



تمرین ششم 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.

سوال اول

- در این تمرین خوشه بندی کنیم و در ابعاد پایین تر نمایش دهیم. مجموعه داده این تمرین در فایل 1.csv قرار دارد.
۱. پیش پردازش های لازم را بر روی داده انجام دهید و آن ها را توضیح دهید.
 ۲. خوشه بندی را برای تعداد خوشه از ۲ تا ۳۰ (۲ و ۳۰ خوشه لحاظ شود) انجام دهید، و یکی از آن ها را بر اساس تکنیکی که سر کلاس آموخته اید انتخاب کنید. پروسه را انتخاب را توضیح دهید.
 ۳. با استفاده از الگوریتم PCA داده ها را ۲ بعد برده و خوشه بندی را در آن فضا نمایش دهید. آیا الگویی مشاهده می کنید؟

سوال دوم

- در این سوال قصد داریم یک مسئله دسته بندی اعداد انگلیسی از ۰ تا ۹ را به صورت شبه با نظارت^۱ حل کنیم. تصاویر 8×8 اند، و ۶۴ ستون اول فایل ها مربوط به پیکسل ها اند و اگر برچسب ها در فایل داده موجود باشد، ستون آخر با برچسب target است. داده های این بخش با پیشوند "2_" مشخص شده اند.
۱. در صورتی که پیش پردازش خاصی می تواند دقت را به صورت کلی افزایش دهد، بر روی مجموعه دادگان پیاده کنید.
 ۲. مدل Logistic Regression را بر روی دادگان آموزشی، 2_train_set.csv، آموزشی دهید و دقت را بر روی دادگان تست، 2_test_set.csv، گزارش کنید.
 ۳. الگوریتم خوشه بندی با تعداد خوشه ۴۰ بر روی دادگان بدون برچسب 2_unlabeled.csv اجرا کنید.
 ۴. با استفاده از PCA خوشه بندی را در فضای دو بعدی نمایش دهید. آیا الگویی مشاهده می شود؟
 ۵. در هر خوشه نزدیک ترین نمونه به مرکز خوشه را بدست آورید و رسم کنید. اکنون به تمام نمونه های خوشه مذکور برچسب نمونه انتخاب شده را بدهید. با این کار شما به صورت شبه ناظر مجموعه داده را برچسب گذاری کرده اید.
 ۶. اکنون مدل را با دادگان آموزش قبلی و دادگان ساخته شده آموزش دهید و خطا را بر روی دادگان تست گزارش دهید. آیا تغییری وجود دارد؟ دلیل چه است؟

¹Semi-Supervised