باسمه تعالی

دانشگاه صنعتی شریف

دانشکده مهندسی برق

مقدمه‌ای بر یادگیری ماشین - گروه ۱
نیم سال دوم ۱۴۰۲-۱۴۰۱

مسابقه/پروژه

مهلت ارسال: ۶ تیر

---

۱ کلیات

هدف این مسابقه/پروژه آشنایی شما با روند حل یک مسئله یادگیری ماشین و کار با داده واقعی است. دقت‌های مسابقه با توجه به نتیجه بقیه افراد سنجیده می‌شوند. همینطور توجه کنید که در تمام قسمت‌ها، نمایش و رسم نمودارهای مناسب برای بیان بهتر نتایج، اهمیت و نمره دارد. برای مسابقه با توجه به دقت مدل شما روی داده تست که در دسترس شما نیست، تا ۲۰ نمره امتیازی برای پروژه شما در نظر گرفته می‌شود. همینطور نیازی به تهیه گزارش جدا نیست ولی دقت کنید که باید در هر قسمت توضیحات کافی داده باشید. بهتر است تمام کد و توضیحات خود را در یک فایل Jupyter Notebook قرار دهید و آپلود کنید.

هم‌فکری و استفاده از منابع اینترنتی با ذکر منبع مجاز است (بدون ذکر منبع مجاز نیست!) ولی کپی‌کردن پاسخ مورد قبول نیست. در صورت تشخیص، نمره‌ای به هر دو فرد تعلق نمی‌گیرد. با توجه به نظر مصحح امکان پرسش شفاهی درباره پاسخ‌ها وجود دارد. استفاده از تمام کتابخانه‌های پایتون مجاز است.

۲ بررسی داده‌ها (۳۰ نمره)

۱.۲ لود کردن داده‌ها (۱۰ نمره)

ابتدا داده‌ها را از فایل CSV بخوانید. بعضی از ویژگی‌های داده‌ها عدد نیستند و نیاز به تبدیل آنها به عدد دارید (برای مثال درباره one-hot encoding جست‌وجو کنید). در این دیتاست شما باید ستون آخر را با استفاده از بقیه ستون‌ها برچسب‌گذاری کنید. ستون آخر یکی از برچسب‌های ۰ یا ۱ را دارد.

۲.۲ بررسی و نمایش (۲۰ نمره)

هیستوگرام مقادیر هر ویژگی از داده‌ها را رسم کنید. سپس دو ویژگی تصادفی از بین ویژگی‌های داده‌ها انتخاب کنید و آن‌ها را با دو رنگ متناظر با دو کلاس رسم کنید. این مرحله (رسم کلاس‌ها بر اساس دو ویژگی تصادفی) را ۵ بار تکرار کنید. همینطور مستقل‌بودن این جفت ویژگی‌های انتخاب‌شده را بررسی کنید (آیا ویژگی‌ها با هم همبستگی دارند؟).

۳ آموزش مدل (۴۰ نمره)

برای انجام مراحل زیر و گزارش دقت باید ۳۰ درصد داده‌ها را به صورت تصادفی به عنوان داده تست جدا کنید و بقیه داده‌ها را به عنوان داده آموزش در نظر بگیرید. برای بررسی دقیق‌تر شما باید این کار را ۵ بار تکرار کنید و میانگین و واریانس امتیازهای خواسته شده را گزارش کنید. یعنی برای ۵ بار شما باید ۷۰ درصد داده‌ها را به صورت تصادفی به عنوان داده آموزش و بقیه را به عنوان داده تست انتخاب کنید، مراحل قسمت زیر رو انجام و مدل خود را آموزش بدهید، و در نهایت میانگین و واریانس امتیازهای خواسته شده را گزارش کنید. **برای مسابقه، کد شما توسط مصحح روی دیتای جدا تست خواهد شد!**

۱.۳ انتخاب مدل و هایپرپارامتر (۲۰ نمره)

دو مدل که به نظر شما میتواند دقت مناسبی روی این دیتاست داشته باشد را انتخاب کنید و با استفاده از داده‌های آموزش و روش cross-validation بهترین هایپرپارامترها برای این مدل‌ها را بدست بیاورید.

۲.۳ آموزش مدل و نتایج (۲۰ نمره)

مدل خود را آموزش دهید. سپس مقادیر دقت طبقه‌بندی، precision، recall و F1 score مدل خود روی داده‌های تست را گزارش کنید (کافی است میانگین و واریانس این مقادیر برای ۵ بار آزمایش خود را گزارش کنید). همینطور برای آخرین مدل آموزش داده‌شده، Confusion Matrix را رسم کنید.

۳.۳ مسابقه (تا ۲۰ نمره امتیازی)

یک فایل پایتون با نام test.py ایجاد کنید که یک فایل به اسم test.csv با فرمت داده آموزش را در مسیر حاضر بخواند و دقت بهترین مدلتان را روی این داده چاپ کند. دقت کنید که فرمت داده کاملا مشابه فرمت داده‌ای است که در اختیار شما قرار داده شده است (اگر پیش‌پردازشی روی داده‌های آموزش انجام دادید، کد آن را برای اجرای دوباره در این قسمت آماده کنید). همینطور باید تمام کتابخانه‌ها و موارد مورد نیازتان را import کرده باشید و کد باید اجرا شود. یعنی شما باید با داده‌ای که در اختیار دارید یک مدل به همراه هایپرپارامترهای آن برای مسابقه انتخاب کنید. مصحح روی داده تست (که در اختیار شما نیست) مدل شما را امتحان می‌کند و دقت شما در مسابقه مورد استفاده قرار می‌گیرد.

۴ بررسی نتایج و توضیح مدل (۳۰ نمره)

۱.۴ بررسی مدل (۲۰ نمره)

مرز تصمیم‌گیری بهترین مدلتان را بررسی کنید. برای این‌کار می‌توانید از ماژول Inspection در کتابخانه sklearn استفاده کنید و مرزهای تصمیم‌گیری مدلتان را بر اساس دو ویژگی انتخابی رسم کنید. همینطور اگر از روش‌های درختی استفاده می‌کنید، درخت مدلتان را رسم کنید.

۲.۴ بررسی اثر ویژگی‌ها (۱۰ نمره)

بهترین طبقه‌بند خود را بر اساس یکی از ویژگی‌ها آموزش دهید و این کار را برای تمام ویژگی‌ها تکرار کنید. دقت بر اساس هر ویژگی را گزارش کنید و بهترین ویژگی‌ها برای محاسبه دقت را گزارش کنید. این‌کار چه اهمیتی در توضیح مدل دارد؟ درباره Feature importance تحقیق کنید.