دانشگاه تربیت مدرس

دانشکده مهندسی برق و کامپیوتر

تمرین پنجم درس یادگیری ماشین

ماشین‌های بردار پشتیبان و یادگیری تجمیعی

استاد محترم درس:

جناب آقای دکتر قادری

دستیاران آموزشی:

محمد مولائی

نیلوفر مقدس

نیمسال اول ۱۴۰۴-۱۴۰۳

نکات مربوط به تحویل پروژه

- ✓ پروژه فقط با زبان برنامه نویسی پایتون قابل قبول میباشد.
- ✓ فایل تحویلی شما ،یک فایل زیپ شده نهایی شامل گزارش کار (فایل pdf) و فایل کد پایتون با پسوند ipynb (Jupyter Notebook)) می‌باشد. لطفا آن را به صورت زیر نام‌گذاری و ارسال نمایید.
- ✓

HW_۵_HMM_[LastName]_[FirstName]

- ✓ گزارش کار خود را در یک فایل pdf تحویل دهید و از گذاشتن صرفا اسکرین شات‌های پشت سرهم از کد در گزارش کار خودداری کنید.
- ✓ توجه داشته باشید که در فایل ارسالی پایتون، خروجی هر سلول (شامل نمودار، خروجی عددی و غیره) حتما ذخیره شده و قابل مشاهده باشد.
- ✓ لازم است حتما نتایج بدست آمده را گزارش و تحلیل کنید.
- ✓ علاوه بر مهارت حل سوالات، نوشتن پاسخ مینی پروژه‌ها در فرمت گزارش فنی (فصل بندی و صفحه بندی مناسب، رعایت اصول نگارش و ...) برای دانشجویان تحصیلات تکمیلی اهمیت دارد، این مورد نیز در ارزشیابی لحاظ می‌شود.
- ✓ در صورت فراموشی در ارسال کد پایتون، هیچ نمره‌ای به شما تعلق نخواهد گرفت.
- ✓ در صورت مشاهده تشابه در هر بخش از انجام پروژه، نمره هر دو نفر صفر لحاظ میگردد.

برای پاسخ به سوالات دانشجویان در مورد مینی‌پروژه ها، دو راه ارتباطی وجود دارد:

1. ایمیل: برای پرسش سوال از طریق ایمیل، در قسمت To، ایمیل دستیار آموزشی این مینی پروژه، آقای مولائی (mhmdmovlaie@gmail.com) را قرار دهید و در قسمت Cc، ایمیل niloofar.moghaddas[at]gmail.com را قرار دهید.
2. گروه تلگرام: می توانید برای پرسش سوال از مینی پروژه چهارم در گروه، آقای مولائی را با شناسه تلگرامی @mhmdmolaee، در پیام خود نام ببرید.

ایمیل دستیار آموزشی: mhmdmovlaie@gmail.com

مقدمه

هدف از انجام این مینی‌پروژه اعمال و بررسی عملکرد الگوریتم‌های SVM و Ensemble Learning بر روی مجموعه‌داده‌ی معرفی شده است.

در این تمرین یک فایل notebook با خروجی هر سلول در اختیار شما قرار داده شده است و شما می‌بایست تلاش کنید تا خروجی هر سلول مجددا تولید کنید (نتایج مدل‌ها الزامی نیست).

- **مجموعه‌داده Bank Marketing**

این مجموعه‌داده برای کمپین بازاریابی یک موسسه‌ی بانکی پرتغالی است. هدف از طبقه‌بندی پیش‌بینی ثبت یا عدم ثبت سپرده (deposit) در این بانک است. هر یک از ویژگی‌های این مجموعه داده در این آدرس توضیح داده شده است.

- **معیار ارزیابی**

با توجه به مجموعه‌داده از معیارهای Accuracy، Precision، Recall و F۱_score استفاده می‌کنیم.

پروژه پنجم: ماشین هسته و یادگیری تجمعی

- **گام اول EDA**:

در این بخش مجموعه‌داده را از دید آماری مورد بررسی قرار دهید و با چالش‌های مجموعه‌داده آشنا شوید. مواردی که می‌بایست بررسی کنید:

الف) آیا در مجموعه داده مقادیر گمشده وجود دارد؟ این مورد منجر به چه مشکلاتی می‌شود؟ چگونه می‌توان با آن مقابله کرد؟

ب) توزیع کلاس خروجی به چه صورت است؟

ج) همبستگی ویژگی‌ها با یک دیگر به چه صورت است؟ همبستگی میان دو ویژگی از رابطه‌ی زیر محاسبه می‌شود.

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \, \sigma_Y}$$

این معیار نشان می‌دهد که آیا ارتباط خطی میان دو ویژگی وجود دارد یا خیر. آیا در صورتی که این معیار بین یکی از ویژگی‌ها و کلاس خروجی نزدیک به صفر بود، می‌توان آن را حذف کرد؟ چرا؟

د) آیا در مجموعه‌داده، داده‌ی پرت وجود دارد؟ این مورد منجر به چه مشکلاتی می‌شود؟ چگونه می‌توان با آن مقابله کرد؟

ه) پس از کاهش ابعاد، با رسم داده‌ها به چه نتیجه‌ایی می‌توان رسید؟ کدام روش کاهش ابعاد مناسب‌تر به نظر می‌رسد؟ چرا؟

در نظر داشته باشید که انجام این مراحل الزامی است و در صورت تمایل می‌توانید دیدهای بیشتر و متفاوت‌تری مجموعه‌داده را بررسی کنید.

- **گام دوم: پیش‌پردازش**

در این مرحله، بعد از آشنایی با مجموعه‌داده، پیش پردازش‌هایی که از نظرتان الزامی هستند را اعمال کنید. انجام مراحل ذکر شده الزامی است.

الف) تقسیم مجموعه‌داده به بخش‌های train و test.

ب) normalize کردن داده‌های عددی

- **گام سوم: یادگیری و انتخاب مدل**

در این بخش در ابتدا الگوریتم SVM و یک مدل یادگیری تجمیعی مانند Random Forest و XGBoost را بر روی داده‌های پیش‌پردازش شده اعمال کنید. نتایج هر مدل را با مشخص کردن پیش پردازش اعمال شده، در یک جدول نمایش دهید (نتایج می‌بایست با استفاده از Cross-validation تعیین شود). سپس چند مدل برتر از میان این مدل‌ها انتخاب کنید و با استفاده از روش‌هایی مانند رای اکثریت یک مدل واحد بسازید. نتیجه این مدل را نیز به انتهای جدول اضافه کنید. بهترین

مدل را با توجه به نتایج انتخاب کنید. توجه داشته باشید که استفاده از مدل‌های نام برده شده الزامی است ولی محدود به این مدل‌ها نیستید.

۱ مدل، پیش‌پردازش و نتایج بدست آمده

الگوریتم	پیش‌پردازش	میانگین ± انحراف معیار Accuracy	میانگین ± انحراف معیار Precision	میانگین ± انحراف معیار Recall	میانگین ± انحراف معیار F۱_score
SVM (rbf)	Missing Values (Method), PCA, etc.	۷۰ ± ۰,۵	…	…	…
جنگل تصادفی	Forward Feature Selection	…	…	…	…

- **گام چهارم: ارزیابی مدل**

توضیح دهید که بهترین مدل را چگونه انتخاب می‌کنید و آن را بر روی مجموعه تست اعمال کنید. توجه کنید که داده‌های تست را **فقط و فقط یک بار** می‌توانید به مدل بدهید، در غیر این صورت نتایج معتبر نخواهند بود.

- **گام پنجم: بحث و نتیجه‌گیری**

نتایج بدست آمده از هر بخش را به صورت مجزا گزارش کنید و علت هرکدام از پیش‌پردازش‌هایی که اعمال کرده‌اید را شرح دهید. همچنین توضیح دهید، به نظرتان چرا بهترین نتیجه را از آن الگوریتم بدست آورده‌اید؟