به نام خدا

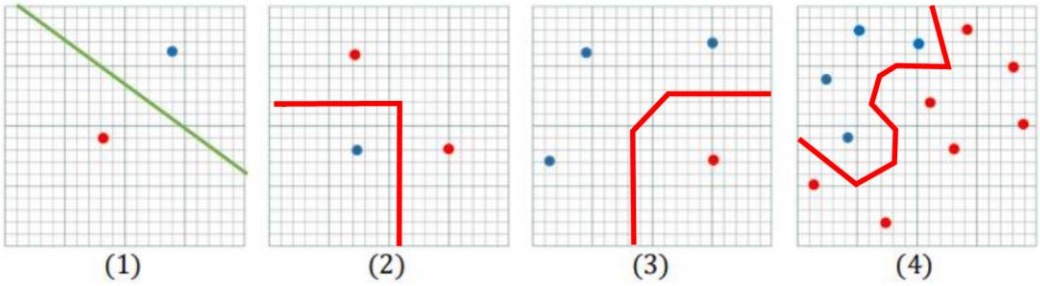
# تمرین دوم درس مبانی داده کاوی

**امیررضا حسینی ۹۸۲۰۳۶۳**

## سوالات تئوری

سوال ۱)

الف) بین هرکدام از کلاس‌ها، عمود منصف دو داده رسم میشود.




ب) در صورتی که بتوان داده‌ها را خوشه بندی کرد و به جای ذخیره کردن تمامی دیتاها، فقط یک دیتا از هر خوشه را به عنوان نماینده ذخیره کرد میتوان به این هدف رسید.

اما در حالاتی که دیتاها پراکنده باشند و تعداد خوشه‌ها زیاد باشد، این روش بهینه نیست و باعث میشود حتی تعداد نمایندگان هر خوشه با تعداد دیتاهای آموزشی اولیه نیز برابر باشد. و حتی پارامترهای خطوط نیز ممکن است زیاد باشد. اما از نظر فاصله حساب کردن در فضای n بعدی حتما بهبود خواهیم داشت.

پس در حالت کلی میتوان گفت به نحوه چنیش داده‌ها مربوط است و در شرایطی ممکن است بهتر و یا بدتر عمل کند.

ج) خیر ندارد، در این رویکرد به دلیل اینکه نحوه کارکرد آن adoptive هست در نتیجه اگر بعد از آموزش داده‌ای جدید وارد شود به راحتی میتوان آن‌را به مدل اضافه کرد و مرزبندی را آپدیت کرد.

سوال ۲)

الف) باعث میشود درخت کوچکتری و بهینه‌تری داشته باشیم که در زمان تست، سرعت بیشتر و فضای کمتری را به همراه خواهد داشت.

پیش هرس یا(Pre-Pruning (Early Stopping Rule در واقع حالتی است که الگوریتم قبل از بوجود آمدن درخت کامل با تمامی شاخه‌ها، متوقف میشود.
یکی از شروط این توقف و جلوگیری از رشد بی‌رویه درخت، توقف به محض مواجه شدن با مقادیر یکسان برای همه ویژگی ها است.
و شرط دیگر میتواند توقف هنگام عضو کلاس واحد بودن تمامی نمونه‌ها از یک ویژگی بخصوص باشد.
البته میتوان از شروط دیگری نیز استفاده کرد مثلا گذاشتن سطح آستانه و ....

در مقابل پس‌هرس یا Post-pruning زمانی رخ میدهد که میگذاریم الگوریتم کامل اجرا شود و درخت به ماکسیمم حالت خود برسد و شکل میگیرد. سپس با استفاده از رویکرد از پایین به بالا، درخت را هرس میکنیم. باتوجه به اینکه در هر مرحله از هرس کردن، درخت را در صورت کوچک‌تر شدن ارور عمومی، با زیردرخت‌های خودش جابجا کنیم.

فایده دیگری که هر دوی این رویکردها برایمان دارد جلوگیری از overfitting برای مدل در یادگیری داده‌های آموزشی هست که باعث میشود بیشتر به پیدا کردن الگوها و روابط بین داده‌ها بپردازد تا حفظ کردن داده‌های آموزشی.

ب) فرض میکنیم که درخت T از روی دیتاست D ساخته شده باشد. میدانیم که هر کدام از دیتاها در نهایت در یک برگ نهایی از درخت T قرار میگیرند. حال برای دیتاهای جدید از دیتاست D′، شروع به گسترش برگ برای هر کدام از leafهای درخت قبلی میکنیم و هر کدام را یکبار به عنوان root node فرض میکنیم.

حال این درخت جدید از ترکیب داده‌های D+D′ بوجود آمده که تعداد بیشتری شاخه و برگ نسبت به درخت اولیه(درخت T) دارد که آنرا T′ مینامیم.

تنها اطلاعاتی که از قبل برای ساختن این درخت از روی درخت T میخواهیم این است که بدانیم هر کدام از داده‌های دیتاست D′ در کدام برگ درخت T قرار میگیرد تا با استفاده از آن، ریشه جدید را از آن قسمت گسترش دهیم.

البته در این مسئله میتوان از رویکردهای bagging و boosting نیز استفاده کرد و از random forrest به منظور رای گیری بین دو درخت T و T′استفاده کرد. به این نحو که درخت دومی بعد از درخت T آموزش داده بشود و سپس نتیجه درخت اولیه و ثانویه باهم دیگر مورد قضاوت قرار گیرد.

درنهایت با هر یک از رویکردهای پیش‌هرس و پس‌هرس باعث میشود درختی با تعداد شاخه‌ها و nodeهای کمتر اما در همان فضای حالت قبلی، پدید بیاید.

سوال ۳)

باتوجه به فرض اولیه روش Naïve Bayes، فرض میکنیم که تمامی خصوصیت‌های X از یکدیگر مستقل‌اند. همچنین در نظر میگیریم که خصوصیت‌های دیتای جدید با نام X در نظر گرفته میشود. باید مقادیر P(No|X) و P(Yes|X) را محاسبه و با یکدیگر مقایسه کنیم.

$$X = (\text{تب} = \text{yes} , \text{سرفه} = \text{no} , \text{سردرد} = \text{yes})$$

$$P(X|No) = P\left(\text{تب} = \text{yes} \middle| \text{سرماخوردگی} = \text{No}\right) \times P\left(\text{سرفه} = \text{No} \middle| \text{سرماخوردگی} = \text{No}\right) \times P\left(\text{سردرد} = \text{Yes} \middle| \text{سرماخوردگی} = \text{No}\right)$$

$$= \frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} = \frac{18}{125}$$

$$P(X|Yes) = P\left(\text{تب} = \text{yes} \middle| \text{سرماخوردگی} = \text{Yes}\right) \times P\left(\text{سرفه} = \text{No} \middle| \text{سرماخوردگی} = \text{Yes}\right) \times P\left(\text{سردرد} = \text{Yes} \middle| \text{سرماخوردگی} = \text{Yes}\right)$$

$$= \frac{3}{5} \times \frac{1}{5} \times \frac{2}{5} = \frac{6}{125}$$

$$P(X|No)P(No) = \frac{18}{125} \times \frac{5}{10} = 0.072$$

$$P(X|Yes)P(Yes) = \frac{6}{125} \times \frac{5}{10} = 0.024$$

$$P(No|X) > P(Yes|X) \Rightarrow Class = No$$

در نتیجه باتوجه از روابط بدست آمده میتوان گفت برای این مریض جدید، برچسب حاوی عبارت سرما خوردگی ندارد است.

سوال ۴)

الف) به دلیل استفاده از فاصله اقلیدسی در این نوع رویکرد باید تمامی ویژگی‌ها دارای دامنه‌های یکسانی باشند.

در نهایت برای استفاده از داده‌ها در فرمول فاصله اقلیدسی، داده‌های هر ویژگی را بین صفر تا یک نرمال میکنیم: (برای شکستن داده‌ها از MinMax Scaler استفاده میکنیم تا محاسبات و حافظه کمتری اشغال کند.)

$$x_n = (x - x_{min})/(x_{max} - x_{min})$$

برای داده‌هایcategorical نیز برای خصوصیت Martial از روش One Hot Encoding استفاده میکنیم تا بتوان در فرمول از آنها به عنوان فاصله استفاده کرد.

| Record | Age(normalized) | Marital_Single | Marital_Married | Martial_Other | Income(normalized) | Risk |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 0.88 | Bad Loss |
| 2 | 0.25 | 0 | 1 | 0 | 0.01 | Bad Loss |
| 3 | 0.136 | 0 | 0 | 1 | 0.19 | Bad Loss |
| 4 | 0.659 | 0 | 0 | 1 | 0 | Bad Loss |
| 5 | 0.068 | 1 | 0 | 0 | 0.92 | Bad Loss |
| 6 | 0.386 | 1 | 0 | 0 | 0.40 | Good Risk |
| 7 | 0.727 | 1 | 0 | 0 | 0.19 | Good Risk |
| 8 | 0.75 | 0 | 1 | 0 | 1 | Good Risk |
| 9 | 0.636 | 0 | 1 | 0 | 0.90 | Good Risk |
| 10 | 1 | 0 | 1 | 0 | 0.48 | Good Risk |

مقدار X ورودی هم نرمال میکنیم:

X=( 0.18,1,0,0,0.24)

سپس با استفاده از فاصله اقلیدسی، فاصله تا هر داده را محاسبه میکنیم.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

| Record | Age(normalized) | Marital_Single | Marital_Married | Martial_Other | Income(normalized) | Distance |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 0.88 | 0.6648 |
| 2 | 0.25 | 0 | 1 | 0 | 0.01 | 1.4345 |
| 3 | 0.136 | 0 | 0 | 1 | 0.19 | 1.4158 |
| 4 | 0.659 | 0 | 0 | 1 | 0 | 1.5123 |
| 5 | 0.068 | 1 | 0 | 0 | 0.92 | 0.6892 |
| 6 | 0.386 | 1 | 0 | 0 | 0.40 | 0.2608 |
| 7 | 0.727 | 1 | 0 | 0 | 0.19 | 0.5493 |
| 8 | 0.75 | 0 | 1 | 0 | 1 | 1.7037 |
| 9 | 0.636 | 0 | 1 | 0 | 0.90 | 1.6259 |
| 10 | 1 | 0 | 1 | 0 | 0.48 | 1.6523 |

سپس چون k=3 در نتیجه ۳ همسایه نزدیکتر را انتخاب میکنیم و از بین آنها رای گیری انجام میدهیم تا در نهایت برچسب داده جدید را پیش‌بینی کنیم.

| Record | Age(normalized) | Marital_Single | Marital_Married | Martial_Other | Income(normalized) | Distance |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 0.88 | 0.6648 |
| 6 | 0.386 | 1 | 0 | 0 | 0.40 | 0.2608 |
| 7 | 0.727 | 1 | 0 | 0 | 0.19 | 0.5493 |

باتوجه به اینکه ۲ تا از ۳ تا برچسب برای داده‌ها مقدار Good Risk را دارند پس در نهایت کلاس نمونه X برابر با Good Risk میشود.

ب) باتوجه به سوال قبل و نحوه استفاده از روش Naïve Bayes خواهیم داشت:

در این سوال به دلیل پیوسته بودن ویژگی‌های Income و Age برای تخمین زدن احتمال رخداد آنها، از توزیع نرمال استفاده میکنیم. با داشتن میانگین و واریانس برای هر کدام از حالات، به احتمال مورد نیاز دست پیدا میکنیم. (برای فرمول واریانس میتوانیم از هرکدام از فرمول‌های popilation و sample که بایاس و فاقد بایاس هستند اسفاده کنیم که اینجا فرض بر دیتایsample هست.)

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$P(X_i \mid Y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(X_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

$$\text{For(Income,Risk=Bad Loss) => mean} = \mu = \frac{(46156.98+24188.1+28787.34+23886.72+47281.44)}{5} = 34060.116$$

$$\text{variance} = \sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n} = 109978875.954$$

$$\text{For(Age,Risk=Bad Loss) => mean} = \mu = \frac{(22+33+28+51+25)}{5} = 31.8$$

$$\text{variance} = \sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n-1} = 105.36$$

$$\text{P(X|Bad Loss)} = \text{P(Income} = 25000\$|\text{Risk} = \text{Bad Loss)} \times \text{P(Age} = 24|\text{Risk} = \text{Bad Loss)} =$$

$$\text{P(Income} = 25000\$|\text{Risk} = \text{Bad Loss)} = \frac{1}{\sqrt{2\pi(109978875.954)}} \exp\left[-\frac{(25000-34060.116)^2}{2\times(109978875.954)}\right] = 0.00002619$$

$$P(Age = 24|Risk = \text{Bad Loss}) = \frac{1}{\sqrt{2\pi(105.36)}} \exp\left[-\frac{(24-31.8)^2}{2\times(105.36)}\right] = 0.0291193$$

$$P(\text{X|Bad Loss}) = 0.00002619 \times 0.0291193 = 0.000000762634467$$

$$P(BadLoss|X) = P(\text{X|Bad Loss})P(BadLoss) = 0.000000762634467 \times \frac{5}{10} = 3.813172335 \times 10^{-7}$$

$$\text{For(Income,Risk=Good Risk) => mean} = \mu = \frac{(33994.9+28716.5+49186.75+46726.5+36120.34)}{5} = 38948.998$$

$$\text{variance} = \sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n} = 60509902.3536$$

$$\text{For(Age,Risk= Good Risk) => mean} = \mu = \frac{(39+54+55+50+66)}{5} = 52.8$$

$$\text{variance} = \sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n-1} = 75.76$$

$$\text{P(X|Good Risk)} = \text{P(Income} = 25000\$|\text{Risk} = \text{GoodRisk)} \times \text{P(Age} = 24|\text{Risk} = \text{GoodRisk)}$$

$$P(Income = 25000\$|Risk = GoodRisk) = \frac{1}{\sqrt{2\pi(60509902.3536)}} \exp\left[-\frac{(25000-38948.998)^2}{2\times(60509902.3536)}\right] = 0.00001027$$

$$P(Age = 24|Risk = GoodRisk) = \frac{1}{\sqrt{2\pi(75.76)}}\exp\left[-\frac{(24-52.8)^2}{2\times(75.76)}\right] = 0.00019222$$

$$P(\text{X|Good Risk}) = 0.00001027 \times 0.00019222 = 0.0000000019740994$$

$$P(\text{Good Risk}|X) = P(\text{X|Good Risk})P(\text{Good Risk}) = 0.0000000019740994 \times \frac{5}{10} = 9.870497 \times 10^{-10}$$

باتوجه به اینکه مقدار $P(BadLoss|X)$ از $P(\text{Good Risk}|X)$ بیشتر است پس برای داده جدید کلاس برابر است با Bad Loss