به نام خدا

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

پاسخ تمرین سری سوم داده کاوی

نیم‌سال اول ۰۰-۹۹

سوال ۱-

دسته بندی مشتاق سریعتر از دسته بندی تنبل است زیرا قبل از دریافت هر گونه نمونه جدید برای دسته بندی ، یک مدل کلی را ایجاد می کند. می توان به ویژگی ها وزن اختصاص داد ، که می تواند دقت دسته بندی را بهبود بخشد. از معایب دسته بندی مشتاق این است که باید به یک فرضیه متعهد شود که کل فضای نمونه را در بر بگیرد ، که موجب می شود زمان بیشتری برای آموزش لازم باشد.

دسته بندی تنبل از فضای فرضیه[۱] قوی تری استفاده می کند ، که می تواند دقت دسته بندی را بهبود بخشد. نسبت به دسته بندی مشتاق ، برای آموزش به زمان کمتری نیاز دارد. یکی از معایب دسته بندی تنبل این است که تمام داده های آموزشی باید ذخیره شوند ، که منجر به هزینه های سنگین ذخیره سازی می شود و به تکنیک های نمایه سازی کارآمد نیاز دارد. از دیگر معایب آن این است که در دسته بندی کندتر است زیرا دسته بندی کننده ای ساخته نمی شوند تا زمانی که نیاز به دسته بندی نمونه های جدید باشد. علاوه بر این ، ویژگی ها همه به یک اندازه وزن دارند که می تواند دقت دسته بندی را کاهش دهد. (ممکن است به دلیل ویژگی های نامرتبط موجود در داده ها مشکلاتی بوجود آید.)

سوال ۲-

الف) در اینجا چون min-sup = 3/5 است و ۵ تراکنش داریم، بنابراین supcount یک frequent itemset باید بزرگتر مساوی ۳ باشد.

Apriori:

Candidate Set C1:

Item Set	Sup Count
M	3
O	3
N	2
K	5
E	4
Y	3
D	1
A	1
U	1
C	2

[۱] hypothesis space

I	1

Item Set L1:

Item Set	Sup Count
M	3
O	3
K	5
E	4
Y	3

Candidate Set C2:

Item Set	Sup Count
M, O	1
M, K	3
M, E	2
M, Y	2
O, K	3
O, E	3
O, Y	2
K, E	4
K, Y	3
E, Y	2

Item Set L2:

Item Set	Sup Count
M, K	3
O, K	3
O, E	3
K, E	4
K, Y	3

Candidate Set C3:

Item Set	Sup Count
M, K, O	0
M, K, E	2
M, K, Y	2
O, K, E	3
O, K, Y	2
K, E, Y	2

Item Set L3:

Item Set	Sup Count
O, K, E	3

Item Set L4 = Candidate Set C4 = {}

complete set of frequent itemsets: {E, K, M, O, Y, EK, EO, KM, KO, KY, EKO}

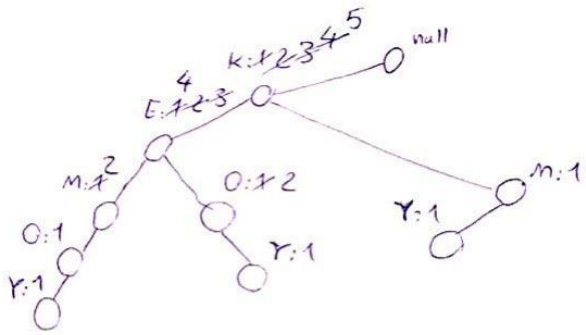
FP-growth:

TID	Items bought
T1	M, O, K, N, E, Y
T2	D, O, N, K, E, Y
T3	M, A, K, E
T4	M, U, C, K, Y
T5	C, O, O, K, I, E

descending order

K, E, M, O, Y
K, E, O, Y
K, E, M
K, M, Y
K, E, O

Itemset	Sup-Count
M	3
O	3
N	2
K	5
E	4
Y	3
D	1
A	1
U	1
C	2
I	1



↳ descending order + removing below 3s

Itemset	Sup-Count
K	5
E	~~3~~ 4
M	3
O	3
Y	3

Item	Conditional Pattern Base	Conditional FP-tree	Frequent Pattern Generated
Y	{{K, E, M, O: 1}, {K, E, O: 1}}	—	—
O	{{K, E, M: 1}, {K, E: 2}}	<K: 3, E: 3>	{K, O: 3}, {E, O: 3}, {K, E, O: 3}
M	{{K, E: 2}, {K: 1}}	<K: 3>	{K, M: 3}
E	{{K: 4}}	<K: 4>	{K, E: 4}
~~K~~			

ب) به طور کلی الگوریتم APriory ، در هر محاسبه candidate itemset را محاسبه کرده و سپس با prune کردن آن، آنهایی که sup کمتر از مینیموم دارند حذف میکند اما در الگوریتم FP-growth دیگر کاندیداها ساخته نشده (عملا برای بهبود Apriory روی کار آمد) و از همان ابتدا هرجیزی ساخته میشود sup بیشتری نسبت به مینیموم دارد. به طور جزئی تر:

۱- FP-growth الگوریتم سریعتری نسبت به Apriory است.

۲- FP=growth تنها ۲ تا database scan نیاز دارد اما Apriory چندین بار برای تولید مجموعه کاندیداها

۳- FP-growth حافظه کمتری مصرف میکند.

۴- FP-growth دقیق تر است.

سوال ۳-

یادگیری نیمه‌نظارتی:
در روش یادگیری نیمه‌نظارتی برای ساخت مدل دسته‌بند از هر دو بخش داده‌های برچسب‌خورده و برچسب‌نخورده استفاده می‌کنیم. وقتی حجم داده‌های برچسب‌نخورده بسیار بالاست و هزینه‌ی برچسب‌گذاری آن‌ها هم بسیار زیاد باشد از این روش‌ها استفاده می‌شود. به طور مثال در تشخیص کاربران یا عملیات‌های مخرب در شبکه‌های اجتماعی که تشخیص و برچسب‌گذاری داده‌ها و عملیات‌ها دشوار است می‌توان از این روش استفاده کرد. یکی از چالش‌های اصلی این روش پیداکردن ویژگی‌های مناسب در داده برای بهبود عملکرد دسته‌بند است.

یادگیری فعالانه:
در روش یادگیری فعالانه از نیروی انسانی برای برچسب گذاری داده‌ها استفاده می‌شود. بعد از جمع‌آوری داده‌ها تمامی داده‌های برچسب‌دار را برای یادگیری به مدل می‌دهیم. از این روش در مواردی که حجم داده بسیار بالا و پراکنده استفاده می‌شود. به طور مثال در مواردی که داده‌ها بر روی سرویس وب وجود دارند می‌توان از کاربران استفاده کرد تا بخشی از داده‌ها را برچسب‌گذاری کنند.

یادگیری انتقالی:
در مواقعی که داده‌های برچسب‌خورده به اندازه کافی نداریم می‌توانیم از مدل‌هایی استفاده کنیم که بر روی تسک‌های دیگر مربوط به همان حوزه مورد نظر آموزش دیده‌اند. با این روش می‌توان از دانش کسب‌شده این مدل‌ها برای تسک مورد نظر جدید استفاده کرد. کاربرد این روش در پردازش زبان و تصویر زیاد است. به طور مثال از مدل‌هایی که قبلا برای دسته‌بندی متن آموزش دیده‌اند برای تشخیص اسپم در ایمیل استفاده می‌کنند.

از چالش‌های مهم این حوزه یافتن یک مدل انتقالی است که بتواند عملکرد خوبی بر روی تسک جدید داشته باشد زیرا برای بهبود عملکرد دانش قبلی مدل باید با تسک جدید ارتباط معناداری داشته باشد.

سوال ۴-

دو دسته اصلی یادگیری گروهی bagging و boosting می باشند.

طی روش bagging از ترکیب نظر همه ی classifier ها که با داده های bootstrap آموزش داده شده اند استفاده می شود. یعنی به صورت همزمان، داده های مربوطه به دسته بندها داده شده و نظر هر دسته بند گرفته می شود. نهایتا رای غالب یا میانگین میان این دسته بندها به عنوان خروجی انتخاب می گردد.

طی روش boosting مدل های مختلف به صورت مستقل بررسی نمی گردند. یعنی دسته بندها به صورت مکمل و پشت هم آموزش داده می شوند برای مثال، بر روی داده هایی که توسط دسته بندهای قبلی به اشتباه دسته بندی می شوند وزن بیشتری در نظر گرفته می شود. در حالت کلی از روش های filtering و یا weighting برای تمرکز بر داده های اشتباه دسته بندی شده یا حذف توجه به داده های درست دسته بندی شده استفاده می گردد. این وزن دهی می تواند در تابع loss مدل ها دیده شود.

سوال ۵-

الف) می توان از روش های one vs one و یا one vs rest استفاده نمود. برای مثال در روش one vs rest ابتدا یکی از دسته ها را با استفاده از SVM از قبل معرفی شده جداسازی نموده و سایر M - 1 دسته را نیز به همین ترتیب جداسازی می کنیم تا به تعداد دسته ی هدف برسیم.

ب) هدف مدل SVM با حاشیه سخت آن است که دو کلاس هدف به دسته های کاملا جداگانه و مجزا و با بیشینه margin تقسیم گردند. پس شروط مساله سختگیرانه تر می باشند و این موضوع باعث حساسیت بالای این روش به نویز و outlier خواهد شد.

ج)

۱. می توان پارامترهای مدل خطی را افزایش داد.

۲. استفاده از تعداد ویژگی های بیشتر.

۳. در صورت امکان، حذف داده هایی همچون نویز که موجب دشواری تقسیم بندی مجموعه داده ورودی به صورت خطی می شوند.

۴. می توان از مدل های چندجمله ای و غیرخطی استفاده نمود. به این ترتیب قابلیت برازش به داده های آموزشی توسط مدل بالا خواهد رفت.

سوال ۶-

سوال 6 :

در ابتدا میزان تکرار تکی هر نوع کتاب خریداری شده را محاسبه می‌کنیم.

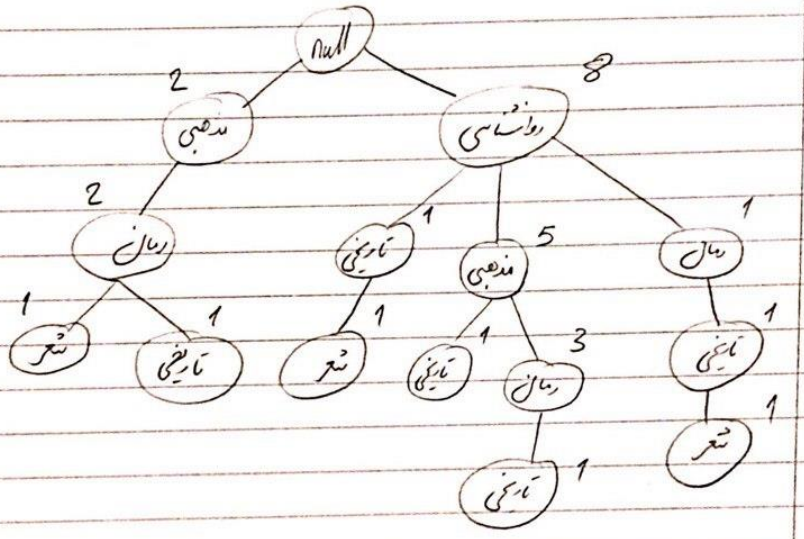
روانشناسی : 8

مذهبی : 7

رمان : 6

تاریخی : 5

شعر : 3



Item	Conditional Pattern Base	Conditional FP. tree	Frequent Patterns
رمان	{مذهبی : 2}	{روانشناسی : 4 ، مذهبی : 3}	{رمان ، روانشناسی : 4}
	{روانشناسی : 1}	{مذهبی : 2}	{رمان ، مذهبی ، روانشناسی : 5}
	{روانشناسی ، مذهبی : 3}		{رمان ، مذهبی ، روانشناسی : 3}

ب) ریاضی ⟹ {روانشناسی، مذهبی} = $\frac{3}{6}$ ✓ ریاضی ⟹ روانشناسی = $\frac{4}{6}$ ✓

روانشناسی ⟹ {ریاضی، مذهبی} = $\frac{3}{8}$ X $< \frac{1}{2}$ روانشناسی ⟹ ریاضی = $\frac{4}{8}$ ✓

مذهبی ⟹ {روانشناسی، ریاضی} = $\frac{3}{7}$ X $< \frac{1}{2}$ مذهبی ⟹ ریاضی = $\frac{5}{7}$ ✓

{روانشناسی، مذهبی} ⟹ ریاضی = $\frac{3}{5}$ ✓ ریاضی ⟹ مذهبی = $\frac{5}{6}$ ✓

{ریاضی، مذهبی} ⟹ روانشناسی = $\frac{3}{5}$ ✓

{روانشناسی، ریاضی} ⟹ مذهبی = $\frac{3}{4}$ ✓