دانشگاه صنعتی اصفهان
دانشکده مهندسی برق و کامپیوتر

# مبانی داده کاوی

## تمرین سری ۴

بهار ۱۴۰۳

# فهرست مطالب

# ۱ سوالات

## ۱.۱ سوال ۱

در مورد ارتباط هر کدام از موارد زیر با رخداد بیش‌برازش توضیح دهید.

۱. حجم مجموعه داده

۲. تعادل داده‌ها (داده‌های بالانس و غیربالانس)

۳. ویژگی‌های نامربوط

۴. تعداد ایپاک‌های آموزش

۵. پیچیدگی مدل

۶. نشت داده‌ها (Data Leakage)

## ۲.۱ سوال ۲

جدول زیر مجموعه کوچکی از نتایج اعتبارسنجی یک مدل رده‌بندی را همراه با مقادیر واقعی آن، نشان می‌دهد. نرخ‌های خطا (error rate) را با استفاده از مقادیر برش (آستانه یا cutoff) ۰٫۲۵، ۰٫۵ و ۰٫۷۵ محاسبه کنید.

| Propensity of 1 | Actual |
|---|---|
| 0.03 | 0 |
| 0.52 | 0 |
| 0.38 | 0 |
| 0.82 | 1 |
| 0.33 | 0 |
| 0.42 | 0 |
| 0.55 | 1 |
| 0.59 | 0 |
| 0.09 | 0 |
| 0.21 | 0 |
| 0.43 | 0 |
| 0.04 | 0 |
| 0.08 | 0 |
| 0.13 | 0 |
| 0.01 | 0 |
| 0.79 | 1 |
| 0.42 | 0 |
| 0.29 | 0 |
| 0.08 | 0 |
| 0.02 | 0 |

شکل ۱: مقادیر واقعی عضویت رده و تمایل (تخمین)ها به ردهٔ ۱ درمجموعهٔ داده‌های اعتبارسنجی

## ۳.۱ سوال ۳

به سوال ۵ در تمرین سری ۳ مراجعه کنید.
در ادامهٔ Notebook خود:

۱. یکی از روش‌های cross-validation را برای بهبود دقت مدل خود به کار بگیرید. [۱] و روش خود را توضیح دهید. مقاله بارگذاری شده در کنار تمرین می تواند به شما در انتخاب یکی از روش‌ها کمک کند.

سپس برای پیاده سازی آن، لازم است پس از فراخوانی کتابخانه مربوطه، از تابع cross_val_score استفاده کنید و نتیجه را چاپ کنید. ورودی‌های این تابع نام مدل شما و مقادیر X_train, y_train یا ... می تواند باشد. توجه کنید که می توان کل مقادیر x و y (که در بخش سوم سوال ۵ تمرین قبلی با نام X و y مشخص شد) را به این تابع وارد کرد. انجام هرکدام در نهایت به روش ارزیابی مدل‌تان بستگی دارد.

## ۴.۱ سوال ۴

نتیجه اجرای یک الگوریتم داده‌کاوی بر روی یک مجموعهٔ داده‌های تراکنشی بدین شرح است: ۸۸ رکورد به عنوان کلاهبردار رده‌بندی شده‌اند (که ۳۰ رکورد آن به درستی رده‌بندی شده است) و ۹۵۲ رکورد نیز به عنوان غیرکلاهبردار رده‌بندی شده اند (که ۹۲۰ رکورد آن به درستی رده‌بندی شده است).

- نرخ خطای طبقه‌بندی نادرست(misclassification rate) را پیدا کنید.

## ۵.۱ سوال ۵

درستی یا نادرستی جملات زیر را تعیین کنید و برای پاسخ خود دلیل بیاورید.

۱. برای داده‌های با نویز بالا، پس–هرس (post-pruning) بهتر از پیش–هرس (pre-pruning) می‌باشد.

۲. برای داده‌های غیر بالانس، پیش–هرس (pre-pruning)، بهتر از پس–هرس (post-pruning) می‌باشد.

۳. وقتی مجموعهٔداده‌ها کوچک است، احتمال بیش‌برازش بالا می‌رود و وقتی مجموعهٔ داده بزرگ است؛ احتمال کم‌برازش بیشتر می‌شود.

۴. درخت‌های تصمیم برای مجموعه داده‌هایی که فیچرها، روابط غیرخطی با متغیر هدف دارند، مناسب هستند.

## ۶.۱ سوال ۶

برای ارزیابی مدلی که باید روی دیتاست تشخیص کلاهبرداری کارت اعتباری (Credit Card Fraud Detection) [۲] آموزش ببیند، کدام روش ارزیابی را پیشنهاد می دهید؟ چرا؟

[1] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html
[2] https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

## ۲ نکات پاسخ دهی

- تمرینات به صورت مرتب و خوانا بارگذاری شوند.
- برای تمرینات غیر عملی که به صورت تایپی ارسال شوند امتیاز تشویقی درنظر گرفته می‌شود.
- کدهای خود را حتما در فایل PDF نیز قرار دهید.
- در سوالات توضیحی، قدرت تحلیل افراد ملاک مقایسهٔ پاسخ‌ها خواهد بود.
- فایل پایتون و یا Notebook برای تمرینات ضمیمه شود و همه به صورت یک فایل zip بارگذاری شوند. فایل zip را با فرمت DM4022_HW4_[StudentNumber].zip نام‌گذاری کنید.
- درصورت وجود ابهام خاص می توانید موارد را با دستیار آموزشی مطرح کنید.