به نام ایزد منان

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

تمرین سری اول
داده‌کاوی

توضیحات:

- پاسخ به تمرین‌ها باید به صورت انفرادی صورت گیرد. حداقل برخورد با پاسخ‌های مشابه، تخصیص نمره کامل منفی به طرفین خواهد بود.
- پاسخ‌های خود را به زبان فارسی و به صورت مرتب، در قالب یک فایل فشرده (.zip) با الگوی زیر در صحفه‌ی درس بارگذاری کنید:

DM_HW[No]_[Student_number].pdf

- لطفاً نظم، ساختار و توالی سوالات را در پاسخ‌ها رعایت کنید.
- تمیزی و خوانایی گزارش تمرین از اهمیت بالایی برخوردار است.
- برای تمرین‌های عملی، علاوه بر کد گزارش کتبی نیز ارسال کنید.
- در صورتی که درمورد این تمرین سوال یا ابهامی داشتید با ایمیل dm.1401.spring@gmail.com با تدریس‌یاران درس در ارتباط باشید.
- مهلت ارسال تمرین تا ساعت ۱۱:۵۵ روز جمعه مورخ ۱۹ فروردین ۱۴۰۱ است.

نیم‌سال دوم ۱۴۰۰-۱۴۰۱

فهرست مطالب

صفحه

بخش نوشتاری

سوال اول

مفاهیم زیر را تعریف کنید.

۱- Dimension
۲- Outlier
۳- Independent variable
٤- Dependent variable
٥- Stratified Sampling

سوال دوم

پیش‌پردازش داده‌ها از جمله موارد پراهمیت در انجام پروژه‌های مبتنی بر یادگیری است و نرمال‌سازی[1] داده‌ها یکی از مهم‌ترین مراحل پیش‌پردازش است. سه مورد از روش‌های نرمال‌سازی را با ذکر مثال توضیح دهید. سپس محدوده‌ی نرمال‌سازی هر یک را مشخص کنید.

سوال سوم

تکنیک ChiMerge یک الگوریتم خودکار گسسته‌سازی[2] تحت نظارت، مبتنی بر ادغام از پایین به بالا است که با استفاده از آماره‌ی x^2 کار خود را انجام می‌دهد. فواصل مجاور با حداقل مقادیر x^2 با هم ادغام می‌شوند تا زمانی که معیار توقف انتخاب شده برآورده شود. به طور خلاصه نحوه عملکرد ChiMerge را شرح دهید.

[1] Normalization

[2] Discretization

سوال چهارم

برای بردارهای داده شده، موارد خواسته شده را بدست آورید.

- $x = [1,1,1,1], y = [2,2,2,2]$
  Cosine similarity, Correlation, Euclidean distance.
- $x = [0,1,0,1], y = [1,0,1,0]$
  Cosine similarity, Correlation, Euclidean distance, Jaccard distance.
- $x = [1,1,0,1,0,1], y = [1,1,1,0,0,1]$
  Correlation, Manhattan distance, Bhattacharya distance.
- $x = [2,-1,0,2,0,-3], y = [-1,1,-1,0,0,-1]$
  Cosine similarity, Correlation.

سوال پنجم

کاهش داده[3] یکی از عملیات‌های اصلی در پیش‌پردازش داده در داده‌کاوی به‌شمار می‌رود. هدف از انجام تکنیک کاهش داده چیست؟ راهبردهای آن را توضیح دهید.

سوال ششم

کاهش بعد[4] یکی از تکنیک‌های رایج در داده‌کاوی است و روش‌های گوناگونی برای آن وجود دارد. تبدیل موجک[5] یکی از تکنیک‌هایی است که برای راهبرد کاهش بعد انجام می‌گیرد. آن را به اختصار توضیح دهید. در ادامه تفاوت feature selection و feature extraction را بیان کنید.

سوال هفتم

اگر داده‌های زیر را به روش box plot نمایش دهیم، min و max چه اعدادی خواهند بود؟ همچنین داده‌های پرت را نیز مشخص کنید.

70,56,71,73,74,144,89,80,90,143,89,80,90,143,146,100,20,44,74

[3] Data Reduction

[4] Dimensionality Reduction

[5] Wavelet Transform

سوال هشتم

در رابطه با noise و outlierها به سوالات زیر پاسخ کامل همراه با توضیح ارائه دهید.

۱- آیا noise همیشه مطلوب هست؟ outlierها چطور؟

۲- آیا noise objects می‌توانند outlier باشند؟

۳- آیا outlierها همیشه noise objects هستند؟

٤- آیا noise می‌تواند یک مقدار معمولی را به یک مقدار غیرمعمول تبدیل کند یا برعکس؟

سوال نهم

در رابطه با cosine measure و correlation به سوالات زیر پاسخ دهید.

۱- محدوده مقادیر ممکن برای cosine measure چقدر است؟

۲- اگر cosine measure دو object برابر یک باشد، آیا آنها یکسان هستند؟

۳- چه رابطه‌ای بین cosine measure و correlation وجود دارد؟

سوال دهم

نمودار quantile–quantile و quantile را با هم مقایسه کنید.

سوال یازدهم

به طور خلاصه نحوه محاسبه عدم تشابه[6] بین اشیا توصیف شده توسط دو ویژگی numeric و nominal را شرح دهید.

[6] Dissimilarity

بخش پیاده‌سازی

پیش‌پردازش داده‌ها برای مدل‌های یادگیری ماشینی یک مهارت اصلی برای هر دانشمند داده[۷] یا مهندس یادگیری ماشین[۸] است. به طور کلی دو کتابخانه مطرح pandas و Scikit-learn برای پیش‌پردازش داده استفاده می‌شود که ما در این بخش به بررسی این کتابخانه‌ها می‌پردازیم.

در یک پروژه علم داده در دنیای واقعی، پیش‌پردازش داده‌ها یکی از مهم‌ترین گام‌های آن است و یکی از عوامل مشترک موفقیت یک مدل است، یعنی اگر پیش‌پردازش داده‌ها و مهندسی ویژگی‌ها[۹] درست باشد، احتمال موفقیت آن مدل در مقایسه با مدلی که داده‌ها برای آن به خوبی پیش‌پردازش نشده‌اند، بیشتر است و نتایج بهتری تولید خواهد کرد.

۰- مجموعه‌داده:

مجموعه‌داده‌ی در نظرگرفته شده برای این تمرین، مجموعه داده‌ی iris است، این مجموعه‌داده به همراه تمرین برای شما قرار داده شده است. این مجموعه داده برای شناسایی سه نوع گل (iris-setosa، iris-versicolor و iris-virginica) جمع‌آوری شده است و برای هر گل چهار ویژگی ذکر شده است که در ادامه این ویژگی‌ها به ترتیب بیان شده‌است:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm

برای بارگذاری این مجموعه داده از قطعه کد زیر استفاده کنید:

```
## loading the dataset
import pandas as pd
df = pd.read_csv(dataset_path,
names=['sepal_length','sepal_width','petal_length','petal_width','target'])
```

که در این قطعه کد dataset_path آدرس مجموعه‌ی داده مذکور در سیستم شماست.

---

[7] Data Scientist

[8] Machine Learning Engineer

[9] Feature Engineering

۱- اهمیت داده‌های ازدست‌رفته[10]:

یک عبارت معروف در یادگیری ماشینی وجود دارد که ممکن است آن را شنیده باشید:

Garbage in, Garbage out.

اگر مجموعه داده‌های شما مملو از NaN و مقادیر زباله باشد، مطمئناً مدل شما نیز نتیجه‌ی قابل قبولی ندارد. بنابراین مقابله با چنین داده‌هایی مهم است.

سوال) ابتدا به دنبال داده‌های NaN در مجموعه‌داده بگردید و ذکر کنید که از هر ویژگی چند سطر فاقد داده هستند. برای اینکار از تابع isna() استفاده کنید.

یک روش برای پرکردن مقادیر از دست‌رفته، پرکردن آن با میانگین، میانه، واریانس آن ستون یا مقداری ثابت است. برای انجام این کار، می‌توانیم از SimpleImputer از sklearn.impute استفاده کنیم. البته در سه مورد اول باید داده‌های ما از نوع عدد باشند و تنها در مورد چهارم (جایگذاری با مقدار ثابت) می‌توان برای داده‌هایی از نوع str نیز استفاده کرد.

اگر تعداد سطرهایی با مقادیر از دست‌رفته کم باشد، یا داده‌های ما به گونه‌ای است که توصیه نمی‌شود مقادیر از دست‌رفته را پر کنید، می توانیم با استفاده از dropna در پاندا، ردیف‌های از دست‌رفته را حذف کنیم.

سوال) داده‌های از دست‌رفته در مجموعه‌داده را با استفاده از dropna حذف کنید.

۲- داده‌های غیرعددی:

به طورکلی در علم داده، مدل‌های ما قادر به درک یک داده‌ی متن نیستند و لازم است که این داده‌ها به عدد تبدیل شود. برای تبدیل ویژگی‌های کلاس‌بندی شده[11] می‌توان از دو روش Label Encoding و یا One Hot Encoding استفاده کرد.

در Label Encoder می‌توانیم مقادیر Categorical را به برچسب‌های عددی تبدیل کنیم.

سوال) با استفاده از Label Encoder در ستون target، Iris-setosa را به ۰، Iris-versicolor را به ۱ و Iris-virginica را به ۲ تبدیل کنید (برای این کار از LabelEncoder در sklearn.preprocessing استفاده کنید). ایرادی که ممکن است این روش داشته باشد را بیان کنید..

روش دیگر استفاده از OneHotEncoder است.

سوال) در رابطه با این روش توضیح دهید و یک مثال برای درک بهتر بیان کنید.

۳- نرمال‌سازی:

---

[10] Missing Values

[11] categorical features

از آزمایش‌های مشخصی ثابت شده است که مدل‌های یادگیری ماشین و یادگیری عمیق در مقایسه با مجموعه داده‌ای که نرمال‌سازی نشده‌اند، در یک مجموعه داده نرمال‌شده عملکرد بهتری دارند. هدف نرمال‌سازی تغییر مقادیر به یک مقیاس مشترک است. چندین راه برای این‌کار وجود دارد.

سوال) با استفاده از `StandardScaler` در `sklearn.preprocessing` اقدام به نرمالسازی داده‌ها کنید. مقدار واریانس و میانگین هر ستون را قبل از نرمالسازی و پس از آن ذکر کنید (دقت کنید که این نرمالسازی را بر روی برچسب داده‌ها انجام ندهید).

۴- تحلیل مولفه‌های اصلی[12]:

برای بسیاری از پروژه‌های یادگیری ماشین، به تجسم داده‌ها به درک بهتر پروژه کمک می‌کند. تجسم داده های ۲ یا ۳ بعدی چندان چالش برانگیز نیست. همچنین در بعضی از پروژه‌های یادگیری ماشین، ویژگی‌های استخراج شده، ویژگی‌های اضافی هستند و می‌توان آن‌ها را کاهش داد. تحلیل مولفه‌های اصلی یا همان PCA به ما کمک می‌کند تا بردار ویژگی‌های خود را از یک فضای n بعدی به k بعدی تبدیل کنیم.

سوال) با استفاده از `PCA` در `sklrean.decomposition` مولفه‌های اصلی داده‌ها را حساب کنید و بردار ویژگی‌ها از یک فضای ۴ بعدی به ۲ بعدی کاهش دهید (پیش‌نیاز این کار نرمالسازی داده‌هاست).

۵- مصورسازی:

همانطور که در قسمت قبل گفته شد، تجسم داده‌ها برای فهم بهتر پروژه به ما کمک خواهد کرد، در این قسمت اقدام به رسم داده‌های مجموعه‌ی داده خود خواهیم کرد.

برای رسم ویژگی‌ها، از ویژگی‌های استخراج شده در قسمت قبل استفاده کنید.

سوال) با استفاده از کتابخانه‌ی `matplotlib` داده‌های مجموعه داده را رسم کنید. دقت کنید برای ویژگی‌ها از ویژگی‌های حاصل از `PCA` استفاده کنید (محور افقی اولین ویژگی و محور عمودی دومین ویژگی‌باشد) و برای هر کلاس رنگ متفاوتی استفاده کنید.

سوال) برای هر چهار ویژگی ارائه شده در مجموعه‌داده، نمودار `box plot` را رسم کنید (داده‌های `missing value` را حذف کنید).

[12] PCA