دانشگاه صنعتی خواجه نصیرالدین طوسی تهران
مبانی سیستم‌های هوشمند (پاییز ۱۴۰۴)، مهدی علیاری شوره‌دلی

مینی‌پروژهٔ ۱

پرسش‌های تحلیلی

۱ پرسش یک

یک مجموعه‌داده دارای ۴ کلاس است:

(الف) روابط مربوط به محاسبه‌ی **Sensitivity** و **Specificity** را بنویسید.

(ب) اگر وضعیت پیش‌بینی کلاس‌های ذکرشده به صورت زیر باشد، مقادیر **FN TN، FP، TP،** را برای هر کلاس بنویسید، سپس **Sensitivity** و **Specificity** آن‌ها را محاسبه کنید.

	C_1'	C_2'	C_3'	C_4'
C_1	45	3	2	1
C_2	3	32	2	6
C_3	2	2	16	10
C_4	0	2	0	20

۲ پرسش دو

از الگوریتم یادگیری پرسپترون با نرخ یادگیری $\eta = 1$ استفاده کنید تا صفحه‌ای (در اینجا: خطِ جداساز دوبعدی) به‌دست آید که مجموعه‌های A^+ و A^- زیر را به‌صورت خطی از هم جدا کند:

$$A^+ = \{(1,1),(0,2),(3,0)\}, \qquad A^- = \{(-2,-1),(0,-2)\}.$$

الف) فرض کنید $w(0)$ بردارِ صفر است. الگوریتم را طوری تغییر دهید که بایاس b همزمان با مؤلفه‌های w محاسبه شود. برای این کار، به هر بردارِ آموزشی یک مؤلفهٔ اضافی «1» اضافه کنید و سپس $w' = (w_1, w_2, b)$ را به‌دست آورید. خطِ جداساز حاصل را همراهِ داده‌ها رسم کنید.

ب) این مسئله را با روش کمترین مربعات حل کنید. خطِ به‌دست‌آمده را در همان شکلِ قسمت (الف) رسم کنید و خطوط را با هم مقایسه نمایید.

پ) این مسئله را با روش فیشر حل کنید و مطابقِ توضیحِ قسمت (ب)، خطِ جداساز را رسم نمایید.
کدام روش جداساز بهتری است ؟ از روی بردار وزن ها تحلیل کنین .

۳ پرسش سوم

فرض کنید یک دیتاست شامل دو ویژگی x_1 و x_2 داریم. دامنه‌ی تغییرات ویژگی‌ها به صورت زیر است:

$$0 < x_1 < 1000 \qquad , \qquad 0 < x_2 < 1$$

اگر بدون انجام هیچ‌گونه پیش‌پردازش یا مقیاس‌بندی، مستقیماً PCA را روی این داده‌ها اعمال کنیم، انتظار می‌رود چه تأثیری بر روی مؤلفه‌های اصلی (Principal Components) ایجاد شود؟

۱. از دیدگاه ریاضی توضیح دهید چرا این اتفاق می‌افتد.
(با اشاره به ماتریس کوواریانس و تجزیه‌ی مقادیر و بردارهای ویژه ـ Eigenvalue Decomposition)

۲. مشخص کنید کدام ویژگی در مؤلفه‌ی اصلی غالب خواهد بود و چرا.

۳. چه مرحله/مراحلی از پیش‌پردازش پیشنهاد می‌شود تا PCA به‌درستی رفتار کند، و دلیل آن را توضیح دهید.

سوالات پیاده‌سازی

۴ پرسش چهارم

تصور کنید یک شرکت مخابراتی مشتریان خود را بر اساس الگوهای استفاده از خدمات، به چهار گروه تقسیم کرده است. اگر بتوان با استفاده از داده‌های جمعیت‌شناختی، عضویت در گروه‌ها را پیش‌بینی کرد، شرکت می‌تواند پیشنهادهای ویژه‌ای برای مشتریان احتمالی ارائه دهد. این مسئله یک مسئله‌ی طبقه‌بندی است. یعنی با داشتن مجموعه‌داده‌ای با برچسب‌های از پیش تعیین‌شده، می‌توان مدلی ارائه کرد که بتواند کلاس مورد نظر یک نمونه‌ی جدید یا ناشناخته را پیش‌بینی کند.

در دنیای مدرن یادگیری عمیق، طراحی و آموزش شبکه‌های عصبی به یکی از مهم‌ترین چالش‌ها در حوزه‌ی یادگیری ماشین تبدیل شده است. در این سؤال از داده‌های جمعیت‌شناختی مانند منطقه‌ی جغرافیایی، سن و وضعیت تأهل برای پیش‌بینی الگوهای مصرف استفاده می‌کنیم. به این منظور از مجموعه‌داده‌ی telecust1000t بهره می‌بریم. در این مجموعه‌داده، فیلد هدف که custcat نام دارد، دارای چهار مقدار ممکن است که به چهار گروه مشتریان مربوط می‌شوند:

- خدمات پایه
- خدمات الکترونیکی
- خدمات پیشرفته
- خدمات کامل

منبع داده‌ها

دیتاست مورد استفاده در این پروژه از طریق لینک زیر در دسترس است: دانلود دیتاست Telecust1000t

۱. بخش اول: تحلیل اکتشافی داده‌ها (EDA)

- فایل داده را با pandas بخوانید و با استفاده از توابع `.info()` و `.describe()`. ساختار کلی و خلاصه‌ی آماری داده‌ها را به‌دست آورید.
- بررسی کنید آیا داده‌های گمشده (NaN) وجود دارند یا خیر. در صورت وجود، سه روش مختلف برای رفع آن‌ها (میانگین، میانه، حدس بر اساس نزدیک‌ترین همسایه و غیره) را پیاده‌سازی و مقایسه کنید.
- نوع ویژگی‌ها را مشخص کنید (عددی یا طبقه‌ای) و تفاوت این دو نوع را توضیح دهید.
- با استفاده از کتابخانه‌های `seaborn` و `plotly`:
  - Heatmap همبستگی ویژگی‌ها را رسم و ویژگی‌هایی با بیشترین همبستگی با متغیر هدف را مشخص کنید.
  - Pairplot یا Plot Scatter چند ویژگی مهم را نمایش دهید.
  - برای دو ویژگی عددی مهم، نمودار Hexbin رسم کنید و ارتباط آن‌ها با خروجی را تفسیر کنید.
  - با استفاده از countplot و plot pie توزیع کلاس‌ها را نمایش دهید و درباره‌ی تعادل داده‌ها توضیح دهید.

۲. بخش دوم: پیش‌پردازش داده‌ها

- ویژگی‌های طبقه‌ای را با یکی از روش‌های Encoding One-Hot یا Encoding Label به داده‌ی عددی تبدیل کنید.
- داده‌های عددی را نرمال‌سازی یا استانداردسازی کنید و هدف از این کار را توضیح دهید.
- در صورت وجود ویژگی‌های غیرمفید یا تکراری، آن‌ها را حذف کرده و دلیل تصمیم خود را بیان کنید.

۳. بخش سوم: انتخاب ویژگی و مدل‌سازی کلاسیک

- برای انتخاب ویژگی‌های مؤثر از دو روش زیر استفاده کنید و نتایج آن‌ها را مقایسه نمایید:
  - رگرسیون لاسو Lasso Regression
  - حذف بازگشتی ویژگی‌ها (RFE)
- با استفاده از ویژگی‌های منتخب، یک مدل رگرسیون لجستیک طراحی و آموزش دهید.
- دقت مدل را روی داده‌های آموزش و آزمون محاسبه کنید.
- ماتریس درهم‌ریختگی (ConfusionMatrix) و نمودار ROC را رسم و مقدار AUC را گزارش کنید.
- با تحلیل ضرایب مدل، مشخص کنید کدام ویژگی‌ها بیشترین اثر را بر خروجی دارند.

۴. بخش چهارم: نمایش ویژگی‌ها با استفاده از کاهش ابعاد

در این بخش باید روی داده‌ای که در اختیار دارید، کاهش ابعاد اعمال کرده و سپس آن را در یک plot scatter نمایش دهید. برای کاهش ابعاد از سه روش PCA، LDA و MLP استفاده می‌شود.

۱) PCA و LDA

- مطابق آموزش‌ها از توابع آماده استفاده کنید و تعداد ویژگی‌ها را به **دو** کاهش دهید؛ یعنی در نهایت هر نمونه دقیقاً دو ویژگی خواهد داشت.
- نگاشت دوبعدیِ به‌دست‌آمده را با یک scatter plot نمایش دهید.

۲) MLP به‌عنوان کاهش‌بُعد

- یک شبکهٔ عصبی برای طبقه‌بندی آموزش دهید (مشابه مدل‌های بخش قبل)، با این تفاوت که **لایهٔ ماقبلِ آخر** باید دو نورون داشته باشد.
- مثال معماری (برای ۱۰ ویژگی ورودی و یک مسألهٔ دودسته):

$$10\,(\text{input}) \;\rightarrow\; 5\,(\text{layer1}) \;\rightarrow\; 2\,(\text{layer2}) \;\rightarrow\; 1\,(\text{Layer3;\ 2-class})$$

- پس از رسیدن به دقت مناسب، داده‌ها را به شبکه بدهید و **خروجی‌های لایهٔ دوم** (لایهٔ ماقبلِ آخر) را استخراج کنید. چون این لایه برای هر نمونه دو عدد تولید می‌کند، می‌توانید آن را به‌عنوان نمایش دوبعدیِ ویژگی‌ها استفاده کنید.

۳) نمایش و مقایسه

- نمایش نتایج هر سه روش (PCA، LDA، MLP) روی نمودارهای scatter دوبعدی.
- عملکرد و تفکیک‌پذیری کلاس‌ها را بین این سه نگاشت دوبعدی **مقایسه** و تحلیل کنید.

۵ پرسش پنجم

بیان مسئله

موضوع پیش‌بینی قیمت مسکن با استفاده از تکنیک‌های یادگیری ماشین اهمیت زیادی دارد، زیرا قیمت مسکن تحت‌تأثیر عوامل متعدد و غیرخطی مانند موقعیت جغرافیایی، سن بنا، شرایط بازسازی و دسترسی به امکانات شهری است. روش‌های سنتی معمولاً قادر به مدل‌سازی این روابط پیچیده نیستند، اما الگوریتم‌های یادگیری ماشین با تحلیل حجم زیادی از داده‌های واقعی می‌توانند الگوهای پنهان را استخراج کرده و پیش‌بینی‌های دقیق‌تری ارائه دهند. این موضوع نه‌تنها برای تصمیم‌گیری‌های اقتصادی و سرمایه‌گذاری اهمیت دارد، بلکه در برنامه‌ریزی شهری و سیاست‌گذاری مسکن نیز نقش کلیدی ایفا می‌کند.

هدف در این بخش، پیاده‌سازی پیش‌بینی خطی و پرسپترون چندلایه به کمک زیر است:

Housing Price Prediction via Improved Machine Learning Techniques (Truong et al., 2020)

بخش اول: مطالعه مقاله

- مقاله مدل خود را برای چه شهری و چه دیتاستی (شامل چه ویژگی‌هایی) انجام داده است؟
- مقاله چه پیش‌پردازش‌هایی روی دیتاست خود انجام داده است؟
- مقاله از چه مدل‌هایی استفاده کرده است؟ (صرفاً نام ببرید)

بخش دوم: دادگان

برای این سؤال، شما از دادگانی متفاوت از مقاله استفاده خواهید کرد. لینک داده‌ها: دانلود داده‌ها

- این دادگان چند نمونه و چند ویژگی دارد؟
- نوع داده‌ی هر ویژگی چیست؟
- هر ویژگی چند مقدار منحصربه‌فرد دارد؟

بخش سوم: تحلیل اکتشافی داده‌ها (EDA)

- تحلیل اکتشافی داده‌ها چیست؟ در رابطه با آن و اهمیتش توضیح دهید.
- پس از دانلود و ذخیره داده‌ها به کمک دستور `gdown!` (در صورت استفاده از کولب)، ۵ سطر اول آن را نمایش دهید.
- این دادگان چند ویژگی عددی و چند ویژگی دسته‌ای دارد؟ نامشان را جداگانه ذخیره کنید.
- به تعداد دلخواه از ویژگی‌های دسته‌ای را با `sns.countplot` نمایش دهید.
- توزیع تعدادی از ویژگی‌های عددی را با `sns.distplot` نمایش دهید. آیا داده پرت مشاهده می‌شود؟ توضیح دهید.
- با دستور `sns.pairplot` رابطه بین ویژگی‌های مختلف را نمایش دهید.

بخش چهارم: پیش‌پردازش

- بررسی کنید آیا سطر تکراری در داده وجود دارد؟ در صورت وجود، حذف کنید.
- داده‌ها را برای وجود داده‌های گمشده بررسی و در صورت وجود، با استدلال رفع کنید.
- چند نوع از کدگذاری‌های ویژگی‌های دسته‌ای را نام ببرید و توضیح دهید هرکدام برای چه کاربردی مناسب‌تر هستند. ویژگی‌های دسته‌ای را با روش مناسب و استدلال کدگذاری کنید.
- چه روش‌هایی برای حذف داده‌های پرت وجود دارد؟ با روش مناسب داده‌های پرت را حذف نمایید.
- داده را به دو بخش آموزش و آزمون (مثلاً ۸۰/۲۰) تقسیم کنید.
- داده‌ها را نرمال‌سازی کنید.

بخش پنجم: انتخاب ویژگی

- ماتریس همبستگی ویژگی‌ها را رسم کنید. کدام ویژگی‌ها بیشترین ارتباط را دارند؟
- با روش PCA تعداد مناسبی از ویژگی‌ها را انتخاب کرده و دلیل انتخاب خود را شرح دهید.

امتیازی: دو روش VIF و RFE را برای شناسایی هم‌خطی چندگانه و انتخاب ویژگی توضیح داده و پیاده‌سازی کنید.

بخش ششم: آموزش مدل

مدل خود را برای پیش‌بینی قیمت خانه با روش‌های زیر آموزش داده و ارزیابی کنید:

- Regression Linear Multiple
- Regression Ridge
- Regression Lasso
- Regression Polynomial
- Perceptron Multi-Layer (با طراحی شبکه مناسب)

امتیازی: Regression Elastic-Net چیست؟ روابط آن را توضیح داده و پیاده‌سازی کنید.

بخش هفتم : استفاده از MLP تحت انتخاب کننده ویژگی

طبق مواردی که در درس آموخته اید حال از لایه پنهان آخر شبکه ویژگی هارو استخراج کنید . تعداد لایه و نرون ها به عهده شماست فقط درنظر داشته باشین تعداد نرون خروجی برابر۴ است چون مسئله ۴ کلاسه است . وحال از اون ویژگی ها در مدل های بخش ششم استفاده کنید . آیا نتایج بهت میشوند یا خیر ؟ تحلیل خود را کامل بیان کنید .

در انجام این تمرین حتماً به نکات زیر توجه کنید:

- موعد تحویل این تمرین، ساعت ۱۸:۰۰ روز ۳۰ مهرماه ۱۴۰۴ است.
- برای گزارش لازم است که پاسخ هر سؤال 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 گزارشِ خود لینک مخزن گیت‌هاب و گوگل‌کولبِ مربوط به مینی‌پروژهٔ خود را درج کنید. درخصوص گیت‌هاب، یک مخزنِ خصوصی درست کنید و آی‌دی‌های `MJAHMADEE`، `ParisaGhorbani`، `AliBagheriNejad` و `erfanY2AJ` را به‌عنوان Collaborator به مخزن اضافه کنید. پروژه‌های گیت‌هاب می‌بایست در انتهای ترم پابلیک شوند. در مقابل، لینک گوگل‌کولب را در حالتی که دسترسی عمومی دارد به اشتراک بگذارید. دفترچهٔ گوگل‌کولب باید به‌صورت منظم و با بخش‌بندیِ مشخص تنظیم شده باشد و خروجیِ سلول‌های اجراشده قابل مشاهده باشد. در گیت‌هاب نیز یک مخزن برای درس و یک پوشهٔ مجزا برای هر مینی‌پروژه ایجاد کنید.
(آموزش پرایوت‌کردن مخزن گیتهاب و آموزش افزودن Collaborator به مخزن گیت‌هاب)
- چرا از دفترچهٔ گوگل‌کولب؟ شما نباید به فراخوانیِ فایلی خارج از محیط نیاز داشته باشید؛ مطابق آموزش‌های ارائه‌شده ملزم هستید از دستور gdown استفاده کنید و مسیرهای فایل‌ها را طوری تنظیم کنید که صرفاً با اجرای سلول‌ها، امکان فراخوانی و خواندن فایل‌ها توسط هر کاربری وجود داشته باشد.
- در تمامی مراحلِ تعریف داده و مدل و هر جای دیگری که مطابق آموزش‌های ویدیویی و به لحاظ منطقی نیاز است، Random State را برابر با دو رقمِ آخرِ شمارهٔ دانشجوییِ خود در نظر بگیرید.
- استفاده از ابزارهای هوشمند (مانند ChatGPT) در کمک‌گرفتن برای بهبود کدها مجاز است؛ امّا لازم است تمام جزئیات مواردی که در خروجی‌های مختلفِ گزارش خود عنوان می‌کنید را به‌خوبی خوانده، درک و تحلیل کرده باشید. استفاده از این ابزارهای هوشمند در نوشتنِ گزارش و تحلیل‌ها ممنوع است.
- در جاهایی که با توجه به دو رقمِ آخر شمارهٔ دانشجوییِ خود محدود به انتخاب عددی معین یا داده‌ای خاص شده‌اید، برای تست‌های اضافه‌تر و نمایشِ بهبود در نتایجِ خود، مجاز هستید از مقادیر دیگر هم استفاده کنید.
- رعایت نکات بالا به حرفه‌ای‌تر شدن شما کمک خواهد کرد و اهمیتی معادلِ مطالبِ درسیِ فراگرفته‌شده دارد؛ بنابراین، در صورت عدم رعایت یک یا چند مورد از این نکات، از نمرهٔ تمرین شما کاسته خواهد شد.
- آی‌دیِ پرسشِ هرگونه سؤال درخصوص مینی‌پروژهٔ شمارهٔ ۱