بسمه تعالی

دانشگاه صنعتی شریف

دانشکده مهندسی برق

مقدمه‌ای بر یادگیری ماشین - گروه ۱

نیم‌سال دوم ۱۴۰۳-۱۴۰۲

مسابقه / پروژه

مهلت ارسال: ۶ تیر

۱ کلیات

هدف این پروژه و مسابقه آشنایی شما با روند حل یک مسأله یادگیری ماشین و کار با داده واقعی است. دقت‌های مسابقه با توجه به نتیجه بقیه افراد سنجیده می‌شوند. همچنین توجه کنید در تمام قسمت‌ها، نمایش و رسم نمودارهای مناسب برای بیان بهتر نتایج اهمیت دارد. نیازی به گزارش مجزا برای پروژه نیست ولی لازم است توضیحات کافی برای هر قسمت ارائه شود. بهتر است توضیحات را به همراه کد در یک فایل Jupyter notebook قرار داده و آپلود کنید.

۲ بررسی داده‌ها (۴۰ نمره)

۱.۲ خواندن داده‌ها (۵ نمره)

داده‌ها را از فایل train_data.csv بخوانید. این داده‌ها مربوط به اعتبارسنجی افراد به منظور دریافت سهام از یک کمپانی بزرگ می‌باشند. دیتاست شامل ۲۰۰۰ سطر است و هر سطر اطلاعات مربوط به هر فرد را بیان می‌کند. به طور مشخص اطلاعات زیر را از هر فرد در اختیار داریم:

- Id: آیدی فرد مد نظر
- Age: سن فرد مد نظر
- MF: جنسیت. M مرد و F زن است.

- LoE: سطح آموزش. Dip دیپلم، Ad. Dip فوق دیپلم، Bach لیسانس، Mst فوق لیسانس، Doct دکتری و P. Doct پسادکتری می‌باشد.
- YoW: سابقه کار بر حسب سال
- YoCW: سابقه کار بر حسب سال در شغل فعلی
- Income: درآمد ماهیانه
- Housing: وضعیت خانه. O صاحب خانه، R اجاره نشین و N بدون خانه.
- Car: داشتن یا نداشتن خودرو
- Res: پذیرفته شدن یا نشدن فرد در اعتبارسنجی (برچسب خروجی).

۲.۲ پیش‌پردازش داده‌ها (۱۵ نمره)

در این مرحله باید پیش‌پردازش‌های مورد نیاز روی داده‌ها انجام شود. این پیش‌پردازش شامل موارد زیر می‌باشد:

- داده‌های پرت حذف شوند. این داده‌ها ممکن است شامل مقادیر عددی پرت و یا NaN در برخی از ستون‌ها باشند.
- داده‌هایی که از جنس string هستند به صورت مناسبی به داده‌های عددی تبدیل شوند.
- نرمالیزاسیون مناسب روی داده‌ها صورت گیرد تا بازهٔ عددی آن‌ها مشابه باشد.
- برچسب نهایی داده‌ها به صورت ۰ یا ۱ باشد.

۳.۲ نمایش داده‌ها (۱۰ نمره)

حال برای ستون‌های مختلف به نحوهٔ مناسبی هیستوگرام داده‌ها را رسم کنید و آن‌ها را تحلیل نمایید. از هیستوگرام متناظر با ستون برچسب در مورد معیار مناسب برای ارزیابی مدل نتیجه‌گیری کنید. همچنین دو ویژگی تصادفی از بین داده‌ها انتخاب کنید و آن‌ها را با دو رنگ متناظر با دو کلاس رسم کنید و مستقل بودن این جفت ویژگی‌های انتخاب شده را با محاسبهٔ ضریب همبستگی آن‌ها بررسی نمایید. این مرحله را ۵ بار تکرار کنید.

۴.۲ تحلیل داده‌ها (۱۰ نمره)

صحت یا عدم صحت گزاره‌های زیر را در این دیتاست با روش آماری t-test تعیین کنید.

- افرادی که سطح آموزشی بالاتری دارند از اعتبار بهتری برخوردار هستند.
- عموماً افراد با درآمد بالاتر یا صاحب‌خانه هستند یا ماشین دارند.
- افرادی که در ۵ سال اخیر شغل خود را تغییر نداده‌اند اعتبار بالاتری دارند.
- مردان با سن بالای ۵۰ سال نسبت به مردان زیر ۳۰ سال نرخ پذیرش کمتری در اعتبارسنجی دارند.
- مردان درآمد بیشتری از زنان دارند.

۳ آموزش و تست مدل (۴۰ نمره)

در انجام مراحل زیر و گزارش دقت‌های مورد نیاز لازم است ۲۰ درصد داده‌ها را به صورت تصادفی به عنوان داده تست جدا کرده و بقیه داده‌ها را به عنوان داده آموزش در نظر بگیرید. برای بررسی دقیق‌تر باید این کار ۵ بار (با random_state های مختلف) تکرار شود و میانگین و واریانس امتیازهای خواسته شده گزارش گردد.

۱.۳ انتخاب مدل و هایپرپارامترها (۱۰ نمره)

مدل مناسبی که به نظر شما می‌تواند دقت خوبی روی دیتاست داده شده داشته باشد انتخاب کنید و هایپرپارامترهای آن را با روش cross-validation به دست آورید. توجه کنید استفاده از همهٔ مدل‌های یادگیری ماشین مجاز است اما مجاز به استفاده از روش‌های یادگیری عمیق نیستید.

۲.۳ آموزش مدل و نتایج (۲۰ نمره)

مدل خود را آموزش دهید. مقادیر accuracy، precision، recall، F1 score و AUPRC را گزارش نمایید (گزارش میانگین و واریانس مقادیر فوق کافی است). همچنین Confusion Matrix را برای آخرین مدل آموزش داده شده رسم کنید. توضیح دهید چرا معیار دقت (accuracy) به تنهایی در این دیتاست نمی‌تواند معیار مناسبی باشد و مشخص کنید بین معیارهای گزارش شده کدام یک و به چه علت برای این دیتاست مناسب‌تر است؟ همچنین زمان آموزش مدل را به دست آورید. این زمان نباید از ۲۰۰ میلی‌ثانیه فراتر رود.

۳.۳ تست مدل (۱۰ نمره)

فایل test_data_1.csv را که در اختیار شماست بخوانید. این فایل شامل ۵۰۰ داده با فرمتی دقیقاً مشابه داده‌های آموزش ولی بدون برچسب است. برنامه‌ای بنویسید که برچسب پیش‌بینی شده شما (۰ به معنای رد شدن یا ۱ به معنای قبول شدن) را برای این داده‌ها به ترتیب چاپ کند (هر برچسب در یک خط جداگانه). در نهایت فایل text به دست آمده را ضمیمه کرده و ارسال کنید. بر اساس برچسب‌های این مجموعه داده، AUPRC مدل (که معیار ارزیابی پروژه و مسابقه است) محاسبه شده و نمره شما تعیین می‌شود.

۴ ساده‌سازی مدل (۲۰ نمره)

- بررسی کنید که کدام ویژگی‌ها همبستگی کمتری با خروجی دارند و سپس با حذف آن‌ها آموزش مدل را تکرار کنید. دقت کنید می‌خواهیم معیار AUPRC از ۸۰ درصد کمتر نشود. چه ستون‌هایی را می‌توان حذف کرد تا معیار مشخص شده همچنان برقرار باشد؟ (۱۰ نمره)
- در مرحله بعد با روش‌های کاهش بعد سعی کنید ویژگی‌های جدیدی با ابعاد پایین‌تر به دست آورید. تا جایی که معیار قسمت قبل برقرار باشد بعد داده‌ها را پایین بیاورید و مشخص کنید حداقل تعداد ویژگی‌ها برای برقراری دقت مطلوب چقدر است. (۱۰ نمره)

۵ مسابقه (۲۵ نمره امتیازی)

یک فایل پایتون به نام test.py ایجاد کنید که از یک فایل به نام test_data_2.csv (که در اختیار شما نیست) داده‌های تست را بخواند و معیار مسابقه یعنی AUPRC بهترین مدلتان را روی آن چاپ کند. دقت کنید فرمت این داده‌های تست دقیقاً مشابه داده‌های آموزش است که در اختیار شماست (یعنی شامل ستون برچسب و سایر ستون‌ها با همان فرمت است). لذا تمام پیش‌پردازش‌هایی که روی داده‌های آموزش انجام داده‌اید باید روی این داده‌ها نیز صورت گیرد. همچنین تمام کتابخانه‌های مورد نیاز import شده و کد بدون هیچ مشکلی با قرار گرفتن در کنار فایل داده تست اجرا شود. مصحح روی این مجموعه داده مدل شما را امتحان کرده و نتایج آن در مسابقه مورد استفاده قرار می‌گیرد.

۶ یادگیری با استفاده از Ensemble (۱۰ نمره امتیازی)

در این قسمت یادگیری باید از طریق مدل Ensemble و به دو روش مطابق توضیحات زیر انجام شود. انجام هر قسمت ۵ نمره امتیازی دارد.

۱.۶ با روش تقسیم داده

در این بخش باید یادگیری با استفاده از روش Ensemble و تعداد مختلفی از weak learner ها را انجام دهید. مراحل زیر را دنبال کنید:

- یک مدل پایه انتخاب کنید (مدل‌های یادگیری عمیق مجاز نیستند) و یادگیری با روش‌های مختلف Ensemble را با تعداد weak learner های مختلف (۵، ۱۰ و ۲۰ تا) انجام دهید.
- تصادفا ۸۰ درصد داده‌های داده شده را داده‌های آموزشی و بقیه را داده تست در نظر بگیرید.
- داده‌های آموزشی را به تعداد weak learner ها تقسیم کنید و هر weak learner را بر روی داده‌های خودش آموزش دهید.
- برای ترکیب نتایج weak learner ها از می‌توانید از روش max voting یا هر روشی به دلخواه خود استفاده کنید.

این فرآیند را ۵ بار تکرار کنید و میانگین و انحراف از معیار مقادیر ارزیابی گفته شده در قسمت‌های قبل را به ازای بهترین مدل خود(بهترین مدل پایه، بهترین تعداد weak learner ها، بهترین روش ترکیب نتایج weak learner ها) گزارش کنید.

۲.۶ با روش تقسیم ویژگی‌ها

در این بخش نیز همانند بخش قبل ولی با رویکردی متفاوت به یادگیری با استفاده از روش Ensemble و تعداد مختلفی از weak learner ها می‌پردازیم. مراحل زیر را دنبال کنید:

- یک مدل پایه انتخاب کنید (مدل‌های یادگیری عمیق مجاز نیستند) و یادگیری با روش‌های مختلف Ensemble را با تعداد weak learner های مختلف (۲، ۳، ۴ تا) انجام دهید.
- تصادفا ۸۰ درصد داده‌های داده شده را داده‌های آموزشی و بقیه را داده تست در نظر بگیرید.
- ویژگی‌های داده‌های آموزشی را بین weak learner ها تقسیم کنید، به طوری که هر weak learner مجموعه‌ای از ویژگی ها دریافت کند و همه داده‌ها برای آن ویژگی‌ها را داشته باشد. سپس هر weak learner را بر روی داده‌های خودش آموزش دهید.

- برای ترکیب نتایج weak learner ها از می‌توانید از روش max voting یا هر روشی به دلخواه خود استفاده کنید.

این فرآیند را ۵ بار تکرار کنید و میانگین و انحراف از معیار مقادیر ارزیابی گفته شده در قسمت‌های قبل را به ازای بهترین مدل خود(بهترین مدل پایه، بهترین تعداد weak learner ها، بهترین روش تقسیم ویژگی‌ها، بهترین روش ترکیب نتایج weak learner ها) گزارش کنید.