مینی‌پروژهٔ شمارهٔ یک

در انجام این مینی‌پروژه حتماً به نکات زیر توجه کنید:

- موعد تحویل این مینی‌پروژه، ساعت ۱۸:۰۰ روز جمعه ۱۷ فرودین‌ماه ۱۴۰۳ است.
- اطلاعات خود را در این گوگل‌شیت مطابق نمونهٔ سطر دوم تکمیل کنید.
- 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.
- برای گزارش لازم است که پاسخ هر سوال و زیربخش‌هایش به‌ترتیب و به‌صورت مشخص نوشته شده باشند. بخش زیادی از نمره به توضیحات دقیق و تحلیل‌های کافی شما روی نتایج بستگی خواهد داشت.
- لازم است که در صفحهٔ اول گزارش خود لینک پوشهٔ گیت‌هاب و گوگل‌کولب مربوط به مینی‌پروژهٔ خود را در حالتی که دسترسی Public دارد به اشتراک گذاشته باشید. دفترچه‌کد گوگل‌کولب باید به‌صورت منظم و با بخش‌بندی مشخص تنظیم شده باشد، و خروجی سلول‌های اجراشده قابل مشاهده باشد. در گیت‌هاب هم برای هر مینی‌پروژه یک پوشهٔ مجزا ایجاد کنید.
- هرجا از دفترچه‌کد گوگل‌کولب شما نیاز به فراخوانی فایلی خارج از محیط داشت، مطابق آموزش‌های ارائه‌شده 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 هرجای دیگری که مطابق آموزش‌های ویدیویی و به لحاظ منطقی نیاز است، Random State را برابر با دو رقم آخر شمارهٔ 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 شمارهٔ 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 نمرهٔ هر سوال به بهترین نتایج کسب‌شده اختصاص خواهد یافت.
- رعایت نکات بالا به حرفه‌ای‌ترشدن شما کمک خواهد کرد و اهمیتی معادل مطالب درسی فراگرفته‌شده دارد؛ بنابراین، درصورت عدم رعایت هریک از این نکات، گزارش شما تصحیح نخواهد شد.

۱ سوال اول

۱. فرآیند آموزش و ارزیابی یک مدل طبقه‌بند خطی را به‌صورت دیاگرامی بلوکی نمایش دهید و در مورد اجزای مختلف این دیاگرام بلوکی توضیحاتی بنویسید. تغییر نوع طبقه‌بندی از حالت دوکلاسه به چندکلاسه در کدام قسمت از این دیاگرام بلوکی تغییراتی ایجاد می‌کند؟ توضیح دهید.

۲. با استفاده از sklearn.datasets، یک دیتاست با ۱۰۰۰ نمونه، ۴ کلاس و ۳ ویژگی تولید کنید و آن را به‌صورتی مناسب نمایش دهید. آیا دیتاستی که تولید کردید چالش برانگیز است؟ چرا؟ به چه طریقی می‌توانید دیتاست تولیدشده خود را چالش‌برانگیزتر و سخت‌تر کنید؟

۳. با استفاده از حداقل دو طبقه‌بند خطی آمادهٔ پایتون (در sklearn.linear_model) و در نظر گرفتن فراپارامترهای مناسب، چهار 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؟

۴. 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 و رنگ متفاوت نمایش دهید.

۵. فرآیندی مشابه قسمت «۲» را با تعداد کلاس و ویژگی دلخواه؛ اما با استفاده از ابزار drawdata تکرار کنید. قسمت‌های «۳» و «۴» را برای این داده‌های جدید تکرار و نتایج را به‌صورتی مناسب نشان دهید.

۲ سوال دوم

۱. با مراجعه به صفحهٔ دیتاست CWRU Bearing با یک دیتاست مربوط به حوزهٔ «تشخیص عیب» آشنا شوید. با جستجوی آن در اینترنت و مقالات، توضیحاتی از اهداف، ویژگی‌ها و حالت‌های مختلف این دیتاست ارائه کنید.
در ادامه، ابتدا به صفحهٔ داده‌های سالم مراجعه کنید و داده‌های کلاس سالم (Normal_X[1]) را دریافت کنید. سپس، به صفحهٔ داده‌های عیب در حالت 12k مراجعه کرده و داده‌های کلاس عیب (IR007_X) را دریافت کنید.

۲. برای تشکیل دیتاست مراحل زیر را انجام دهید:

آ) از هر کلاس M نمونه با طول N جدا کنید (M حداقل ۱۰۰ و N حداقل ۲۰۰ باشد). یک ماتریس از داده‌های هر دو کلاس به همراه برچسب مربوطه تشکیل دهید. می‌توانید پنجره‌ای به طول N در نظر بگیرید و در نهایت یک ماتریس $M \times N$ از داده‌های هر کلاس استخراج کنید.

ب) در مورد اهمیت استخراج ویژگی در یادگیری ماشین توضیحاتی بنویسید. سپس، با استفاده از حداقل ۸ عدد از روش‌های ذکرشده در جدول ۱، ویژگی‌های دیتاست قسمت «۲ـآ» را استخراج کنید و یک دیتاست جدید تشکیل دهید.

جدول ۱: ویژگی‌های پیشنهادی برای استخراج از دیتاست.

Feature	Formula	Feature	Formula
Standard Deviation	$x_{std} = \sqrt{\frac{\sum_{i=1}^{N}(x(i)-\bar{x})^2}{N}}$	Shape Factor	$\mathrm{SF} = \frac{x_{rms}}{\frac{1}{N}\sum_{i=1}^{N}\lvert x(i)\rvert}$
Peak	$x_p = \max\lvert x(i)\rvert$	Impact Factor	$\mathrm{IF1} = \frac{x_p}{\frac{1}{N}\sum_{i=1}^{N}\lvert x(i)\rvert}$
Skewness	$x_{\mathrm{ske}} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x(i)-\bar{x})^3}{x_{\mathrm{std}}^3}$	Square Mean Root	$x_{smr} = \left(\frac{1}{N}\sum_{i=1}^{N}\sqrt{\lvert x(i)\rvert}\right)^2$
Kurtosis	$x_{kur} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x(i)-\bar{x})^4}{x_{std}^4}$	Mean	$\mathrm{Mean} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Crest Factor	$\mathrm{CF} = \frac{x_p}{x_{rms}}$	Absolute Mean	$\mathrm{Abs\ Mean} = \frac{1}{n}\sum_{i=1}^{n}\lvert x_i\rvert$
Clearance Factor	$\mathrm{CLF} = \frac{x_p}{x_{smr}}$	Root Mean Square	$\mathrm{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} {x_i}^2}$
Peak to Peak	Maximum - Minimum	Impulse Factor	$\mathrm{IF2} = \frac{AbsMax}{\frac{1}{n}\sum_{i=1}^{n}\lvert x_i\rvert}$

ج) ضمن توضیح اهمیت فرآیند بُرزدن (مخلوط کردن)[2]، داده‌ها را مخلوط کرده و با نسبت تقسیم دلخواه و معقول به دو بخش «آموزش» و «ارزیابی» تقسیم کنید.

د) 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؟

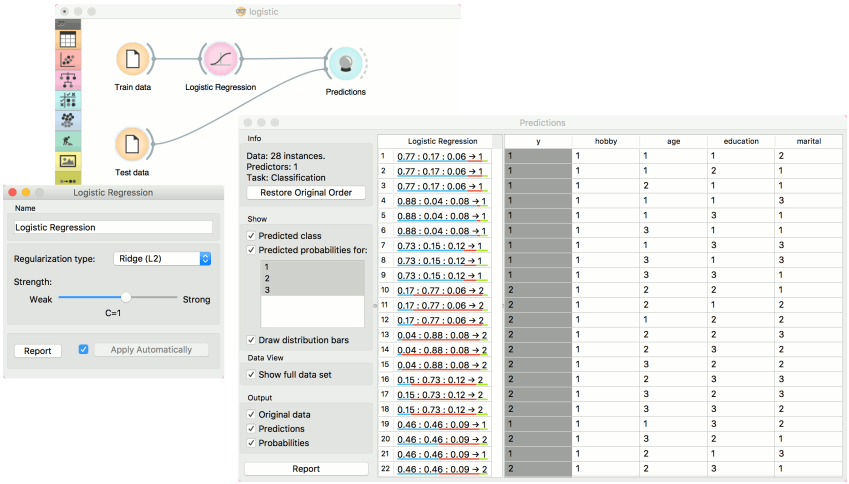
۳. بدون استفاده از کتابخانه‌های آمادهٔ پایتون، مدل طبقه‌بند، تابع اتلاف و الگوریتم یادگیری و ارزیابی را کدنویسی کنید تا دو کلاس موجود در دیتاست به خوبی از یکدیگر تفکیک شوند. نمودار تابع اتلاف را رسم کنید و نتیجهٔ ارزیابی روی داده‌های تست را با حداقل ۲ شاخصه محاسبه کنید. نمودار تابع اتلاف را تحلیل کنید. 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. فرآیند آموزش و ارزیابی را با استفاده از یک طبقه‌بند خطی آمادهٔ پایتون (در sklearn.linear_model) انجام داده و نتایج را مقایسه کنید. در حالت استفاده از دستورات آمادهٔ سایکیت‌لرن، آیا راهی برای نمایش نمودار تابع اتلاف وجود دارد؟ پیاده‌سازی کنید.

[1] X، باقی‌ماندهٔ تقسیم دو رقم شمارهٔ دانشجویی شما بر ۴ است.

[2] Data Shuffling

۵. در مورد نرم‌افزار داده‌کاوی Orange و قابلیت‌های آن تحقیق کنید و سعی کنید این سوال یا یک مثال ساده‌تر را با استفاده از این نرم‌افزار پیاده‌سازی کنید (راهنمایی: می‌توانید از پیوندهای ۱، ۲ و ۳ کمک بگیرید). پاسخ به این قسمت از سوال اختیاری و امتیازی است. می‌توانید عملکرد خود را به‌صورت تصویری و یا ویدیویی هم نشان دهید. مقدار نمرهٔ امتیازی، وابسته به جامعیت مثال بررسی‌شده و استفاده از ویژگی‌های مختلف این ابزار است.



شکل ۱: نمایی از رگرسیون لجستیک در نرم‌افزار داده‌کاوی Orange.

۳ سوال سوم

یک دیتاست در زمینهٔ آب و هوا با نام Weather in Szeged 2006-2016 را در نظر بگیرید. در این دیتاست هدف آن است که ارتباط بین Humidity با Temperature و هم‌چنین ارتباط بین Humidity و Apparent Temperature پیدا شده و با کمک داده‌های Humidity و Temperature تخمین انجام شود.

۱. ابتدا هیت‌مپ ماتریس همبستگی و هیستوگرام پراکندگی ویژگی‌ها را رسم و تحلیل کنید.

۲. روی این دیتاست، تخمین LS و RLS را با تنظیم پارامترهای مناسب اِعمال کنید. نتایج به‌دست‌آمده را با محاسبهٔ خطاها و رسم نمودارهای مناسب برای هر دو مدل با هم مقایسه و تحلیل کنید.

۳. در مورد Weighted Least Square توضیح دهید و آن را روی دیتاست داده‌شده اِعمال کنید.

۴. در مورد الگوریتم QR-Decomposition-Based RLS تحقیق کنید. پاسخ به این قسمت از سوال اختیاری و امتیازی است.

منابع

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

[2] R. Magar, L. Ghule, J. Li, Y. Zhao, and A. B. Farimani, “FaultNet: A Deep Convolutional Neural Network for Bearing Fault Classification,” IEEE Access, vol. 9. Institute of Electrical and Electronics Engineers (IEEE), pp. 25189–25199, 2021. doi: 10.1109/access.2021.3056944.