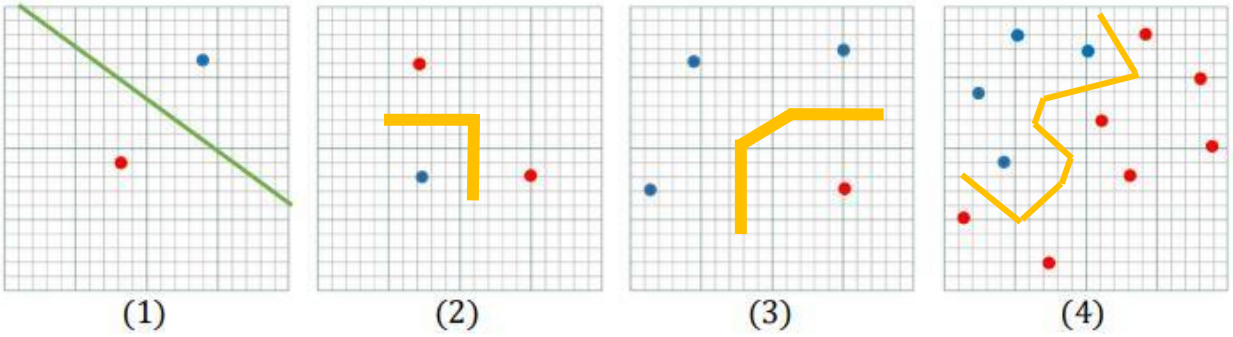
به نام خدا

تمرین دوم داده کاوی

حدیث غفوری 9825413

سوال 1

a.



b.

به طور کلی، ذخیره مرزهای تصمیم گیری برای تسک های طبقه بندی می تواند نسبت به ذخیره تمام داده های آموزشی برای رویکردهای نزدیکترین همسایه از نظر حافظه کارآمدتر باشد. این به این دلیل است که مرزهای تصمیم گیری به جای ذخیره یک مجموعه بالقوه بزرگ از نقاط داده، بر تعریف مناطقی از فضای ویژگی که با کلاس های مجزا مرتبط هستند، متکی است.

با این حال، مقدار حافظه مورد نیاز برای ذخیره سازی مرز تصمیم به پیچیدگی مسئله طبقه بندی و الگوریتم انتخاب شده برای تولید مرز تصمیم بستگی دارد.

برخی از الگوریتم‌ها ممکن است به حافظه بیشتری برای ذخیره‌سازی مرز تصمیم نسبت به سایرین نیاز داشته باشند، و برخی از مسائل طبقه‌بندی ممکن است به مرزهای تصمیم پیچیده‌تری نیاز داشته باشند که حافظه فشرده‌تری دارند.

بنابراین، همیشه تضمین نمی شود که استفاده از مرزهای تصمیم منجر به بهبود حافظه مورد نیاز برای ذخیره سازی شود، اما اغلب یک رویکرد امیدوارکننده است که میتواند در نظر گرفته شود.

c.

خیر، KNN این مشکل را ندارد.

برخلاف درخت‌های تصمیم، KNN یک الگوریتم یادگیری تنبل یا lazy learning algorithm است، به این معنی که مدلی نمی‌سازد یا هیچ فرضی در مورد داده‌ها در مرحله آموزش ایجاد نمی‌کند.

در عوض، تمام داده های آموزشی را ذخیره می کند و از آن برای پیش بینی نمونه های جدید استفاده می کند. هنگامی که داده های آموزشی جدید معرفی می شود، KNN به سادگی پایگاه داده خود را بدون هیچ گونه تغییری در الگوریتم یا داده های آموزشی قبلی به روز می کند.

این باعث می‌شود KNN در مدیریت داده‌های آموزشی جدید در مقایسه با درخت‌های تصمیم انعطاف‌پذیرتر باشد.

سوال 2

a.

در الگوریتم های درختی، هرس تکنیکی است که برای بهبود دقت و کارایی درخت تصمیم استفاده می شود. هرس شامل حذف شاخه ها یا گره های غیر ضروری از درخت برای کاهش پیچیدگی آن است که به جلوگیری از overfit شدن و بهبود توانایی های تعمیم یا generalizationکمک می کند.

دو نوع تکنیک هرس وجود دارد: پیش هرس و پس هرس.

پیش هرس شامل توقف زود هنگام ایجاد درخت است قبل از اینکه خیلی پیچیده شود.

این را می توان با تعیین حداکثر عمق درخت یا نیاز به حداقل تعداد نمونه در هر برگ به دست آورد. مزیت پیش هرس این است که از نظر محاسباتی کارآمد است و می تواند به کاهش overfittingکمک کند. با این حال، اگر درخت به شدت هرس شود و اطلاعات مهم دور ریخته شود، می تواند منجر به underfittingشود.

از طرف دیگر، پس هرس شامل ایجاد کل درخت تصمیم و سپس حذف شاخه ها یا گره هایی است که غیر ضروری به نظر می رسند.

این را می‌توان از طریق تکنیک‌هایی مانند کاهش خطای هرس، که شامل پیش‌بینی برچسب کلاس فرزند هر گره با اکثر نمونه‌هایی است که از طریق گره جریان می‌یابد، به دست آورد. اگر عملکرد درخت هرس شده به طور قابل توجهی بدتر از درخت اصلی نباشد، درخت هرس شده به عنوان مدل نهایی انتخاب می شود. مزیت روش پس هرس این است که می تواند به دقت بالاتر و مدل های قوی تر منجر شود، اما می تواند از نظر محاسباتی نیز گران باشد.

در نهایت، انتخاب پیش هرس یا پس هرس به مجموعه داده، اندازه درخت تصمیم و منابع محاسباتی موجود بستگی دارد. به طور کلی، پس هرس موثرتر در نظر گرفته می شود، اما پیش هرس می تواند گزینه خوبی برای مجموعه داده های بزرگ یا منابع محاسباتی محدود باشد.

b.

می توانیم از یک رویکرد یادگیری افزایشی استفاده کنیم که در آن نقاط داده جدید را از 'D به درخت تصمیم T موجود اضافه می کنیم.

برای این کار، می توانیم درخت تصمیم T را طی کنیم و گره های برگ را با نقاط داده جدید به روز کنیم. این فرآیند را یادگیری آنلاین (online learning) یا یادگیری افزایشی (incremental learning)می نامند.

گسترش T به اتریبیوت های انتخاب شده در درخت اصلی T، ساختار درخت و شباهت بین داده های اصلی Dو داده های جدید 'D بستگی دارد.

با این حال، افزودن بیش از حد داده های جدید به درخت تصمیم موجود می تواند منجر به overfittingشود، بنابراین نظارت بر عملکرد درخت بر روی یک مجموعه داده validation یا training برای اطمینان از اینکه هنوز پیش بینی های دقیق انجام می دهد، مهم است.

سوال 3

اگر اطلاعات مساله را با متغیرتصادفی x نشان دهیم:

X = (fever = yes, cough = no, headache = yes)

طبق قانون بیز:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

میتوانیم احتمال اینکه فرد سرماخوردگی دارد یا نه را بدست بیاوریم و احتمال هرکدام بیشتر بود، نتیجه ی پیشبینی همان میشود.

$$P\ (cold = yes \mid X) = \frac{P(X \mid cold = yes) * P(cold{=}yes)}{P(X)}$$

به دل مستقل بودن اتریبیوت ها:

$$P\ (X \mid cold = yes) = P(fever = yes \mid cold = yes) * P(cough = no \mid cold = yes) * P(headache = yes \mid cold = yes)$$

از ۵نفری که سرماخوردند، ۳نفر تب دارند پس :

$$P(\text{fever} = \text{yes} \mid \text{cold} = \text{yes}) = \frac{3}{5}$$

از ۵نفری که سرماخوردند،۱نفر سرفه نمیکند پس :

$$P(\text{cough} = \text{no} \mid \text{cold} = \text{yes}) = \frac{1}{5}$$

از ۵نفری که سرماخوردند،۲نفر سردرد دارند پس :

$$P(\text{headache} = \text{yes} \mid \text{cold} = \text{yes}) = \frac{2}{5}$$

احتمال اینکه فردی سرماخورده باشد از کل رکوردها:

$$P(\text{cold} = \text{yes}) = \frac{5}{10}$$

$$P(\text{cold} = \text{yes} \mid X) = \frac{P(X \mid \text{cold} = \text{yes}) * P(\text{cold}=\text{yes})}{P(X)} = \frac{\frac{3}{5}*\frac{1}{5}*\frac{2}{5}*\frac{5}{10}}{P(X)} = \frac{\frac{3}{125}}{P(X)}$$

$$P(\text{cold} = \text{no} \mid X) = \frac{P(X \mid \text{cold} = \text{no}) * P(\text{cold}=\text{no})}{P(X)}$$

$$P(X \mid \text{cold} = \text{no}) = P(\text{fever} = \text{yes} \mid \text{cold} = \text{no}) * P(\text{cough} = \text{no} \mid \text{cold} = \text{no}) * P(\text{headache} = \text{yes} \mid \text{cold} = \text{no})$$

از ۵نفری که سرمانخوردند، ۲نفر تب دارند پس :

$$P(\text{fever} = \text{yes} \mid \text{cold} = \text{no}) = \frac{2}{5}$$

از ۵نفری که سرمانخوردند ،۳نفر سرفه نمیکند پس :

$$P(\text{cough} = \text{no} \mid \text{cold} = \text{no}) = \frac{3}{5}$$

از ۵نفری که سرمانخوردند ،۳نفر سردرد دارند پس :

$$P(\text{headache} = \text{yes} \mid \text{cold} = \text{no}) = \frac{3}{5}$$

احتمال اینکه فردی سرمانخوردهباشد از کل رکوردها:

$$P(\text{cold} = \text{no}) = \frac{5}{10}$$

$$P(\text{cold} = \text{no} \mid X) = \frac{P(X \mid \text{cold} = \text{no}) * P(\text{cold}=\text{no})}{P(X)} = \frac{\frac{3}{5}*\frac{3}{5}*\frac{2}{5}*\frac{5}{10}}{P(X)} = \frac{\frac{9}{125}}{P(X)}$$

به دلیل مساوی بودن مخرج ها، فقط صورت‌ها را مقایسه میکنیم و چون مقدار صورت در کسر P (cold = no | X) بیشتر است از کسر P (cold = yes| X) پس نتیجه ی پیشبینی این است که فرد سرماخوردگی ندارد.

سوال 4

a.

X=(30, Single, $30000) , k=3

ابتدا باید داده های مربوط به اتریبیوت های سن و درامد را نرمال کنیم. به کمک روش MinMaxScaler عملیات نرمالایز کردن را انجام میدهیم:

مقادیر بدست امده تا ۲رقم اعشار درنظرگرفته شده اند.

X_std = (X - X.min) / (X.max- X.min)

```
def minimaxScale(x,min_val,max_val):
    return (x-min_val) / (max_val - min_val)
```

In [39]:
```
import numpy as np
arr_age = [22,33,28,51,25,39,54,55,50,66]
arr_income = [36120.34,46726.50,49186.75,28716.50,33994.90,47281.44,23886.72,28787.34,24188.10,46156.98]
```

In [40]:
```
max_age = np.max(arr_age)
max_age
```

Out[40]: 66

In [41]:
```
min_age = np.min(arr_age)
min_age
```

Out[41]: 22

In [43]:
```
for age in arr_age:
    print(f'scaled of age={age} is {minimaxScale(age,min_age,max_age)}')
```

```
scaled of age=22 is 0.0
scaled of age=33 is 0.25
scaled of age=28 is 0.13636363636363635
scaled of age=51 is 0.6590909090909091
scaled of age=25 is 0.06818181818181818
scaled of age=39 is 0.38636363636363635
scaled of age=54 is 0.7272727272727273
scaled of age=55 is 0.75
scaled of age=50 is 0.6363636363636364
scaled of age=66 is 1.0
```

```
In [44]: max_income = np.max(arr_income)
         max_income

Out[44]: 49186.75

In [45]: min_income = np.min(arr_income)
         min_income

Out[45]: 23886.72

In [46]: for income in arr_income:
             print(f'scaled of income={income} is {minimaxScale(income,min_income,max_income)}')

scaled of income=36120.34 is 0.483541719120491
scaled of income=46726.5 is 0.9027570323039142
scaled of income=49186.75 is 1.0
scaled of income=28716.5 is 0.1909001688930803
scaled of income=33994.9 is 0.39953233257035664
scaled of income=47281.44 is 0.9246913936465689
scaled of income=23886.72 is 0.0
scaled of income=28787.34 is 0.19370016557292616
scaled of income=24188.1 is 0.01191223883924238
scaled of income=46156.98 is 0.8802463870596202
```

ما یک ورودی جدید داریم اما هنوز کلاسی ندارد. برای دانستن کلاس آن، باید فاصله ورودی جدید تا سایر ورودی‌های مجموعه داده را با استفاده از فرمول فاصله اقلیدسی محاسبه کنیم.

فرمول فاصله ی اقلیدسی:

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

مقایسه برای اولین سطر با داده ی داده شده:

X=(30, Single, $30000) که اسکیل شده ی این داده هم برابر:

Age = 0.18181818181818182 , income = 0.24163133403399123 که اعداد را تا ۲رقم اعشار درنظر میگیریم پس

D1 = $\sqrt{(q1-p1)^2+(q2-p2)^2+(q3-p3)^2} = \sqrt{(0.18-0)^2+(1-1)^2+(0.24-0.88)^2} = 0.66$

D2 = $\sqrt{(0.18-0.25)^2+(1-0)^2+(0.24-0.01)^2} = 1.02$

D3 = $\sqrt{(0.18-0.13)^2+(1-0)^2+(0.24-0.19)^2} = 1.00$

D4 = $\sqrt{(0.18-0.65)^2+(1-0)^2+(0.24-0)^2} = 1.13$

D5 = $\sqrt{(0.18-0.06)^2+(1-1)^2+(0.24-0.92)^2} = 0.69$

D6 = $\sqrt{(0.18-0.38)^2+(1-1)^2+(0.24-0.39)^2} = 0.25$

D7 = $\sqrt{(0.18-0.72)^2+(1-1)^2+(0.24-0.19)^2} = 0.54$

D8 = $\sqrt{(0.18-0.75)^2+(1-0)^2+(0.24-1)^2} = 1.37$

D9 = $\sqrt{(0.18-0.63)^2+(1-0)^2+(0.24-0.9)^2} = 1.27$

D10 = $\sqrt{(0.18-1)^2+(1-0)^2+(0.24-0.48)^2} = 1.31$

از مرتب کردن این فاصله ها به صورت صعودی متوجه میشویم که ۳ همسایه با کمترین فاصله رکوردهای ۶و ۷ و ۱ به ترتیب از کمترین فاصله به بیشترین هستند.

در این ۳ رکورد کلاس ها به ترتیب good و good و bad است پس طبق رای گیری و وتینگ: کلاسی که پیشبینی میشود good است.

b.

بدست اوردن کلاس به کمک قانون بیز

X = (Age = 24, Income = 25000)

طبق فرمول توزیع نرمال برای متغیرهای پیوسته:

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

میانگین و واریانس را طبق فرمول های زیر محاسبه میکنیم:

Population mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Population variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

$$\text{P (income =25000 | Risk = Bad loss)} = \frac{e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}}}{\sqrt{2\pi\sigma_{ij}^2}} = 0.00003$$

$$\mu = \frac{46156.98+24188.10+28787.34+23886.72+47281.44}{5} = \frac{170{,}300.58}{5} = 34060.116$$

$$\sigma^2 = 109978875.9541440308$$

Normal distribution = 0.00003

P (income =25000 |Risk = Good risk) = 0.00001

$$\mu = \frac{33994.90 + 28716.50 + 49186.75 + 46726.50 + 36120.34}{5} = \frac{194{,}744.99}{5} = 38948.998$$

$\sigma^2 = 60509902.3536159992$

Normal distribution = 0.00001

احتمال اینکه در هرکلاسی، سن ۲۴ باشد چقدر است؟

ابتدا باید میانگین و واریانس اتریبیوت سن را حساب کنیم:

برای کلاس Bad Loss یعنی داده های 22 33 28 51 25:

$\mu = 31.8$

$\sigma^2 = 105.36$

برای کلاس Good Risk یعنی داده های 39 54 55 50 66:

$\mu = 52.8$

$\sigma^2 = 75.76$

$P(Age = 24 \mid Risk = Bad\ Loss) = 0.02912 = normal\ distribution$

$P(Age = 24 \mid Risk = Good\ Risk) = 0.00019 = normal\ distribution$

کسر ۱

$$P(Risk = Bad\ Risk \mid X) = \frac{P(X \mid Risk = Bad\ Risk) * P(Risk = Bad\ Risk)}{P(X)}$$

$P(X \mid Risk = Bad\ Risk) = P(Age = 24 \mid Risk = Bad\ Risk) * P(Income = 25000 \mid Risk = Bad\ Risk) = 0.02912 * 0.00003 = 0.0000008736$

کسر ۲

$$P(Risk = Good\ Risk \mid X) = \frac{P(X \mid Risk = Good\ Risk) * P(Risk = Good\ Risk)}{P(X)}$$

$$P(X \mid Risk = Good\ Risk) = P(Age = 24 \mid Risk = Good\ Risk) * P(Income = 25000 \mid Risk = Good\ Risk) = 0.00019 * 0.00001 = 0.0000000019$$

به دلیل مساوی بودن مخرج کسرهای ۱و۲ و همچنین برابر بودن احتمال $P(Risk = Bad\ Risk)$ و $P(Risk = Good\ Risk)$ به دلیل توزیع یکسان پس فقط کافی است برای بدست اوردن کلاس، احتمال های $P(X \mid Risk = Bad\ Risk)$ و $P(X \mid Risk = Good\ Risk)$ را مقایسه کنیم که عدد حاصل از اولی بزرگتر است :

$$0.0000000019 < 0.0000008736$$

پس نتیجه ای که پیشبینی میشود کلاس Bad loss است.