به نام خدا

تمرین سری اول داده کاوی

توضیحات:

- پاسخ به تمرین‌ها باید به صورت انفرادی صورت گیرد و در صورت مشاهده هرگونه تقلب نمره صفر برای کل تمرین منظور خواهد شد.
- تمیزی و خوانایی گزارش تمرین از اهمیت بالایی برخوردار است.
- لطفا گزارش تمرین خود را در قالب یک فایل PDF با نام **«HW…_StudentNumber.pdf»** در سایت درس در مهلت معین بارگزاری نمایید.
- در صورت داشتن اشکال می‌توانید از طریق ایمیل **datamining.spring2021@gmail.com** با تدریس‌یار درس در ارتباط باشید.

نیم‌سال دوم ۰۰-۹۹

بخش نوشتاری

سوال ۱) مفاهیم زیر را به طور کامل تعریف نمایید.

۱- Unsupervised Learning
۲- Supervised Learning
۳- Semi-supervised Learning
۴- Outlier
۵- Dimension
۶- Training, Validating and Testing Data
۷- Data Warehousing
۸- Missing Values
۹- Independent Variable

سوال ۲) کاهش بعد[۱]، یکی از تکنیک‌های رایج در داده‌کاوی برای بهره‌وری بهتر از داده‌هاست. بسیاری از ویژگی‌ها، اثربخشی چندانی ندارند و حضور یا عدم حضور آنها، حائز اهمیت نیست. لذا با استفاده از تکنیک‌هایی می‌توان ابتدا پیدا کرد که چه ویژگی‌هایی را می‌توان حذف کرد و سپس ویژگی‌های مهم را از آنها جدا کرد. به حداقل دو مورد از این تکنیک‌ها اشاره کرده و برای یکی از آنها، با ذکر مثال، تکنیک را توضیح دهید. در ادامه فرق انتخاب ویژگی[۲] و استخراج ویژگی[۳] را توضیح دهید.

سوال ۳) معیارهای ارزیابی دقت[۴]، فراخوانی[۵] و امتیاز اف[۶] را برا اساس ماتریس درهم ریختگی[۷] معرفی کنید.

سوال ۴) فرض کنید که همبستگی بین دو متغیر صفر است. مفهوم آن چیست؟ باتوجه به تعریف در سوال اول، آیا این متغیرها مستقل از یکدیگر هستند؟

---

[۱] Dimensionality Reduction
[۲] Feature Selection
[۳] Feature Extraction
[۴] Precision
[۵] Recall
[۶] F-score
[۷] Confusion Matrix

سوال ۵) پیش‌پردازش داده‌ها از جمله موارد پراهمیت در انجام پروژه‌های مبتنی بر یادگیری است. در مبحث یادگیری زبان طبیعی، از آنجایی که جنس داده‌ها بر خلاف اکثر مواقع، داده‌های متنی است، در نتیجه نوع پیش‌پردازش متفاوتی را در بر می‌گیرد که به مواردی همچون حذف ایست‌واژه‌ها، پاک کردن فضاهای خالی و غیره می‌توان اشاره کرد اما در درس داده‌کاوی، بیشتر تمرکز بر روی داده‌های عددی است. مجموعه تکنیک‌هایی همچون پاک‌سازی داده‌ها[۸]، ادغام داده‌ها[۹] و تبدیل داده‌ها[۱۰] می‌توان اشاره کرد. این سه تکنیک ذکر شده را توضیح دهید.

سوال ۶) الگوریتم Apriori را بر روی تراکنش‌های زیر اجرا کنید. فرض کنید آستانه پشتیبانی[۱۱] برابر با ۳۳٪ و آستانه اطمینان[۱۲] ۶۰٪ می‌باشد. تمامی مراحل تولید مجموعه آیتم‌های کاندید را نشان دهید و در نهایت مجموعه آیتم‌های پرتکرار را بدست آورید. همچنین تمام قواعد انجمنی قابل تولید از این مجموعه آیتم‌ها را نوشته، آن‌هایی که مطمئن هستند را مشخص کرده و بر اساس اطمینان مرتب کنید.

ID	Items
T1	نان، الویه، پنیر
T2	نان، الویه
T3	نان، کره، مربا
T4	مربا، کره
T5	مربا، پنیر
T6	نان، کره، مربا

سوال ۷)

الف) با استفاده از تراکنش‌های سوال قبل و با همان آستانه پشتیبانی، یک درخت الگو پرتکرار[۱۳] بسازید. نشان دهید با هر تراکنش چگونه درخت گسترش می‌یابد.

ب) با استفاده از الگوریتم FP-Growth، مجموعه آیتم‌های پرتکرار در این درخت را بیابید.

---

[۸] Data Cleaning
[۹] Data Integration
[۱۰] Data Transformation
[۱۱] Support
[۱۲] Confidence
[۱۳] FP-Tree

بخش پیاده‌سازی

پیش‌پردازش

هدف از انجام این بخش، آشنایی با تکنیک‌های پیش‌پردازش داده‌ها و استفاده از داده‌هاست. به طور کلی دو کتابخانه مطرح در این قسمت استفاده می‌شود:

۱) Pandas
۲) Seaborn

ابتدا دستورات موجود در این کتابخانه‌ها را یادگرفته سپس به قسمت بعدی مراجعه نمایید. بازی فیفا ۲۰۲۱ که چند ماهی بیشتر از انتشار آن نمی‌گذرد، از سری بازی‌های جذاب کنسول‌های بازی است :) اما در این قسمت ما صرفاً قرار است از اطلاعات بازیکن‌های این بازی استفاده کنیم نه خود بازی! :دی

مجموعه داده ضمیمه شده در این تمرین، مجموعه داده اطلاعات بازیکنان این بازی است. حال موارد زیر را به ترتیب بر روی مجموعه داده‌ها اعمال نمایید.

۱) مجموعه داده players.csv را خوانده و ابتدا و انتهای آن را نمایش دهید.

۲) طبق تعریف بخش نوشتاری، Missing Value ها را پیدا کنید.

۳) میانگین، حداکثر و حداقل وزن بازیکنان را بدست آورید.

۴) کدام کشور دارای بیشترین بازیکن و کدام کشور دارای کمترین بازیکن است؟ تعداد هر کدام را گزارش کنید.

۵) آینده‌دارترین بازیکن‌ها را بر اساس Growth > 4 و Potential > 84 بدست آورده و گزارش کنید.

۶) نمودار این بازیکنان آینده‌دار را بر اساس موقعیت‌شان در بازی را گزارش کنید.

۷) کدام باشگاه دارای بیشترین تعداد بازیکن آینده‌دار است؟ چه تعداد بازیکن؟

۸) مجموع ارزش بازیکن‌های آینده‌دار باشگاه فوتبال چلسی[۱۴] چقدر است؟

۹) تعداد بازیکنانی که در سال ۲۰۲۱ قراردادشان با باشگاه‌شان تمام می‌شود و همچنین در تیم ملی کشورشان حضور ندارند، چندتاست؟ گزارش کنید.

۱۰) موقعیت، درآمد و باشگاه فعلی مهدی طارمی را گزارش کنید.

---

[۱۴] Chelsea

قوانین انجمنی

در این قسمت قصد داریم ابتدا با محیط نرم‌افزاری ابزار Weka آشنا شویم. سپس نحوه استفاده از کتابخانه Weka را فرا خواهیم گرفت.

۱) دستورات فایل weka_guide.pdf را دنبال کنید و به سوال انتهای آن پاسخ دهید.

۲) از کد AssociationRulesMining.java به عنوان template برای انجام موارد بعدی استفاده کنید. همچنین می‌توانید از کدهای موجود در پوشه Examples کمک بگیرید.

۳) با استفاده از کتابخانه وکا، برنامه‌ای به زبان جاوا بنویسید که مجموعه داده supermarket را لود کند و به ازای مقادیر پارامترهای زیر و با رسم نمودار زمان اجرای دو الگوریتم را با هم مقایسه کند.

Sup_{min} = [0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.11, 0.12, 0.13, 0.14, 0.15]
$Conf_{min}$ = 0.9
Number of Rules = 100000

۴) (امتیازی) با استفاده از کتابخانه وکا، برنامه‌ای به زبان جاوا بنویسید که مجموعه داده supermarket را بارگذاری کند و به ازای مقادیر پارامترهای زیر و با استفاده از الگوریتم FP-Growth در یک فرآیند دو مرحله‌ای ابتدا تمامی الگوهای پرتکرار[۱۵] را تولید نماید و سپس از میان آن‌ها تمامی الگوهای بسته[۱۶] و الگوهای بیشینه[۱۷] را در یک فایل txt خروجی دهد. نیازی به در نظر گرفتن الگوهای پرتکرار با طول ۱ نمی‌باشد.

Sup_{min} = 0.1

نکته: گزارش نویسی صحیح بخش قابل توجهی از نمره قسمت پیاده‌سازی را در بر می‌گیرد پس در انجام آن کوتاهی نکنید. نگارش صحیح، زیبایی ظاهر و کامل بودن گزارش ارزش کمتری از صحت علمی گزارش ندارد.

[۱۵] Frequent Pattern
[۱۶] closed-pattern
[۱۷] max-pattern