



تمرین پنجم یادگیری ماشین

توجه:

- کد باید فقط در زبان پایتون باشد.
- استفاده از کتابخانه‌های از پیش آماده مجاز است، مگر اینکه نقیض آن گفته شود. در گزارش کتابخانه‌ها ذکر شوند.
- لطفاً علاوه بر ضمیمه کردن کد و تصاویر، نتایج را تحلیل و در فایل گزارش خود ضمیمه کنید.
- فایل گزارش به فرمت pdf و به زبان فارسی باشد.

تمرین را در سایت کوئرا بارگذاری کنید.

سوال اول

مجموعه داده این سوال در پوشه 'Dataset' قرار دارد. دو مجموعه داده به نام های `train_set.csv` و `test_set.csv` در اختیار شما قرار گرفته شده است. متغیر وابسته 'income' نام دارد، سایر متغیرها مستقل اند. مجموعه داده `test_set.csv` فاقد متغیر وابسته است.

۱. نوع هر متغیر، اسمی یا عددی، را مشخص کنید.
 ۲. متغیرهای عددی را استاندارد سازی کنید.
- در سوالات بعدی با این مجموعه داده (استانداردسازی شده) کار خواهید کرد.

سوال دوم

۱. دسته‌بند Logistic Regression انتظار دارد که ورودی‌ها به صورت عددی باشد. چه راه حلی برای این مشکل پیشنهادی می‌کنید؟ توضیح دهید و آن را بر روی داده پیاده کنید.
۲. اکنون با استفاده از دسته‌بند Logistic Regression قصد دسته‌بندی را داریم. با استفاده از Repeated 5 Fold Cross-Validation با دو تکرار ترکیب‌های نرخ یادگیری ثابت، 10^{-5} ، 10^{-2} و 1، نرخ منظم سازی کاهش وزن 10^{-5} ، 10^{-2} و 0، و تعداد تکرار 500، 1000 و 2000 را بررسی کنید و دقت‌ها را با استفاده از boxplot رسم کنید (۹ ترکیب خواهید داشت).
- توجه داشته باشید که حتماً از تابع خطای `cross_entropy` و الگوریتم کاهش گرادینان استفاده شود، و همچنین پارامترهای دیگر دخیل نباشد. در کتابخانه‌ها معمولاً پارامترهای دیگر برای بهبود نتایج دخیل است، در نتیجه حتماً توضیحات کتابخانه را مطالعه کنید و آن‌ها را غیر فعال کنید.
۳. بر اساس پلات رسم شده، کدامیک از مدل‌ها به نظر شما بهترین است؟ آیا می‌شود تنظیمات آموزشی بهتری را ارائه داد؟ (منظور از تنظیمات آموزشی نرخ کاهش وزن، نرخ یادگیری، و تعداد تکرار است).
۴. با استفاده از مدل بهینه که انتخاب کردید، پیش‌بینی را بر روی داده `test_set.csv` انجام دهید و آن را در فایل به نام `prediction_lr.csv` ذخیره و ضمیمه کنید.

سوال سوم

۱. مدلی بر پایه بیزین ساده طراحی کنید، که متغیرهای اسمی را از طریق توزیع اسمی و متغیرهای عددی را از طریق توزیع گاوسی مدل سازی کند. سپس با استفاده از روش Repeated 5 Fold Cross-Validation با دو تکرار دقت ارزیابی را محاسبه کنید و Boxplot را در کنار Boxplot های سوال قبل رسم کنید.

۲. این مدل در مقایسه با Logistic Regression چگونه عمل کرده است.

۳. یشبینی را بر روی داده test_set.csv انجام دهید و آن را در فایلی به نام prediction_b.csv ذخیره و ضمیمه کنید.

سوال چهارم (امتیازی)

یکی از ضعف های مدل بیز ساده را توضیح و راه حلی را پیشنهاد دهید. لزومی ندارد که وارد جزئیات ریاضیاتی شوید، کافی است مفهومی توضیح دهید.