دانشگاه صنعتی شریف
دانشکده‌ی مهندسی صنایع

مبانی داده‌کاوی کاربردهای آن
بهار ۱۴۰۴

استاد: مانا مس‌کار
مسئول تمرین: محمدسبحان کسائی و محمدمهدی منتظری هدش

تمرین شمارهٔ پنج **مهلت تحویل: ۱ تیر ۱۴۰۴**

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است. پس از آن، می‌توانید از شناوری مجاز خود استفاده کنید.
- توضیحات و تحلیل‌های شما در فایل PDF حتماً به زبان **فارسی** باشد و در غیر این صورت، نمرهٔ کل قسمت مربوطه از شما کسر خواهد شد. همچنین رعایت اصول نگارشی، قسمتی از بارم‌بندی را تشکیل می‌دهد.
- تمرین را در قالب یک فایل **ZIP** با نام `DataMining_Assignment5_GroupX.zip` ارسال کنید. این فایل باید شامل **دو Notebook** که هرکدام مربوط به هر بخش تمرین است و یک فایل **PDF** باشد. در فایل نوت‌بوک‌ها، کدهای اجرایی همراه با خروجی‌ها قرار داده شود و توضیحات مختصر در کنار کدها نوشته شود. خروجی‌ها را ذخیره کنید تا نیازی به اجرای مجدد نباشد. در فایل PDF، تحلیل و تفسیر نتایج و توضیحات خواسته‌شده آورده شود. همچنین به عنوان جایگزین فایل PDF، می‌توانید تحلیل‌های خود را تنها در فایل کد مورد نظر انجام داده و از آپلود PDF خودداری بفرمایید (همچنان بارم‌بندی تحلیل فارسی، رعایت اصول نگارشی و تمیز بودن نوشتهٔ تحلیل‌ها، برقرار است). بیشترین نمره به بخش تحلیل‌های شما اختصاص دارد؛ زیرا تحلیل داده‌ها مهم‌تر از اجرای کد است. استفاده از **هوش مصنوعی** برای **کدنویسی** مجاز است، اما تحلیل‌ها و توضیحات باید کاملاً توسط دانشجویان مربوطه صورت گیرد.
- حدالامکان از آوردن کد در فایل PDF خودداری بفرمایید و بخش‌های تمرین قابل تمیز باشد.
- تنها برای بخش یک و سوال سه بخشِ دو استفاده از پایتون مجاز می‌باشد.
- سوالات خود را از طریق آیدی تلگرامی @Realsobhanka مطرح بفرمایید.

بخش ۱: Expected Value

دانلود دادهٔ این بخش

۱. (۵۰ نمره) تصور کنید شما در شرکتی که به مشتریان خود، نرم‌افزار افزایش بهره‌وری می‌فروشد، به عنوان یک دانشمند داده فعالیت دارید. به منظور بهبود فروش، شرکت می‌تواند به مشتریان بالقوهٔ خود اشتراک‌های متفاوتی به مدت یک سال ارائه دهد.

(آ) **اشتراک اول:** این اشتراک Full Demo نام دارد. این اشتراک، تمامی قابلیت‌های نرم‌افزار را به طور کامل در اختیار مشتریان قرار می‌دهد اما هزینهٔ آماده‌سازی بالاتری دارد و هزینهٔ آن به ازای هر اشتراک، ۵۰ دلار می‌باشد. همچنین این اشتراک احتمال خرید مشتریان را معمولاً بیشتر می‌کند. شرکت می‌تواند برای این دسته از مشتریان، پشتیبانی رایگان نیز فراهم کند. هزینهٔ پشتیبانی، تابعی از تعداد کارمندان شرکت می‌باشد که در جدول ۱ می‌توان رابطهٔ آن‌ها را مشاهده نمود.

جدول ۱: هزینهٔ پشتیبانی بر تعداد کارمندان

تعداد کارمندان	هزینهٔ پشتیبانی (دلار)
۱ – ۲۰	$۱۵۰
۲۱ – ۵۰	$۲۵۰
۵۱ – ۲۰۰	$۴۰۰
+۲۰۱	$۴۵۰

(ب) **اشتراک دوم:** این اشتراک Lite Demo نام دارد. این اشتراک، نسخه‌ای محدودتر نسبت به اشتراک قبلی را در اختیار مشتریان قرار می‌دهد و هزینهٔ آماده‌سازی پایین‌تری (به ازای هر اشتراک، ۱۰ دلار) دارد. این اشتراک متقابلاً به طور معمول منجر به احتمال خرید کمتری نسبت به اشتراک قبلی می‌شود.

همچنین در نظر داشته باشید که در صورت خرید، یک هزینهٔ ثابت ۲۰۰ دلاری به جهت نصب وجود دارد.

هدف شما این است که با توجه به داده‌های تاریخی جمع‌آوری شده از شرکت، یک مدل پیش‌بینی بسازید. این مدل باید بتواند احتمال خرید یک مشتری را بر اساس ویژگی‌های خود مشتری (مانند اندازهٔ شرکت، صنعت و سابقه خرید قبلی) و جزئیات پیشنهادی که به او می‌دهید (نوع اشتراک، ارائه پشتیبانی و قیمت) تخمین بزنند. پس از برآورد احتمال، باید به تصمیمات استراتژیک زیر برای هر مشتری پاسخ دهید:

(آ) آیا اصلاً اشتراکی پیشنهاد دهیم یا خیر؟

(ب) اگر بله، کدام اشتراک؟

(ج) آیا پشتیبانی رایگان را هم در پیشنهاد لحاظ کنیم؟ (تنها برای اشتراک نوع اول در دسترس می‌باشد)

(د) چه قیمتی برای نرم‌افزار تعیین کنیم؟ (کران بالای قیمت را ۱۰۰۰۰ دلار در نظر بگیرید)

بخش ۲: Clustering

دانلود دادهٔ این بخش

۱. (۲۰ نمره) تابع هزینه الگوریتم خوشه‌بندی k-means به شکل زیر تعریف می‌شود:

$$\mathrm{L} = \sum_{j=1}^{k} \sum_{x_i \in S_j} \|x_i - \mu_j\|^2 \qquad (1)$$

که در آن S_j مجموعه‌ای از نمونه‌ها است که به مرکز خوشه‌ی μ_j نزدیک‌تر از هر خوشه‌ی دیگر هستند. منظور از $x_1, x_2, \ldots, x_n$ نمونه‌ها و $\mu_1, \mu_2, \ldots, \mu_k$ مراکز خوشه‌ها هستند.

(آ) مرحله‌ای از الگوریتم را در نظر بگیرید که برچسب داده‌ها ثابت است و مراکز خوشه‌ها μ_j به‌روزرسانی می‌شوند. نشان دهید که برای کمینه کردن تابع هزینه در این مرحله، کافی است میانگین هر خوشه به‌عنوان مرکز آن خوشه قرار گیرد.

(ب) آیا الگوریتم k-means نسبت به مقداردهی اولیه‌ی مراکز خوشه‌ها حساس است؟ آیا این الگوریتم تضمین می‌کند که همگرا می‌شود؟ توضیح دهید.

(ج) در مرحله‌ای از الگوریتم k-means که در آن میانگین خوشه‌ها μ_i ثابت‌اند و برچسب‌های نقاط داده بروزرسانی می‌شوند، گاهی ممکن است یک نقطه X_j به چندین مرکز خوشه با فاصله‌ی مساوی نزدیک باشد. اگر X_j در همان خوشه‌ای که در تکرار قبلی بوده باقی بماند، توضیح دهید چرا بهتر است این گزینه انتخاب شود؟ اگر این اصل رعایت نشود، چه مشکلی ممکن است پیش بیاید؟

۲. (۱۰ نمره) با استفاده از ماتریس فاصلهٔ داده‌شده، خوشه‌بندی سلسله‌مراتبی را با قانون به‌روزرسانی single linkage انجام داده، نمودار dendrogram را رسم کنید و تعداد خوشهٔ مناسب را مشخص نمایید.

جدول ۲: ماتریس فاصله

	A	B	C	D	E
A	۰	۲	۶	۱۰	۹
B	۲	۰	۵	۹	۸
C	۶	۵	۰	۴	۵
D	۱۰	۹	۴	۰	۳
E	۹	۸	۵	۳	۰

۳. (۱۰ نمره) داده‌های مشتریان یک فروشگاه در اختیار شما قرار گرفته است. تنها با استفاده از دو ویژگی Annual Income (k$) و Spending Score (1-100):

(آ) داده‌ها را به صورت خوشه‌بندی سلسله‌مراتبی، خوشه‌بندی کنید.

(ب) با استفاده از نمودار elbow، تعداد خوشه‌های مناسب را تعیین کنید. تفاوت‌های استفاده از این نمودار هنگام خوشه‌بندی به روش معمول (مانند k-means) و خوشه‌بندی سلسله‌مراتبی چیست؟

(ج) داده‌ها را با توجه به شمارهٔ خوشه رنگ‌بندی کرده و در یک نمودار با استفاده از دو ویژگی ذکر شده نشان دهید. خوشه‌بندی شما به صورت بصری نیز باید قابل قبول باشد.