دانشگاه صنعتی شریف
دانشکدهٔ مهندسی صنایع

مبانی داده‌کاوی و کاربردهای آن
بهار ۱۴۰۴

استاد: مانا مس‌کار
مسئول تمرین: محمدسبحان کسائی

تمرین شمارهٔ دو **مهلت تحویل: ۱۳ فروردین‌ماه ۱۴۰۴**

- مهلت ارسال پاسخ تا ساعت ۵۹ : ۲۳ 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_Assignment2_GroupX.zip` ارسال کنید. این فایل باید شامل یک **Notebook**، یک **PDF** و فایل دادهٔ تمیزشدهٔ بخش ۳ باشد. در فایل نوت‌بوک، 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.
- حدالمکان از آوردن کد در فایل PDF، خودداری بفرمایید.
- تنها برای سوالِ ۲ بخشِ ۲ و بخش ۳ نیاز به کدنویسی دارید.
- سوالات خود را از طریق آیدی تلگرامی @RealSobhanKa یا آدرس ایمیل Sobhan.kasaei@sharif.edu مطرح بفرمایید.

بخش ۱: پرکردن داده‌های گم‌شده

یکی از مراحل ضروری پیش‌پردازش داده‌ها قبل از استفاده به عنوان ورودی مدل‌های یادگیری ماشین، **پر کردن داده‌های گم‌شده**[۱] است. در این سوال، به بررسی این مهم پرداخته می‌شود.

۱. (۱۰ نمره) در پرکردن داده‌های گم‌شده، مهم است که سعی شود توزیع داده‌ها تغییر داده نشود تا مدل در هنگام یادگیری، توزیع اصلی داده‌ها را فرا گیرد. فرض کنید داده‌های مربوطه، دارای چولگی به راست باشند، کدامین آماره‌ها (میانگین، میانه، مد) به جهت پرکردن داده‌های گم‌شده مناسب‌تر می‌باشند؟ چرا؟ اگر چولگی به چپ باشد چطور؟

۲. (۱۵ نمره) فرض کنید $m+n$ ستون به عنوان داده ستون‌های ویژگی در اختیار شما قرار گرفته‌است و همگی نسبت به یکدیگر مستقل می‌باشند. n ستون آن حاوی داده‌های گم‌شده می‌باشد و می‌خواهید آن‌ها را با مدل‌های یادگیری پر کنید. در ابتدا با استفاده از m ستون دیگر، یکی از ستون‌های حاوی مقادیر گم‌شده را پر می‌کنید و پس از آن، این ستون پرشده را به ستون‌های ویژگی خود برای پر کردن ستون بعدی اضافه می‌کنید (برای مثال برای پر کردن ستون بعدی، از $n+1$ ستون استفاده خواهید کرد). این کار چه آسیبی به دقت پرکردن داده‌ها یا حتی پیش‌بینی نهایی وارد می‌کند؟ راهنمایی: دقت هر مدل i را $1-\alpha_i$ در نظر بگیرید.

۳. (۷ نمره) در هنگام پر کردن داده‌های گم‌شده، باید به جلوگیری از نشت داده[۲] توجه داشت. این مورد را به طور کامل

[۱] Values Missing
[۲] Leakage Data

شرح داده و در هنگام پرکردن داده‌ها ذکر کنید چگونه باید از این عامل جلوگیری کرد؟

بخش ۲: مدیریت داده‌های پرت

دانلود دادهٔ این بخش

همانند، پرکردن داده‌های گم‌شده، اگر مدیریت درستی بر روی داده‌های پرت صورت نگیرد، موجب می‌شود دقت مدل کاهش یافته و مدل نهایی مقاوم به این نوع داده‌ها نباشد.

۱. (۲۵ نمره) تابع هدفی برای مسئلهٔ رگرسیون خطی بر روی داده‌های پیوسته به صورت زیر تعریف می‌کنیم و قصد پیش‌بینی y_i را داریم:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad (1)$$

در این عبارت N تعداد داده‌ها، y_i داده‌های در دست و $\hat{y}_i$ پیش‌بینی مدل می‌باشد.

تابع هدف را به گونه‌ای تغییر دهید تا نسبت به داده‌های پرت مقاوم عمل کند؛ عبارتی خارج از پرانتز کم یا زیاد نکنید و تنها شکل تابع فعلی را تغییر دهید.

تابعی نیز با بدون درنظر گرفتن محدودیت بالا نیز پیشنهاد دهید و علت مقاوم شدن هر دو تابع پیشنهادی را به طور کامل شرح دهید.

اگر تابع فعلی نسبت به داده‌های پرت مقاوم است، چرا همیشه از این تابع‌ها استفاده نمی‌گردد و تابع پراستفادهٔ مسئلهٔ رگرسیون پیوسته، تابع ذکرشده می‌باشد؟ مزایا و معایب آن‌ها را مقایسه کنید.

۲. (۱۵ نمره) اگر تعداد داده‌های موجود به دلایلی کم باشند (مانند هزینهٔ بالایی نمونه‌گیری بیشتر)، باید سعی کرد تا از حذف داده‌ها پرهیز نمود و از داده‌های در دست بیشترین استفاده را داشت. در نگاه اول ممکن است بعضاً داده‌هایی پرت در نظر گرفته باشند و این خطا به علت عدم شناسایی درست توزیع داده‌ها باشد. با استفاده دو تکنیک پیاله‌کردن[۳] و تبدیل[۴] داده‌ها، مدل رگرسیون خطی بر روی داده‌ها برازش داده به‌طوری که معیار R^2 (معیاری برای بیان خوبی برازش بر روی داده‌ها می‌باشد که در این قسمت، کاری با مفهوم آن نداریم) برای مدل هر پیاله، بیشتر از ۶۰ درصد باشد. دقت کنید داده‌ای **نباید** حذف گردد.

بخش ۳: دیابت و جمع‌بندی!

دانلود داده‌های این بخش

۱. (۱۸ نمره) داده‌های دردست، مربوط به میزان قند خون و وضعیت دیابت افراد می‌باشد. هدف پیدا کردن مقادیر مرزی برای تبدیل داده‌های میزان قند خون به سه دستهٔ احتمال دیابت «کم»، «متوسط» و «زیاد» می‌باشد. با استفاده از الگوریتم «Chi Merge»، مقادیر مرزی خواسته‌شده را بدست آورده و داده‌ها را برچسب بزنید. مقادیر موجود در هر بازه را تحلیل کنید؛ آیا نتایج منطقی‌اند؟

۲. (۴۰ نمره) با توجه به آنچه تاکنون آموخته‌اید، دادهٔ مربوطه را تمیز کنید. این تمیزکاری‌ها شامل:

- اصلاح نام ستون‌ها
- اصلاح نوع داده‌ها[۵]
- یکدست کردن مقادیر ستون‌ها
- مدیریت داده‌های پرت، گم‌شده و ...

می‌باشند. شایان ذکر است اگر موارد دیگری وجود دارد، باید مدیریت آن‌ها نیز صورت گیرد و تنها مثال‌هایی برای راهنمایی آورده شد.

---

[۳] Binning
[۴] Transformation
[۵] Data Type