تمرین پنجم یادگیری ماشین

توجه:

۱. استفاده از کتابخانه‌های از پیش آماده مجاز است.

۲. لطفاً علاوه بر ضمیمه کردن کد، نتایج را تحلیل و در فایل گزارش خود ضمیمه کنید.

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 متغیر ها مستقل اند. مجموعه داده test_set.csv فاقد متغیر وابسته است.

سوال اول

۱. نوع هر متغیر، اسمی یا عددی، را مشخص کنید.

۲. متغیر های 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 ورودی ها به صورت عددی باشد. چه راه حلی برای این مشکل پیشنهادی می کنید؟ توضیح دهید و آن را بر روی داده پیاده کنید.

۲. اکنون با استفاده از دسته‌بند Logistic Regression قصد دسته بندی را داریم. با استفاده از Repeated 5 Fold Cross-Validation با دو تکرار ترکیب های 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 خطای ها را با استفاده از boxplot رسم کنید (۹ترکیب خواهید داشت).

توجه داشته باشید که حتما از خطای cross_entropy و الگوریتم کاهش گرادیان استفاده شود، و همچنین پارامتر های دیگر دخیل نباشد. در کتابخانه ها معمولا پارامتر های دیگر برای بهبود نتایج دخیل است، در نتیجه حتما توضیحات کتابخانه را مطالعه کنید و آن ها را غیر فعال کنید.

۳. بر اساس پلات رسم شده، کدامیک از مدل ها به نظر شما بهترین است؟ آیا می شود تنظیمات آموزشی بهتری را ارائه داد؟

۴. با استفاده از مدل بهینه که انتخاب کردید، پیشبینی را بر روی داده test_set.csv انجام دهید و آن را در فایلی به نام prediction_lr.csv ذخیره و ضمیمه کنید.

سوال سوم

۱. مدلی بر پایه بیزین ساده طراحی کنید، که متغیر های اسمی را از طریق توزیع اسمی و متغیر های عددی را از طریق توزیع گاووسی مدل سازی کند. سپس با استفاده از روش Repeated 5 Fold Cross-Validation با دو تکرار خطای 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. پیشبینی 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 جزییات ریاضیاتی شوید، کافی است مفهومی توزیع بدهید.