fn}$ و Specificity=$\frac{tn}{tn+fp}$ مشهود است. همچنین عبارت (tp+fn)/N برابر با نسبت تعداد مثبت‌های واقعی به کل داده‌ها و عبارت (tn+fp)/N برابر با نسبت تعداد منفی‌های واقعی به کل داده‌ها هستند و به ترتیب می‌توان آن‌ها را Positive-ratio و Negative-ratio نام گذاری کرد. درنهایت مقدار Accuracy به شکل زیر درمی‌آید:

$$Accuracy = Sensitivity * Positive_{Ratio} + Specificity * Negative_{Ratio}$$

ب

فرمول میانگین هارمونیک به شکل زیر می‌باشد:

$$Harmonic\_mean = (\frac{x_1^{-1} + x_2^{-1} + \cdots + x_n^{-1}}{n})^{-1}$$

با جایگذاری ۱ مقدار Precision و β^2 مقدار Recall در فرمول فوق می‌توان نشان داد که عبارت به دست آمده برابر با F-measure می‌باشد:

$$Harmonic_{mean} = \left(\frac{\beta^2 \left(\frac{1}{R}\right) + \left(\frac{1}{P}\right)}{\beta^2 + 1} \right)^{-1} = \frac{\beta_2 + 1}{\frac{\beta^2 P + R}{PR}} = \frac{(\beta_2 + 1)PR}{\beta^2 P + R} = F\_measure$$

پ

برای محاسبه معیارهای مبتنی بر Accuracy برای یک مسئله که تاپل های یک دیتاست بتوانند متعلق به بیش از یک کلاس باشند(Multi-label-classification) می‌توان از میانگین Accuracyها به عنوان معیاری مناسب استفاده کرد. این روش بدین صورت است که Accuracy را برای هر یک از کلاس‌ها به صورت مستقل محاسبه می‌کنیم و دیگر کلاس‌ها در نظر گرفته نمی‌شوند سپس میانگین Accuracyهای به دست آمده را محاسبه می‌کنیم. در صورتی که تعداد مثبت و منفی کلاس‌ها با یکدیگر تفاوت داشته باشد نیز می‌توان از میانگین وزن دارد برای محاسبه Accuracy کل مسئله استفاده کرد.

ت

برای مسئله Class Imbalance می‌توان از معیارهای تجمیع با روش Micro-Averaging استفاده کرد و این مقادیر معتبر هستند چرا که این روش‌ها مقادیر Precision، Recall و ... را به صورت وزن‌دار محاسبه می‌کند و در نتیجه هر کلاس به اندازه خود در معیارها تاثیرگذار خواهد بود در حالی که در صورت محاسبه به روش Macro-Averaging وزن کلاس‌های بزرگ و کوچک یکسان در نظر گرفته می‌شود.