به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس داده‌کاوی

تمرین تشریحی ۳

اردیبهشت ماه ۱۴۰۱

* فهرست

سؤال ۱

درست یا نادرست بودن هر کدام از عبارات زیر را **با ذکر دلیل** مشخص نمایید.

* توجه: اگر یک عبارت درست است، علت درست بودن آنرا توضیح دهید، ولی اگر نادرست است، ابتدا علت نادرست بودن آنرا توضیح دهید و سپس آنرا اصلاح نمایید.

الف) در یک درخت تصمیم که از Gain Ratio به عنوان معیار انتخاب ویژگی استفاده می‌کند، ممکن است که یک ویژگی گسسته وجود داشته باشد که بیش از ۱ بار تست شود.

ب) اگر یکی از ویژگی‌های دیتاست D_1 به‌صورت یک شناسه واحد عمل کند، آنگاه **ارتفاع** درخت تصمیم ساخته شده از روی D_1 که از Information Gain به عنوان معیار انتخاب ویژگی استفاده می‌کند، همواره برابر با یک خواهد بود.

پ) فرض کنید که دیتاست D_2 دارای $k \geq 1$ ویژگی گسسته باشد، به طوریکه ویژگی iام، $1 \leq i \leq k$، $3i$ مقدار متمایز داشته باشد. حال اگر از Gini Index به عنوان معیار انتخاب ویژگی برای ساخت درخت تصمیم مربوط به D_2 استفاده کنیم، آنگاه به $\frac{1}{7}(8^{k+1} - 1) - 2k$ حالت ممکن می‌توان ریشه درخت را انتخاب کرد.

ت) ممکن است تعداد تاپل‌های یک دیتاست از تعداد برگ‌های درخت تصمیم ساخته شده از روی آن دیتاست کمتر باشد. False

ث) برای ارزیابی یک سیستم هوشمند که وظیفه‌ی تشخیص زودهنگام ابتلا/عدم ابتلای افراد به بیماری سرطان را برعهده دارد، sensitivity مهم‌تر از specificity است. True

ج) در الگوریتم bagging، اگر دسته‌بندهای پایه دچار underfitting باشند، آنگاه ترکیب آنها (یعنی دسته‌بند bagged) نیز دچار underfitting خواهد بود. False

چ) در روش K-Nearest Neighbor، اگر مقدار K خیلی بزرگ باشد، با اطمینان بیشتری می‌توان کلاس یک شی داده شده را تخمین زد. False

ح) برای استفاده از دسته‌بند Naïve Bayes در یک مسئله‌ی دسته‌بندی با m کلاس و v ویژگی دودویی، $m + m \times v$ پارامتر را بایستی تخمین زد. True

سؤال ۲

با توجه به دیتاست زیر، به سوالات داده شده پاسخ دهید.

* توجه: لطفا تمام محاسبات لازم را به طور کامل و مرحله به مرحله در گزارش خود بنویسید. تمام محاسبات بایستی با دقت ۳ رقم اعشار انجام شده باشند.

ردیف	ویژگی‌ها			کلاس
	A1	A2	A3	
۱	x	y	x	T
۲	z	z	x	T
۳	x	w	y	F
۴	y	z	x	T
۵	y	x	z	F
۶	x	y	x	T
۷	x	w	y	F
۸	z	w	w	F
۹	x	y	x	F
۱۰	y	x	x	F
۱۱	y	x	w	T
۱۲	x	y	x	T
۱۳	x	z	z	F
۱۴	z	x	w	T
۱۵	z	y	x	F
۱۶	w	x	x	F

الف) با استفاده از الگوریتم Sequential Covering، مجموعه قواعد ممکن را از روی دیتاست داده شده تولید نمایید. برای ارزیابی کیفیت هر قاعده از معیار FOIL_Gain استفاده نمایید.

ب) درخت تصمیم دیتاست داده شده را با استفاده از معیار انتخاب ویژگی Gain ratio رسم نمایید.

پ) مجموعه قواعد ممکن را از روی درخت تصمیم رسم شده در بخش ب، استخراج نمایید.

ت) مجموعه قواعد تولید شده در بخش الف و پ را از نظر نقاط قوت و منفیشان با همدیگر مقایسه نمایید.

ث) با استفاده از قواعد تولید شده در بخش الف و پ کلاس تاپل‌های

$X1 = (A3 = x, A2 = w, A1 = y)$

$X2 = (A3 = w, A2 = y, A1 = z)$

را پیش‌بینی نمایید. آیا هر دو روش کلاس یکسانی را پیش‌بینی می‌کنند؟

سؤال ۳

در رابطه با معیارهای ارزیابی مدل‌های دسته‌بندی به سوالات زیر پاسخ دهید:

الف) نشان دهید که معیار accuracy، تابعی از specificity و sensitivity است.

ب) نشان دهید که معیار f-measure، میانگین هارمونیک precision و recall است.

پ) اگر تاپل‌های یک دیتاست بتوانند متعلق به بیش از یک کلاس باشند، آنگاه چطور می‌توان از معیارهای مبتنی بر accuracy برای ارزیابی عملکرد یک مدل بر روی این دیتاست استفاده کرد؟

ت) از کدام یک از معیارهای ارزیابی می‌توان برای مسئله‌ی class imbalance استفاده کرد؟

ملاحظات (حتما مطالعه شود)

گزارش شما باید در قالب یک فایل با فرمت PDF (گزارش تایپ شده) و با عنوان DM_HW3_StudentID تحویل داده شود.

- خوانایی گزارش شما از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.
- مهلت تحویل تمرین به هیچ عنوان تمدید نخواهد شد. تمرین تا یک هفته بعد از مهلت تعیین شده با جریمه تحویل گرفته می‌شود که جریمه تاخیر تحویل تمرین تا یک هفته ۳۰ درصد است.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

(پیمان رستمی) pe.rostami@ut.ac.ir

مهلت تحویل بدون جریمه: ۳۰ اردیبهشت ۱۴۰۱

مهلت تحویل با تاخیر، با جریمه ۳۰ درصد: ۶ خرداد ۱۴۰۱