دانشگاه صنعتی اصفهان
دانشکده مهندسی برق و کامپیوتر

مبانی داده کاوی

پاسخنامه تمرین سری ۳

بهار ۱۴۰۳

فهرست مطالب

۱ پاسخ سوالات

۱.۱ پاسخ سوال ۱

آ: بی نظارت

ب: با نظارت

ج: در صورتی که بخواهیم به گروه خاصی از مشتریان وام اعطا کنیم تمرکز بر خوشه بندی است و روش بدون نظارت است اما در صورتی که به یک مشتری خاص میخواهیم وام اعطا کنیم می توانیم از روش بانظارت استفاده کنیم.

د: با نظارت

ه: با نظارت

و: بی نظارت

ز: با نظارت

ح: بی نظارت

ط: با نظارت

ی: با نظارت

۲.۱ پاسخ سوال ۲

مطابق نکات مطرح شده در درس، Gain Ratio و Gini Index هردو در درخت‌های تصمیم برای کشف بهترین راه برای تقسیم داده‌ها استفاده می‌شوند. هدف هر دوی آنها ایجاد گروه‌های خالصتر است، اما به روش‌های متفاوت. در معیار Gain Ratio هم اطلاعات به دست آمده از تقسیم بر روی یک ویژگی و هم اطلاعات ذاتی خود ویژگی را (با هدف رسیدگی به سوگیری نسبت به ویژگی هایی با انجام تعداد زیادی تقسیم) در نظر گرفته می شود و Gini Index ناخالصی یا تصادفی بودن نحوه تقسیم یک ویژگی در داده‌ها را به کلاس‌ها اندازه‌گیری می‌کند و بر ایجاد پارتیشن‌های خالص‌تر با به حداقل رساندن ناخالصی Gini تمرکز می‌کند. به بیان دیگر Gain Ratio ، روند به دست آوردن اطلاعات را برای جلوگیری از برازش و سوگیری را تسهیل می کند، اما Gini Index بر ایجاد گره های همگن با مقادیر ناخالصی کمتر برای بهبود کیفیت تقسیم در درختان تصمیم تمرکز دارد.

مثال‌ها:

https://tungmphung.com/information-gain-gain-ratio-and-gini-index/

۳.۱ پاسخ سوال ۳

پاسخ صحیح یکی از دانشجویان در ادامه آمده است:

$$GI_{Car} = 1 - \frac{3}{10}^2 - \frac{7}{10}^2 = 0.42$$

$$GI_{Vacation} = 1 - \frac{5}{10}^2 - \frac{5}{10}^2 = 0.5$$

$$GI_{Season} = 1 - \frac{4}{10}^2 - \frac{3}{10}^2 - \frac{3}{10}^2 = 0.66$$

ابتدا برای یافتن بهترین ویژگی برای برای شروع درخت GI را در حالتی که هرکدام از ویژگی ها معیار شروع باشند محاسبه می کنیم:

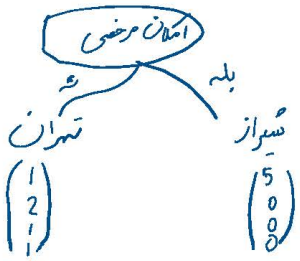
شهر	فصل			امکان مرخصی		خودرو شخصی	
	بهار	تابستان	پاییز	بله	خیر	بله	خیر
شیراز	3	1	2	5	1	3	3
تهران	0	2	0	0	2	2	0
اصفهان	0	0	1	0	1	1	0
رشت	1	0	0	0	1	1	0

$$GI_{Car} = \frac{3}{10}\left(1 - \frac{3^2}{3} - \frac{0^2}{3} - \frac{0^2}{3} - \frac{0^2}{3}\right) + \frac{7}{10}\left(1 - \frac{3^2}{7} - \frac{2^2}{7} - \frac{1^2}{7} - \frac{1^2}{7}\right) = 0 + 0.4857 = 0.4857$$

$$GI_{Vacation} = \frac{5}{10}\left(1 - \frac{1^2}{5} - \frac{2^2}{5} - \frac{1^2}{5} - \frac{1^2}{5}\right) + \frac{5}{10}\left(1 - \frac{5^2}{5} - \frac{0^2}{5} - \frac{0^2}{5} - \frac{0^2}{5}\right) = 0.36 + 0 = 0.36$$

$$GI_{Season} = \frac{3}{10}\left(1 - \frac{2^2}{3} - \frac{0^2}{3} - \frac{1^2}{3} - \frac{0^2}{3}\right) + \frac{3}{10}\left(1 - \frac{1^2}{3} - \frac{2^2}{3} - \frac{0^2}{3} - \frac{0^2}{3}\right) + \frac{4}{10}\left(1 - \frac{3^2}{4} - \frac{0^2}{4} - \frac{0^2}{4} - \frac{1^2}{4}\right) = 0.4167$$

پس با توجه کمتر بودن مقدار *Gini Index* ویژگی امکان مرخصی، از این ویژگی برای شروع درخت استفاده می کنیم.



در ادامه ویژگی بعدی را انتخاب می کنیم:

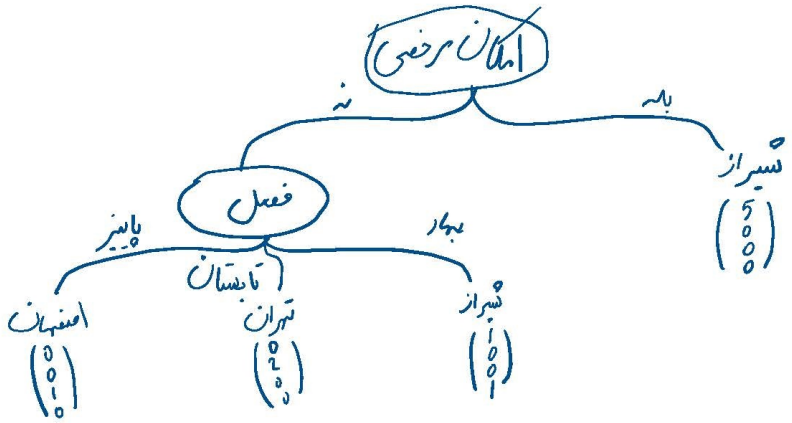
فصل	امکان مرخصی	خودروی شخصی	شهر
تابستان	بله	بله	شیراز
تابستان	نه	بله	تهران
بهار	بله	بله	شیراز
پاییز	بله	نه	شیراز
پاییز	نه	بله	اصفهان
پاییز	بله	نه	شیراز
بهار	نه	نه	شیراز
بهار	نه	بله	رشت
بهار	بله	بله	شیراز
تابستان	نه	بله	تهران

شهر	فصل			خودرو شخصی	
	بهار	تابستان	پاییز	بله	خیر
شیراز	1	0	0	0	1
تهران	0	2	0	2	0
اصفهان	0	0	1	1	0
رشت	1	0	0	1	0

$$I_{Car} = \frac{1}{5}\left(1 - \frac{1^2}{1} - \frac{0^2}{1} - \frac{0^2}{1} - \frac{0^2}{1}\right) + \frac{4}{5}\left(1 - \frac{0^2}{4} - \frac{2^2}{4} - \frac{1^2}{4} - \frac{1^2}{4}\right) = 0 + 0.5 = 0.5$$

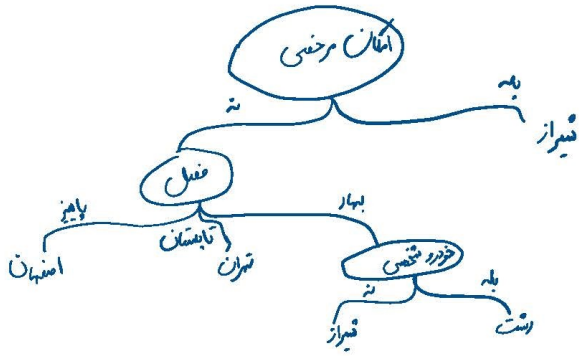
$$GI_{Season} = \frac{2}{5}\left(1 - \frac{1^2}{2} - \frac{0^2}{2} - \frac{0^2}{2} - \frac{1^2}{2}\right) + \frac{2}{5}\left(1 - \frac{0^2}{2} - \frac{2^2}{2} - \frac{0^2}{2} - \frac{0^2}{2}\right) + \frac{1}{5}\left(1 - \frac{0^2}{1} - \frac{0^2}{1} - \frac{1^2}{1} - \frac{0^2}{1}\right) = 0.2 + 0 + 0 = 0.2$$

با توجه به کمتر بودن *gini index* برای تقسیم بر اساس ویژگی فصل از این ویژگی استفاده می کنیم.



حال تنها ویژگی خودروی شخصی می ماند که با تقسیم بر اساس این ویژگی در این مرحله دسته ها کاملا خالص بدست می شوند و مقدار $Gini$ برابر با صفر می شود.

درخت نهایی به شکل زیر است:



۴۰۱ پاسخ سوال ۴

درخت تصمیم ساخته شده ممکن است بیش از حد با داده‌های آموزشی مطابقت داشته باشد. یعنی ممکن است شاخه‌های زیادی وجود داشته باشد که برخی از آنها شاید ناهنجاری هایی را در داده‌های آموزشی به دلیل نویز یا نقاط پرت، نشان دهند. هرس درختان با حذف کمترین انشعابات (از نوع قابل اعتماد یا به کارگیری معیارهای آماری) به حل مشکل overfitting می‌پردازد. این روند در نهایت منجر به ایجاد یک درخت تصمیم فشردهتر و قابل اعتمادتر می‌شود که در طبقه بندی داده ها، سریعتر و دقیقتر است.

اشکال استفاده از مجموعه جداگانهای از تاپل‌ها برای ارزیابی هرس، این است که ممکن است آن مجموعه، نماینده تاپل‌های آموزشی مورد استفاده برای ایجاد درخت تصمیم اصلی نباشد. اگر مجموعهٔ مجزای تاپل‌ها دارای انحراف باشد، استفاده از آنها برای ارزیابی درخت هرس شده، معیار خوبی برای محاسبهٔ دقت طبقه بندی درخت هرس شده، نخواهد بود. علاوه بر این، استفاده از مجموعه جداگانهای از تاپل‌ها برای ارزیابی هرس به این معنی است که تاپل‌های کمتری برای ایجاد و آزمایش درخت وجود دارد. لذا این مسئله به طور کلی یک اشکال در یادگیری ماشین به حساب می‌آید.