



مینی پروژه شماره سه

در انجام این مینی پروژه حتماً به نکات زیر توجه کنید:

- موعد تحویل این مینی پروژه، ساعت ۱۸:۰۰ روز یکشنبه ۲۰ خردادماه ۱۴۰۳ است.
- اطلاعات خود را در **این گوگل شیت** مطابق نمونه سطر دوم تکمیل کنید.
- برای این مینی پروژه ملزم به ارائه گزارش متنی شامل توضیحات کامل هر قسمت هستید. هم گزارش و هم کدهای خود را در گیت‌هاب و سامانه دانشگاه بارگذاری کنید.
- برای گزارش لازم است که پاسخ هر سوال و زیربخش‌هایش به ترتیب و به صورت مشخص نوشته شده باشند. بخش زیادی از نمره به توضیحات دقیق و تحلیل‌های کافی شما روی نتایج بستگی خواهد داشت.
- لازم است که در صفحه اول گزارش خود لینک پوشه گیت‌هاب و گوگل کولب مربوط به مینی پروژه خود را در حالتی که دسترسی Public دارد به اشتراک گذاشته باشید. دفترچه‌کد گوگل کولب باید به صورت منظم و با بخش‌بندی مشخص تنظیم شده باشد، و خروجی سلول‌های اجرا شده قابل مشاهده باشد. در گیت‌هاب هم برای هر مینی پروژه یک پوشه مجزا ایجاد کنید.
- هر جا از دفترچه‌کد گوگل کولب شما نیاز به فراخوانی فایلی خارج از محیط داشت، مطابق آموزش‌های ارائه شده ملزم هستید از دستور **gdown** استفاده کنید و مسیرهای فایل‌ها را طوری تنظیم کنید که صرفاً با اجرای سلول‌های کد، امکان فراخوانی و خواندن فایل‌ها توسط هر کاربری وجود داشته باشد.
- در تمامی مراحل تعریف داده و مدل و هر جای دیگری که مطابق آموزش‌های ویدیویی و به لحاظ منطقی نیاز است، Random State را برابر با دو رقم آخر شماره دانشجویی خود در نظر بگیرید.
- استفاده از ابزارهای هوشمند (مانند ChatGPT) در کمک‌گرفتن برای بهبود کدها مجاز است؛ اما لازم است تمام جزئیات مواردی که در خروجی‌های مختلف گزارش خود عنوان می‌کنید را به خوبی خوانده، درک و تحلیل کرده باشید. استفاده از این ابزارهای هوشمند در نوشتن گزارش و تحلیل‌ها ممنوع است.
- در جاهایی که با توجه به دو رقم آخر شماره دانشجویی خود محدود به انتخاب عدد، متغیر و یا داده‌ای خاص شده‌اید، برای تست‌های اضافه‌تر و نمایش بهبود در نتایج خود، مجاز هستید از مقادیر دیگر هم استفاده کنید. ۱۵ تا ۲۰ درصد از نمره هر سوال به بهترین نتایج کسب شده اختصاص خواهد یافت.
- رعایت نکات بالا به حرفه‌ای‌تر شدن شما کمک خواهد کرد و اهمیتی معادل مطالب درسی فراگرفته شده دارد؛ بنابراین، در صورت عدم رعایت هریک از این نکات، گزارش شما تصحیح نخواهد شد.

به انتخاب خود به یکی از پرسش‌های یک یا دو پاسخ دهید.

۱ پرسش یک

هدف از این سوال آزمایش الگوریتم SVM در نمونه‌های مختلف روی دیتاست معروف گل‌زنابق^۱ است. مراحل زیر را یک به یک انجام دهید و موارد خواسته شده در گزارش خود به همراه کدها ارسال کنید.

^۱ IRIS Dataset

آ. در مرحله اول دیتاست را فراخوانی کنید و اطلاعاتی نظیر ابعاد، تعداد نمونه‌ها، میانگین، واریانس و همبستگی ویژگی‌ها را به دست آورید و نمونه‌های دیتاست را به تصویر بکشید (مثلاً با استفاده از t-SNE). سپس، با توجه به اطلاعات عددی، آماری و بصری بدست آمده، تحلیل کنید که آیا کاهش ابعاد می‌تواند در این دیتاست قابل استفاده باشد یا خیر.

ب. با استفاده از الگوریتم SVM، با هسته خطی، داده‌ها را طبقه‌بندی کنید و ماتریس درهم‌ریختگی آن را بدست آورید و مرزهای تصمیم‌گیری را در فضای دوبعدی (کاهش بُعد از طریق یکی از روش‌های آموخته‌شده با ذکر دلیل) ترسیم کنید.

ج. بخش قبلی را با استفاده از هسته‌های چند جمله‌ای و با استفاده از کتابخانه scikit-learn از درجه یک تا ۱۰ پیاده سازی کنید و نتایج را با معیارهای مناسب گزارش کرده و مقایسه و تحلیل کنید. در نهایت، با استفاده از کتابخانه imageio جداسازی ویژگی‌های اصلی را (کاهش بُعد از طریق یکی از روش‌های آموخته‌شده با ذکر دلیل) برای درجات ۱ تا ۱۰ در قالب یک GIF به تصویر بکشید و لینک دسترسی مستقیم به فایل GIF را درون گزارش خود قرار دهید.

د. حال الگوریتم SVM را برای مورد قبلی، بدون استفاده از کتابخانه scikit-learn و به صورت From Scratch پیاده‌سازی کنید. در این بخش لازم است که یک کلاس SVM تعریف کنید. این کلاس می‌بایست حداقل دارای سه تابع (متد) Predict، Fit، Polynomial_kernel باشد. متد Polynomial_kernel می‌بایست با دریافت درجه‌های ۱ تا ۱۰، هسته‌های چندجمله‌ای را محاسبه کند. دقت الگوریتم را با افزایش درجه گزارش کنید و نتایج حاصل را با بخش قبلی مقایسه کنید. در این قسمت نیز جداسازی ویژگی‌های اصلی را برای درجات ۱ تا ۱۰ در قالب یک GIF به تصویر بکشید پیوند دسترسی مستقیم آن را در گزارش خود قرار دهید.

۲ پرسش دو

مقاله [GenSVM: A Generalized Multiclass Support Vector Machine](#) را مطالعه کنید و به صورت مختصر بخش‌های مربوط به طرح مشکل و مسأله، معرفی روش و جمع‌بندی در آن را گزارش کنید. توضیح دهید که ایده نوآورانه این مقاله چیست.

سپس، این مقاله را پیاده‌سازی کنید و نتایج بدست آمده را با نتایج مقاله مقایسه و تحلیل کنید. استفاده از کتابخانه‌ها و تمامی ابزارهای کدنویسی بلامانع است. برای پیاده‌سازی می‌توانید از دیتاست [IRIS](#) هم استفاده کنید.

۳ پرسش سه

مقاله [Credit Card Fraud Detection Using Autoencoder Neural Network](#) برای پیاده‌سازی این قسمت در نظر گرفته شده است. پس از مطالعه مقاله به سوالات زیر پاسخ دهید.

آ. بزرگ‌ترین چالش‌ها در توسعه مدل‌های تشخیص تقلب چیست؟ این مقاله برای حل این چالش‌ها از چه روش‌هایی استفاده کرده است؟

ب. در مورد معماری شبکه ارائه‌شده در مقاله به صورت مختصر توضیح دهید.

ج. مدل ارائه‌شده را پیاده‌سازی کرده و با استفاده از این دیتاست آموزش دهید. برای جلوگیری از بیش‌برازش، آموزش مدل را طوری تنظیم کنید که در انتهای آموزش، بهترین وزن‌های مدل بر اساس خطای قسمت اعتبارسنجی بازگردانده شود.

د. ماتریس درهم‌ریختگی را روی قسمت آزمون داده‌ها رسم کنید و مقادیر Recall، Precision، Accuracy و f1score را گزارش کنید. فکر می‌کنید در مسائلی که توزیع برچسب‌ها نامتوازن است، استفاده از معیاری مانند Accuracy به تنهایی عمل‌کرد مدل را به درستی نمایش می‌دهد؟ چرا؟ اگر نه، کدام معیار می‌تواند به عنوان مکمل استفاده شود؟

ه. با آستانه‌های مختلف برای Oversampling عمل‌کرد مدل را بررسی کرده و نمودار Recall & Accuracy را مانند شکل ۷ مقاله ترسیم کنید.

و. مدل را با استفاده از داده‌های نامتوازن و بدون حذف نویز، آموزش داده و موارد بخش قبلی را گزارش کنید و نتایج دو مدل را با هم مقایسه کنید.

دقت داشته باشید که هدف رسیدن به نتایج عیناً مشابه با مقاله نیست و در صورتی که مقادیر هریک از فرآپارامترهای مدل در مقاله ذکر نشده باشد، در انتخاب آن‌ها آزادی عمل کامل دارید.

منابع

[1] <https://github.com/MJAHMADEE/MachineLearning2024W>