<<به نام خدا>>

تمرین اول درس داده‌کاوی

هدف: آشنایی با مفاهیم پایه تحلیل داده و رگرسیون

تذکر۱ : ملاک اصلی انجام تمارین بخش پیاده سازی، گزارش است و ارسال کد بدون گزارش فاقد ارزش است. لذا برای این بخش یک فایل گزارش مستقل در قالب pdf تهیه کنید و در آن برای هر سوال، تصاویر ورودی، تصاویر خروجی و توضیحات مربوط به آن را ذکر کنید. سعی کنید

توضیحات کامل و جامعی تهیه کنید. همچنین قابل توجه است که زبان مورد قبول برای بخش پیاده‌سازی، تنها **پایتون** می‌باشد.

تذکر۲ : مطابق قوانین دانشگاه هر نوع کپی برداری و اشتراک کار دانشجویان غیر مجاز بوده و شدیدا برخورد خواهد شد. استفاده از

کدها و توضیحات اینترنت به منظور یادگیری **الزاما با ذکر منبع** بلامانع است.

راهنمایی: در صورت نیاز میتوانید سواالت خود را در خصوص پروژه از تدریسیار درس، از طریق ایمیل زیر بپرسید .

E-mail: zahra.dehghanian97@gmail.com

ارسال: فایل های کد و گزارش خود را در قالب یک فایل فشرده با فرمت StudentID_DM01.zip تا تاریخ ۱۴/۰۸/۱۴۰۰ ارسال نمایید.

* شایان ذکراست هر روز تاخیر باعث کسر **۲۰٪ نمره** خواهد شد.

بخش نوشتاری

۱) در آزمایشگاه ژنتیک دکتر فراهانی، بر روی شباهت دو ژن G1 و G2 تحقیق می شود. مقدار فعالیت این دو ژن در ۱۰ بازه زمانی اسکن شده و در جدول زیر آمده است.

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
g1	-5	-4	-3	-2	-1	1	2	3	4	5
g2	25	16	9	4	1	1	4	9	16	25

a. با استفاده از معیار شباهت correlation شباهت این دو ژن را بررسی کنید.

b. با استفاده از معیار شباهت Mutual Information شباهت این دو ژن را مقایسه کنید

c. آیا دو نتیجه بدست آمده متفاوت از یکدیگرند؟ علت را توضیح دهید.

۲) برای بردارهای داده شده موارد خواسته شده را بدست بیاورید.

a. x = [1,1,1,1] , y=[2,2,2,2]
cosine, correlation, euclidean

b. x = [0,1,0,1] , y=[1,0,1,0]
cosine, correlation, euclidean, jaccard

c. x = [1,1,0,1,0,1] , y=[1,1,1,0,0,1]
correlation, manhattan, bhattacharya

۳) در یک فرم نظرسنجی ، ۱۰۰ سوال چهارگزینه ای وجود دارد. این فرم در بین ۲۵۰ نفر داوطلب توزیع و پاسخ داده می شود. حین جمع بندی این فرم، تحلیلگر تصمیم به استفاده از قوانین انجمنی برای بررسی بیشتر می‌گیرد. برای این منظور نیاز به ویژگی های باینری می باشد. چگونه می‌توان این داده ها را به فرمت مناسب تبدیل کرد؟ پس از تبدیل به چه تعداد ویژگی می رسید؟

۴) مفاهیم زیر را تعریف کنید.

a. Dimension
b. Outlier
c. Independent variable
d. نمونه گیری طبقه ای (Sampling Stratified)
e. پیش‌پردازش

۵) کاهش بعد، یکی از تکنیکهای رایج در داده کاوی است و روشهای گوناگونی برای آن وجود دارد. یک از آنها را انتخاب کرده و با ذکر مثال آن را توضیح دهید. در ادامه تفاوت selection feature و extraction feature را بیان کنید.

۶) فرض کنید همبستگی بین دو متغیر، صفر است. مفهوم آن چیست؟ با توجه به تعریف متغیرهای مستقل، آیا این متغیرها، مستقل هستند؟

۷)) در مسائل یادگیری ماشین اعم از نظارت شده و نظارت نشده، تعیین معیار فاصله اهمیت زیادی دارد. در ادامه مهمترین روش های سنجش فاصله آمده است. خانه های خالی جدول را پر کنید.

نام روش	فرمول اصلی	مناسب برای بعد بالا	حساس به مقادیر بردار	بازه خروجی
Euclidean Distance				
Cosine Similarity				
Hamming Distance				
Manhattan Distance				

۸) سوال ۳) اگر داده های زیر را به روش box plot نمایش دهیم، min و max چه اعدادی خواهند بود؟

۷۴, ۴۴, ۲۰, ۱۰۰, ۱۴۶, ۱۴۳, ۹۰, ۸۰, ۸۹, ۱۴۴, ۷۴, ۷۳, ۷۱, ۵۶, ۷۰

۹) جستجو کنید.

a. یکی از مشکلات عمده در روش های Regression، مشکل multicollinearity است. درمورد آن جستجو کرده و توضیح دهید. همچنین دو روش Ridge Regression و Lasso Regression که برای جلوگیری از این معضل استفاده میشوند، این دو روش را نیز مورد بررسی قرار دهید.

b. جهت یافتن داده های پرت، روش های زیادی وجود دارد. روش های z-Score، IQR را توضیح دهید.

c. یکی از چالش های اجتناب ناپذیر مسائل یادگیری ماشین، وجود missing value ها است. اگر داده های ما کم باشند و نخواهیم که داده های miss شده را ignore کنیم، برای حل این مشکل حداقل 3 روش جستجو کنید و مزایا و معایب هر یک را با هم مقایسه نمایید.

بخش پیاده سازی:

۱) در این سوال قرار است انواع رگرسیون را پیدا سازی کنید، داده های این سوال در فایل 'data.mat' و 'data.npz' قرار دارد. این داده ها بر اساس این رابطه تولید شده اند:

$$y = 4{x_2}^2 x_1 + 2{x_2}^2 + 3x_1 + 1$$

در فایل ارائه شده، y_{test} خروجی متناظر با x_{2test} , x_{1test} است و در نهایت باید خروجی کدخود را با این آرایه مقایسه کنید. تابع هزینه تابع مجموع مربع خطا یا SSE می باشد.

a. رگرسیون خطی با استفاده از Gradient Descent پیاده سازی کنید
b. رگرسیون خطی با استفاده از Stochastic Gradient Descent پیاده سازی کنید
c. نتایج دو بخش قبل را با هم مقایسه کنید.
d. برای داده های تست، خروجی کد خودتان و مقدار صحیح خروجی y_{test} به صورت نمودار سه بعدی نمایش دهید.
e. مقدار تابع خطا رو داده های تست و آموزش را گزارش کنید.

۲) در این سوال هدف آشنایی با تکنیک های پیش پردازش داده و استفاده از داده است. مجوعه داده csv.player ،مجوعه اطلاعات بازیکنان بازی fifa21 میباشد. موارد زیر را برروی این مجموعه داده پیاده سازی کنید و نتایج را در گزارش ذکر کنید.

a. مجموعه داده csv.player را خوانده و داده های ابتدایی و انتهایی را نمایش دهید.
b. طبق تعریف، missing value‌ها را پیدا کنید.
c. میانگین، حداقل و حداکثر وزن بازیکنان را بدیست آورید.
d. کدام کشور دارای کمترین و کدام کشور دارای بیشترین بازیکن است؟ تعداد بازیکنان این دو کشور را گزارش کنید.
e. بازیکنانی که Growth < ۳ و Potential< ۸۴ بدست آورده و گزارش کنید.
f. نموداربازیکنان بدست آمده قسمت ۵ را بر اساس موقعیت بازیشان گزارش کنید.
g. کدام باشگاه دارای بیشترین تعداد بازیکن آیندهدار است. تعداد بازیکنان آن را گزارش کنید.
h. تعداد بازیکنانی که در سال ۲۰۲۱ قراردادشان با باشگاه تمام میشود و در تیم ملی کشورشان حضور ندارند.
i. ارزش بازیکنان زیر ۲۴ سال باشگاه چلسی را گزارش کنید.
j. موقعیت، درآمد و باشگاه فعلی Hazard. E را گزارش کنید.