به نام خدا

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده مهندسی کامپیوتر و فناوری اطلاعات

پاسخ تمرین سری دوم داده کاوی

نیم‌سال اول ۰۰-۹۹

سوال ۱-

برچسب کلاس	A = T	A = F
+	4	0
-	3	3

برچسب کلاس	B = T	B = F
+	3	1
-	1	5

الف)

$$I(t) = \text{Entropy}(t) = -\sum_{j} p(j \mid t)\log_2(p(j \mid t))$$

$$I(\text{parent}) = -\frac{4}{10}\log_2\left(\frac{4}{10}\right) - \frac{6}{10}\log_2\left(\frac{6}{10}\right) = 0.529 + 0.442 = 0.971$$

$$I(A = T) = -\frac{4}{7}\log\left(\frac{4}{7}\right) - \frac{3}{7}\log\left(\frac{3}{7}\right) = 0.461 + 0.524 = 0.985$$

$$I(A = F) = 0$$

$$I(B = T) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.311 + 0.5 = 0.811$$

$$I(B = F) = -\frac{1}{6}\log_2\left(\frac{1}{6}\right) - \frac{5}{6}\log_2\left(\frac{5}{6}\right) = 0.431 + 0.219 = 0.650$$

$$\Delta = I(\text{parent}) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

$$\Delta_A = 0.971 - \frac{7}{10}I(A = T) - \frac{3}{10}I(A = F) = 0.971 - 0.7 * 0.985 - 0.3 * 0 = 0.2815$$

$$\Delta_B = 0.971 - \frac{4}{10}I(B = T) - \frac{6}{10}I(B = F) = 0.971 - 0.4 * 0.811 - 0.6 * 0.650 = 0.2566$$

باید معیار A انتخاب شود چرا که بهره اطلاعاتی بیشتری به ما می دهد

ب)

$$\mathrm{GINI(t)} = 1 - \sum_j \mathrm{p(j \mid t)}^2$$

overall gini before splitting:

$$GINIorig = 1 - 0.4^2 - 0.6^2 = 0.48$$

$$\mathrm{GINI(A = T)} = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 \approx 1 - 0.326 - 0.184 = 0.49$$

$$GINI(A{=}F) = 0$$

$gain\ in\ gini\ after\ splitting\ on\ A$:

$$\Delta = GINIorig - 7/10\ GINI(A = T) - 3/10\ GINI(A = F)$$
$$= 0.48 - 0.7 * 0.49 - 0 = 0.137$$

$$GINI(B = T) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 1 - 0.5625 - 0.0625 = 0.375$$

$$\mathrm{GINI(B = F)} = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 \approx 1 - 0.028 - 0.694 = 0.278$$

gain in gini after splitting on B:

$$\Delta = GINIorig - \frac{4}{10}\ \mathrm{GINI(B = T)} - \frac{6}{10}\ \mathrm{GINI(B = F)}$$
$$= 0.48 - 0.4 * 0.375 - 0.6 * 0.278 \approx 0.1632$$

پس باید معیار B انتخاب شود.

ج) بله ، حتی اگر این اندازه گیری ها دامنه مشابه و رفتار یکنواختی داشته باشند ، اما gain های مربوطه ، Δ ، که تفاوت اندازه گیری ها هستند ، لزوماً رفتار یکسان ندارند ، همانطور که در نتایج قسمت (الف) و (ب) نشان داده شد.

سوال ۲-

الف)

$$P(A = 1|-) = 2/5 = 0.4, P(B = 1|-) = 2/5 = 0.4,$$
$$P(C = 1|-) = 1, P(A = 0|-) = 3/5 = 0.6,$$
$$P(B = 0|-) = 3/5 = 0.6, P(C = 0|-) = 0;\ P(A = 1|+) = 3/5 = 0.6,$$
$$P(B = 1|+) = 1/5 = 0.2, P(C = 1|+) = 2/5 = 0.4,$$
$$P(A = 0|+) = 2/5 = 0.4, P(B = 0|+) = 4/5 = 0.8,$$
$$P(C = 0|+) = 3/5 = 0.6$$

ب) ابتدا فرض می کنیم که $P(A = 0, B = 1, C = 0) = K$ در اینصورت خواهیم داشت:

$$P(+|A = 0, B = 1, C = 0) = \frac{P(A = 0, B = 1, C = 0|+) \times P(+)}{P(A = 0, B = 1, C = 0)}$$
$$= \frac{P(A = 0|+)P(B = 1|+)P(C = 0|+) \times P(+)}{K}$$
$$= 0.4 \times 0.2 \times 0.6 \times 0.5/K = 0.024/K$$

$$P(-|A = 0, B = 1, C = 0) = \frac{P(A = 0, B = 1, C = 0|-) \times P(-)}{P(A = 0, B = 1, C = 0)}$$
$$= \frac{P(A = 0|-) \times P(B = 1|-) \times P(C = 0|-) \times P(-)}{K} = 0/K$$

پس برچسب کلاس + خواهد بود.

سوال ۳-

Accuracy یا دقت زمانی که داده های موجود در کلاس های مختلف در تعادل نیستند معیار و ملاک مناسبی برای ارزیابی مدل کلاس بندی ما نخواهد بود. به عنوان مثال موردی را فرض کنید که ما میخواهیم مدل کلاس بندی باینری خود را با داده هایی آموزش دهیم که ۹۵ درصد آنها برچسب مثبت و ۵ درصد آنها برچسب منفی دارند. در این حالت مدل آموزش داده شده اکثر داده های تست را مثبت برچسب گذاری خواهد کرد. اگر داده های تست ما اکثرشان مثبت باشند، مدل ما بر دقت یا Accuracy دقت بسیار بالایی را گزارش میدهد و اگر داده ها اکثرا منفی باشند، مدل دقت بسیار پایینی را گزارش خواهد کرد.

مثال دیگری را فرض کنید که ما از یک مدل خیلی ساده بدون نیاز به محاسبه استفاده کنیم به عنوان مثال مدل زیر را در نظر بگیرید:

- def is_positive (x):

    - return false

در صورتی که داده های تست را به تکه کد بالا بدهیم، و ۹۵ درصد داده هایمان به صورت تصادفی منفی باشد، امتیاز نهایی مدل ما بر حسب اندازه گیری دقت مقدار ۹۵ درصد گزارش میشود. مسلما ما این را از ارزیابی مدل خود نمیخواهیم.

۳ معیار جایگزین برای این روش ارزیابی وجود دارد که میتواند این مشکل را حل کند:

تا به حال با ترم های زیر آشنا شده اید: (از برچسب مثبت و منفی برای راحتی انتقال مفهوم استفاده میکنیم)

TP (True positive): داده هایی که به عنوان مثبت برچسب زده شده اند واقعا مثبت هستند.

TN (True Negative): داده هایی که به عنوان منفی برچسب زده شده اند و واقعا منفی هستند.

FP (False Positive): داده هایی که به عنوان مثبت برچسب زده شده اند اما در حقیقت منفی هستند.

FN (False Negative): داده هایی که به عنوان منفی برچسب زده شده اند اما در حقیقت مثبت هستند.

حال معیار های زیر را در نظر بگیرید:

- درستی یا صحت (Precision):

$$Precision = \frac{TP}{TP + FP}$$

وقتی مشکل بالا (مدل ساده ی شرطی ifرا) توسط Precision ارزیابی میکنیم، مدل هیچکدام را به عنوان مثبت شناسایی نکرده است (صورت ۰ مشیود) و دقت ۰ به دست می‌آید که منطقی است. (برای مدل بسیار ساده شرطی if منطقی است که مدل خوبی نیست و باید دقت نزدیک ۰ داشته باشد)

- فراخوانی (Recall):

$$Recall = \frac{TP}{TP + FN}$$

در این مورد برای مثال زده شده باز هم صورت ۰ و مخرج ۱ میشود (FN = 5%*allData) که نتیجه ۰ میشود و این ارزیابی منطقی است.

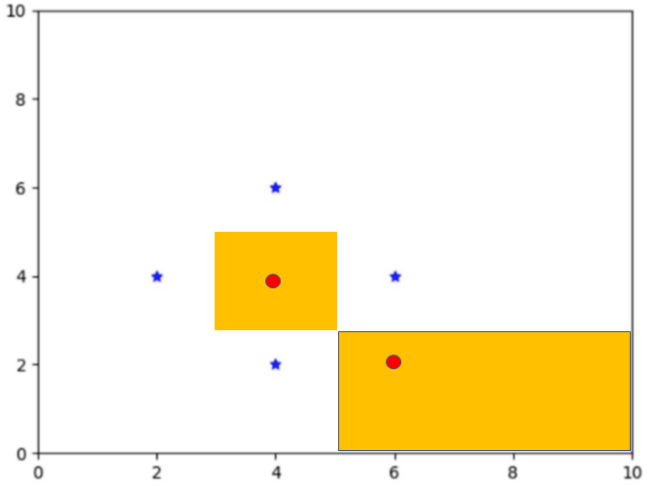
-F1-Score : ترکیبی است از Precision و Recall و به صورت زیر محاسبه میشود:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

این ارزیابی هم برای مثال زده شده ۰ میشود که منطقی است.

سوال ۴-

الف) با توجه به فرض K = ۱ در صورت سوال، دسته ی هر نقطه مشابه با نزدیک ترین داده به آن خواهد شد. مرزهای تصمیم گیری می توانند با ترسیم عمود منصف بر روی خط واصل نقاط موجود و نزدیک ترین همسایه که می تواند بیش از یک عدد باشد بدست آید.



ب) برچسب آبی (ستاره).

ج) بله. برای حل مسئله رگرسیون می توانیم از میانگین k نزدیک ترین همسایه برای مقدار خروجی استفاده کنیم. راهکارهای مشابه دیگری نیز برای جایگزین نمودن مقدار گسسته برچسب کلاس ها با مقدار خروجی مسئله رگرسیون به شرط تبعیت از تعریف KNN مورد قبول است.

د) می توان برای بدست آوردن مقدار بهینه k مدل دسته بندی خود را به مقادیر مختلف آن مورد ارزیابی قرار دهیم تا زمانی که به بهترین حالت دقت دست یابیم. برای این کار می توان یک subset از داده های آموزشی اصلی و یا توزیعی مشابه با آن ها را در نظر گرفت تا پیچیدگی زمانی کاهش یابد.

ه) خیر. در مجموعه داده های بزرگ و با افزایش ابعاد داده، هزینه محاسبه فاصله میان نقاط جدید و تمامی نقاط موجود بالا خواهد رفت و چون روش KNN یک تکنیک یادگیری تنبل است، حافظه بسیار زیادی برای نگهداری تمام داده ها مصرف می شود.

و) حالت آموزش با فرض ذخیره بودن داده ها: $O(1)$

حالت آزمایش با فرض تعداد داده آموزشی n و ابعاد آنها d: ابتدا برای بدست آوردن فاصله تا نقاط:

$O(nd)$ و سپس $O(nk)$ برای یافتن نزدیکترین داده ها. پس در کل: $O(nd + nk)$

ی) فاصله اقلیدسی همان کوتاه ترین مسیر مستقیم (بردار متصل کننده) میان دو نقطه یا نرم ۲ است و فاصله منهتن مجموع مقادیر حقیقی یا نرم ۱ بین دو نقطه. فاصله منهتن عموما در مسائل گسسته که برای جابجایی نیاز به حرکت گسسته در مولفه ها داریم به کار می آید.

سوال ۵- روش cross validation برای ارزیابی مدل آموزشی به کار می رود که هر بار تعدادی از داده ها که در فاز آموزشی استفاده نمیشوند را به عنوان داده ی آزمایشی انتخاب می کند. بدین ترتیب ارزیابی دقیق تر خواهد شد. مزیت این روش تشخیص امکان overfit شدن یا selection bias مدل می باشد.

یکی از دسته بندی های اصلی روش های cross validation به سه دسته ی K-fold، leave one out و hold-out می باشد. (دسته بندی کلی تر می تواند روش های exhaustive و non-exhaustive باشد که این تکنیک ها زیر مجموعه آن ها خواهند بود)

روش K-fold: مجموعه آموزش به k دسته هم اندازه تقسیم می شود و یکی ازین دسته ها برای ارزیابی استفاده می گردد. این کار k بار انجام می شود تا نهایتا همه k دسته یک بار به عنوان داده ارزیابی استفاده گردند.

leave one out: این روش یک حالت خاص از روش فوق به حساب می آید که طی آن مقدار k برابر با تعداد داده ها خواهد بود.

hold out: ارزیابی به صورت کلی انجام میگیرد. در واقع طی این روش داده ها به دو دسته ی آموزشی و آزمایش تقسیم خواهند شد و پس از آموزش مدل بر روی دسته ی آموزشی، آن را بر روی داده های آزمایشی مورد ارزیابی قرار خواهیم داد.