به نام خدا

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

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.
- گزارش تمرین خود را در قالب یک فایل PDF با نام **«HW1_StudentNumber.pdf»** به همراه کد های بخش پیاده سازی (فایل های ipynb. یا py.) در فایلی به نام **«HW1_StudentNumber.zip»** قرار داده و در سایت درس در مهلت معین بارگزاری نمایید.
- در صورت داشتن اشکال می‌توانید از طریق ایمیل **datamining.fall2020@gmail.com** با تدریس‌یاران درس در ارتباط باشید.
- همچنین لازم بذکر است که اگر مواردی در کلاس تدریس نشده انتظار می رود که خود دانشجویان جستجو کنند و انجام دهند.

نیم‌سال اول ۰۰-۹۹

سوال ۱- ویژگی های زیر را به صورت دودویی[۱] ، گسسته یا پیوسته طبقه بندی کنید. همچنین آنها را به صورت کیفی (اسمی یا ترتیبی) یا کمی (بازه[۲] یا نسبت) طبقه بندی کنید. برخی از موارد ممکن است بیش از یک حالت داشته باشند ، بنابراین اگر فکر می کنید ابهامی وجود دارد به طور خلاصه استدلال خود را نشان دهید.

مثال: سن به سال

پاسخ: گسسته، کمی-نسبت

الف) زاویه اندازه گیری شده بین ۰ و ۳۶۰ درجه

ب) ارتفاع از سطح دریا

ج) تراکم ماده به گرم در سانتی متر مکعب

د) مدل های طلا و نقره و برنز در بازی المپیک

ح) روشنایی اندازه گیری توسط نور سنج

ت) روشنایی اندازه گیری توسط قضاوت ناظر انسانی

سوال ۲- با توجه به تفاوت های نویز و outlier به سوالات زیر با دلیل پاسخ دهید.

الف) آیا نویز می تواند مطلوب باشد؟ outlier چطور؟

ب) آیا اشیا نویز[۳] می توانند outlier باشند؟

ج) آیا اشیا نویز همیشه outlier هستند؟

د) آیا Outlier ها همیشه نویز هستند؟

سوال ۳- با توجه به وکتور های x,y معیار های گفته شده برای آنها را محاسبه کنید.

---

۱ binary
۲ interval
۳ Noise object

الف) x=(0, -1, 0, 1), y=(1, 0, -1, 0) معیار های cosine, correlation, Euclidean

ب) x=(2, -1, 0, 2, 0, -3), y=(-1, 1, -1, 0, 0, -1) معیار های cosine, correlation, Euclidean

ج) x=(1, 1, 0, 1, 0, 1), y=(1, 1, 1, 0, 0, 1) معیار های cosine, correlation, Jaccard

سوال ۴- برای رفع هر یک از چالش های زیر دو راه حل ارائه دهید.

الف) وجود تاپل هایی با مقادیر ویژگی های بخصوص حذف شده در مجموعه داده.

ب) پردازش مجموعه داده های با ابعاد بسیار بالا تا هزار ویژگی و نحوه انتخاب ویژگی های پراهمیت.

ج) مجموعه داده های نامتوازن (اختلاف بالا میان تعداد تاپل های با برچسب متفاوت)

د) رخداد بیش برازش در مدل های یادگیری

ه) رخداد کم برازش در مدل های یادگیری

سوال ۵- به سوالات زیر در ارتباط با رگرسیون خطی پاسخ دهید.

الف) آیا این الگوریتم نسبت به outlierها حساسیت دارد؟ توضیح دهید.

ب) معیار اصلی اندازه گیری خطا در این الگوریتم چیست و چرا؟

ج) جدول زیر مقادیر متناظر قد و وزن ده نفر را نشان می دهد. معادلات خطوط کمترین مربعات را به طور تقریبی بدست آورید.

قد (m)	1.58	1.60	1.62	1.65	1.68	1.70	1.74	1.75	1.77	1.80
وزن (kg)	57.5	58.2	59.5	62.1	63.4	64.5	66.2	67.7	69.4	71.3

سوال ۶- در یک نظرسنجی انجام شده از ۲۰۰ هزار نفر در آمریکا، تمایل آنها به یکی از دو حزب جمهوری‌خواه و دموکرات پرسیده شده است. در این نظرسنجی افراد متعلق به سه طبقه اقتصادی اصلی ضعیف، متوسط و مرفه حضور داشته‌اند و تعداد افراد هر طبقه که به یکی از دو حزب رای داده‌اند در جدول زیر مشخص شده است.

تعداد نفر	طبقه اقتصادی	حزب مورد علاقه
20,000	ضعیف	دموکرات
35,000	متوسط	دموکرات
45,000	مرفه	دموکرات
50,000	ضعیف	جمهوری‌خواه
20,000	متوسط	جمهوری‌خواه
30,000	مرفه	جمهوری‌خواه

الف) مقدار Entropy هر یک از متغیرهای طبقه اقتصادی و حزب مورد علاقه را بدست آورید.

ب) مقدار Mutual Information دو متغیر طبقه اقتصادی و حزب مورد علاقه را بدست آورید.

ج) آیا این دو متغیر از هم مستقل‌اند؟ تحلیل خود را از جواب بدست آمده ارائه دهید.

سوال ۷- الف) انواع مختلف پیش‌پردازش داده و موارد استفاده آن‌ها را به اختصار توضیح دهید.
ب) با دو روش بیشینه‌کمینه[4] (min=0, max=1) و z-score داده‌های زیر را نرمال کنید.

200, 300, 400, 600, 1000

[4] Data Mining Concepts and Techniques, Jiawei Han

پیاده سازی:

هدف از این بخش از تمرین آشنایی با کتابخانه های مورد استفاده در پایتون برای داده کاوی می‌باشد. در قسمت اول این تمرین عملیاتی جهت پیش پردازش داده ها و visualization انجام میشود.
پیشنهاد میشود از Jupytr Notebook برای پیاده سازی کد های پایتون خود استفاده کنید.
در قسمت دوم با رگرسیون خطی کار خواهید کرد.
کتابخانه های مورد استفاده:

1. Numpy
2. Pandas
3. Matplotlib
4. Scikit-learn

قسمت اول:

فایل csv موجود در پوشه data با نام covid.csv شامل اطلاعات افراد مبتلا به COVID-19 در کره جنوبی میباشد.

۱- این فایل را خوانده و در یک جدول نمایش دهید.

۲- داده ها را با مشاهده سطر و ستون های آن شرح دهید. تعداد داده ها و نام ستون ها را نمایش دهید.

۳- مقادیر max, mean و std را در ستون birth_year به دست آورده و نمایش دهید.

۴- بررسی کنید که مقدار null در داده ها وجود دارد یا خیر. در صورت وجود با استفاده از متد مناسب آن را از بین ببرید.

۵-در این بخش مصور سازی داده ها را انجام می‌دهید. با انتخاب ستون مناسب از داده ها، scatter plot, matrix plot و histogram plot را نمایش دهید.

۶- بررسی کنید که آیا این مجموعه داده دارای outlier هست یا خیر. در صورت وجود علت خود را بیان کنید و برای روشی برای حل آن ارائه دهید.

قسمت دوم (رگرسیون خطی):

برای این بخش یک مجموعه داده از تعدادی دانش آموز در پوشه data با نام student.csv قرار دارد. هدف از این قسمت پیش بینی نمره نهایی دانش آموز (G3 attribute) با استفاده از رگرسیون خطی میباشد. اطلاعات مربوط به این مجموعه داده را میتوانید در این لینک مشاهده کنید.

داده ها باید به دو بخش train و test تقسیم کنید (نسبت تقسیم ۸۰ به ۲۰ باشد و می‌توانید از متد های آماده استفاده کنید) و روی داده های train رگرسیون خطی انجام دهید. برای سادگی این قسمت فقط ستون هایی که مقادیر عددی دارند را استفاده کنید (در حالت کلی میتوان ستون هایی که مرتبط هستند و مقدار عددی ندارند را به عدد تبدیل کرد).
سپس نمره نهایی را (G3) برای داده های test پیش بینی کنید و دقت (accuracy) مدل آموزش داده شده را به دست آورید.