مینی‌پروژهٔ شمارهٔ یک - بخش اول

نکات مهم و موعد تحویل مینی‌پروژه

- برای این مینی‌پروژه ملزم به ارائهٔ گزارش متنی شامل توضیحات کامل هر قسمت هستید. هم گزارش و هم کدهای خود را در گیت‌هاب و سامانهٔ دانشگاه بارگذاری کنید.
- در تمامی مراحل تعریف داده و مدل و هرجای دیگری که مطابق آموزش ویدیویی و به لحاظ منطقی نیاز است، Random State را برابر با دو رقم آخر شمارهٔ دانشجویی خود در نظر بگیرید.
- موعد تحویل این تمرین، ساعت ۲۳:۵۹ روز سه‌شنبه مورخ ۱۴۰۲/۰۸/۳۰ است.
- استفاده از دستیارهای هوشمند (مانند ChatGPT) آزاد است؛ اما حتماً باید برنامه‌ها و جزئیات پروژه‌های تحویلی خود را فهمیده باشید.

۱ سوال اول

۱. با استفاده از sklearn.datasets، یک دیتاست با ۱۰۰۰ نمونه، ۲ کلاس و ۲ ویژگی تولید کنید.

۲. با استفاده از حداقل دو طبقه‌بند آمادهٔ پایتون و در نظر گرفتن فراپارامترهای مناسب، دو کلاس موجود در دیتاست قسمت قبلی را از هم تفکیک کنید. ضمن توضیح روند انتخاب فراپارامترها (مانند تعداد دورهٔ آموزش و نرخ یادگیری)، نتیجهٔ دقت آموزش و ارزیابی را نمایش دهید. برای بهبود نتیجه از چه تکنیک‌هایی استفاده کردید؟

۳. مرز و نواحی تصمیم‌گیری برآمده از مدل آموزش‌دیدهٔ خود را به همراه نمونه‌ها در یک نمودار نشان دهید. اگر می‌توانید نمونه‌هایی که اشتباه طبقه‌بندی شده‌اند را با شکل متفاوت نمایش دهید.

۴. از چه طریقی می‌توان دیتاست تولیدشده در قسمت «۱» را چالش‌برانگیزتر و سخت‌تر کرد؟ این کار را انجام داده و قسمت‌های «۲» و «۳» را برای این داده‌های جدید تکرار و نتایج را مقایسه کنید.

۵. اگر یک کلاس به داده‌های تولیدشده در قسمت «۱» اضافه شود، در کدام قسمت‌ها از بلوک دیاگرام آموزش و ارزیابی تغییراتی ایجاد می‌شود؟ در مورد این تغییرات توضیح دهید. آیا می‌توانید در این حالت پیاده‌سازی را به راحتی و با استفاده از کتابخانه‌ها و کدهای آمادهٔ پایتونی انجام دهید؟ پیاده‌سازی کنید.

۲ سوال دوم

۱. با مراجعه به این پیوند با یک دیتاست مربوط به حوزهٔ «بانکی» آشنا شوید و ضمن توضیح کوتاه اهداف و ویژگی‌هایش، فایل آن را دانلود کرده و پس از بارگذاری در گوگل‌درایو خود، آن را با دستور gdown در محیط گوگل‌کولب قرار دهید. اگر تغییر فرمتی برای فایل این دیتاست نیاز می‌بینید، این کار را با دستورهای پایتونی انجام دهید.

۲. ضمن توضیح اهمیت فرآیند بُرزدن (مخلوط کردن)[1]، داده‌ها را مخلوط کرده و با نسبت تقسیم دلخواه و معقول به دو بخش «آموزش» و «ارزیابی» تقسیم کنید.

۳. بدون استفاده از کتابخانه‌های آمادهٔ پایتون، مدل، تابع اتلاف و الگوریتم یادگیری و ارزیابی را کدنویسی کنید تا دو کلاس موجود در دیتاست به خوبی از یکدیگر تفکیک شوند. نمودار تابع اتلاف را رسم کنید و نتیجهٔ دقت ارزیابی روی داده‌های تست را محاسبه کنید. نمودار تابع اتلاف را تحلیل کنید. آیا می‌توان از روی نمودار تابع اتلاف و قبل از مرحلهٔ ارزیابی با قطعیت در مورد عملکرد مدل نظر داد؟ چرا و اگر نمی‌توان، راه‌حل چیست؟

۴. حداقل دو روش برای نرمال‌سازی داده‌ها را با ذکر اهمیت این فرآیند توضیح دهید و با استفاده از یکی از این روش‌ها، داده‌ها را نرمال کنید. آیا از اطلاعات بخش «ارزیابی» در فرآیند نرمال‌سازی استفاده کردید؟ چرا؟

۵. تمام قسمت‌های «۱» تا «۳» را با استفاده از داده‌های نرمال‌شده تکرار کنید و نتایج پیش‌بینی مدل را برای پنج نمونه داده نشان دهید.

۶. با استفاده از کدنویسی پایتون وضعیت تعادل داده‌ها در دو کلاس موجود در دیتاست را نشان دهید. آیا تعداد نمونه‌های کلاس‌ها با هم برابر است؟ عدم تعادل در دیتاست می‌تواند منجر به چه مشکلاتی شود؟ برای حل این موضوع چه اقداماتی می‌توان انجام داد؟ پیاده‌سازی کرده و نتیجه را مقایسه و گزارش کنید.

۷. فرآیند آموزش و ارزیابی مدل را با استفاده از یک طبقه‌بند آمادهٔ پایتونی انجام داده و این‌بار در این حالت چالش عدم تعادل داده‌های کلاس‌ها را حل کنید.

۳ سوال سوم

۱. به این پیوند مراجعه کرده و یک دیتاست مربوط به «بیماری قلبی» را دریافت کرده و توضیحات مختصری در مورد هدف و ویژگی‌های آن بنویسید. فایل دانلودشدهٔ دیتاست را روی گوگل‌درایو خود قرار داده و با استفاده از دستور gdown آن را در محیط گوگل‌کولب بارگذاری کنید.

۲. ضمن توجه به محل قرارگیری هدف و ویژگی‌ها، دیتاست را به‌صورت یک دیتافریم درآورده و با استفاده از دستورات پایتونی، ۱۰۰ نمونه‌داده مربوط به کلاس «۱» و ۱۰۰ نمونه‌داده مربوط به کلاس «۰» را در یک دیتافریم جدید قرار دهید و در قسمت‌های بعدی با این دیتافریم جدید کار کنید.

۳. با استفاده از حداقل دو طبقه‌بند آمادهٔ پایتون و در نظر گرفتن فراپارامترهای مناسب، دو کلاس موجود در دیتاست را از هم تفکیک کنید. نتیجهٔ دقت آموزش و ارزیابی را نمایش دهید.

۴. در حالت استفاده از دستورات آمادهٔ سایکیت‌لرن، آیا راهی برای نمایش نمودار تابع اتلاف وجود دارد؟ پیاده‌سازی کنید.

۵. یک شاخصهٔ ارزیابی (غیر از Accuracy) تعریف کنید و بررسی کنید که از چه طریقی می‌توان این شاخص جدید را در ارزیابی داده‌های تست نمایش داد. پیاده‌سازی کنید.

منابع

[1] https://github.com/MJAHMADEE/MachineLearning2023

[1] Data Shuffling