دانشگاه صنعتی خواجه نصیرالدین طوسی تهران
مبانی سیستم‌های هوشمند (پاییز ۱۴۰۳)، مهدی علیاری شوره‌دلی

مینی‌پروژهٔ شمارهٔ دو

در انجام این مینی‌پروژه حتماً به نکات زیر توجه کنید:

- موعد تحویل این مینی‌پروژه، ساعت ۱۸:۰۰ روز ۷ دی‌ماه ۱۴۰۳ است.
- 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 مخزن گیت‌هاب را و گوگل‌کولب مربوط به مینی‌پروژهٔ خود را درج کنید. درخصوص گیت‌هاب، یک مخزن خصوصی درست کنید و آی‌دی `MJAHMADEE` را به‌عنوان Collaborator به مخزن اضافه کنید. پروژه‌های گیت‌هاب می‌بایست در انتهای ترم پابلیک شوند. درمقابل، لینک گوگل‌کولب را در حالتی که دسترسی عمومی دارد به اشتراک بگذارید. دفترچه‌کد گوگل‌کولب باید به‌صورت منظم و با بخش‌بندی مشخص تنظیم شده باشد، و خروجی سلول‌های اجراشده قابل مشاهده باشد. در گیت‌هاب هم یک مخزن برای درس و یک پوشهٔ مجزا برای هر مینی‌پروژه ایجاد کنید.
  (آموزش پرایوت‌کردن مخزن گیتهاب و آموزش افزودن Collaborator به مخزن گیت‌هاب)
- 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.
- آی‌دی پرسش هرگونه سوال درخصوص مینی‌پروژهٔ شمارهٔ دو

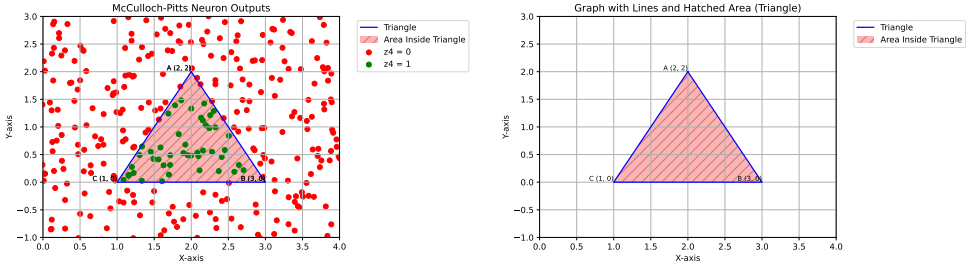
۱ پرسش یک

۱. فرض کنید در یک مسألهٔ طبقه‌بندی دوکلاسه، دو لایهٔ انتهایی شبکهٔ شما فعال‌ساز ReLU و سیگموید است. چه اتفاقی می‌افتد؟

۲. یک جایگزین برای ReLU در معادله ۱ آورده شده است. ضمن محاسبهٔ گرادیان آن، حداقل یک مزیت آن نسبت به ReLU را توضیح دهید.

$$\mathrm{ELU}(x) = \begin{cases} x & x >= 0 \\ \alpha\left(e^{x} - 1\right) & x < 0 \end{cases} \qquad (۱)$$

۳. به کمک یک نورون ساده یا پرسپترون یا نورون McCulloch-Pitts[۱] شبکه‌ای طراحی کنید که بتواند ناحیهٔ هاشورزدهٔ داخل مثلثی که در نمودار شکل ۱(آ) نشان داده شده را از سایر نواحی تفکیک کند. پس از انجام مرحلهٔ طراحی شبکه (که می‌تواند به‌صورت دستی انجام شود)، برنامه‌ای که در این دفترچه‌کد و در کلاس برای نورون McCulloch-Pitts آموخته‌اید را به‌گونه‌ای توسعه دهید که ۲۰۰۰ نقطهٔ رندوم تولید کند و آن‌ها را به‌عنوان ورودی به شبکهٔ طراحی‌شده توسط شما دهد و نقاطی که خروجی «۱» تولید می‌کنند را با رنگ سبز و نقاطی که خروجی «۰» تولید می‌کنند را با رنگ قرمز نشان دهد. خروجی تولیدشده توسط برنامهٔ شما باید به‌صورتی که در شکل ۱(ب) نشان داده شده است باشد (به محدودهٔ عددی محورهای x و y هم دقت کنید). اثر اضافه‌کردن دو تابع فعال‌ساز مختلف به فرآیند تصمیم‌گیری را هم بررسی کنید.



(آ) نمودار هاشورزدهٔ مورد سوال

(ب) خروجی مطلوب برنامه

شکل ۱: نمودارهای مربوط به بخش «۳» سوال اول و خروجی برنامه.

۲ پرسش دو

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 مسئله طبقه‌بندی است. یعنی با داشتن مجموعه داده‌ای با برچسب‌های از پیش تعیین‌شده، می‌توان مدلی ارائه کرد که بتواند کلاس مورد نظر یک نمونه 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 سوال از داده‌های جمعیت‌شناختی مانند منطقه جغرافیایی، سن و وضعیت تأهل برای پیش‌بینی الگوهای مصرف استفاده می‌کنیم. به این منظور از مجموعه داده telecust1000t بهره می‌بریم. در این مجموعه داده 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

هدف ما ساخت یک طبقه‌بند برای پیش‌بینی کلاس نمونه‌های ناشناخته است. در این پروژه از یک شبکه عصبی چندلایه (MLP) استفاده خواهیم کرد. در این راستا، پارامترهایی همچون تعداد لایه‌های مخفی، تعداد نورون‌ها، استفاده از تکنیکهایی

[۱] تشخیص اینکه با کدام روش می‌توانید این کار را انجام دهید با شماست.

مانند Dropout و L2-Regularization و انتخاب بهترین روش بهینه‌سازی، نقش مهمی در دستیابی به عملکرد بهینه مدل دارند. در این سوالات، طراحی و آموزش مدل‌های شبکه عصبی با تنظیمات مختلف بررسی شده و تأثیر تنظیمات گوناگون هایپرپارامترها، منظمکننده‌ها و روش‌های بهینه‌سازی بر روی مجموعه داده telecust1000t تحلیل می‌شود.

حال با توجه به این توضیحات به سوالات زیر پاسخ دهید.

۱. فایل csv مجموعه داده را با استفاده از کتابخانه pandas بخوانید.

۲. با استفاده از کتابخانه seaborn یا مشابه، هیت مپ مربوط به این مجموعه داده را تولید و تحلیل کرده و سپس هیستوگرام دو ویژگی که بیشترین همبستگی را با فیلد هدف دارند رسم کنید.

۳. داده ها را با استفاده از MinMaxScaler نرمالایز کنید و به سه دسته train، test، validation تقسیم کنید.

۴. دو شبکه عصبی چندلایه (MLP) برای مجموعه داده طراحی کنید که به ترتیب مدل اول شامل یک لایه مخفی و مدل دوم شامل دو لایه مخفی باشد و از بهینه‌ساز SGD برای آموزش شبکه ها استفاده کنید. سپس با توجه به شرایط زیر و با استفاده از مجموعه داده های train و validation عمل کنید:

- به بررسی تاثیر تعداد نورون ها بر عملکرد مدل بپردازید (حداقل دو حالت).
- به بررسی تأثیر اضافه‌کردن لایهٔ نرمال‌سازی دسته بپردازید.
- بهترین مدل‌های بند قبل را در نظر گرفته و تاثیر Dropout را در آنها بررسی کنید.
- بهترین مدل‌های بند قبل را در نظر گرفته و از L2-Regularization با نرخ 0.0001 در آن‌ها بهره ببرید و عملکرد مدل را ارزیابی کنید.
- بهترین مدل‌های بند قبل را با استفاده از بهینه‌ساز Adam یا RMSprop دوباره آموزش دهید و نتایج را تحلیل کنید. استفاده از بهینه‌ساز ADOPT نمرهٔ امتیازی دارد!

۵. ارزیابی بهترین مدل‌های بخش قبل را برروی داده های test انجام داده و تحلیل کنید. سپس 10 نمونه را از داده‌ای test بصورت تصادفی انتخاب کرده و خروجی های شبکه ها را با مقادیر واقعی مربوطه گزارش کنید.

۶. بهترین مدل‌های به‌دست‌آمده از دو شبکهٔ طراحی‌شده در بندهای قبل را به یکی از روش‌هایی که آموخته‌اید ترکیب کنید و نتایج تست را برای این حالت هم محاسبه کرده و نمایش دهید. آیا نتایج بهبود پیدا کردند؟ چرا؟

۳ پرسش سه

به این دفترچه‌کد مراجعه کنید و با اجرای سلول اول، ۵ دادهٔ تصویری مربوط به حروف الفبای فارسی که در شکل ۲ نشان داده شده است را دریافت کنید و سپس به سوالات زیر پاسخ دهید. دقت داشته باشید که در هر مرحله ارائهٔ توضیحات متنی و دیداری مناسب لازم است. مثلاً می‌توانید ورودی نویزی و خروجی‌پیش‌بینی شده را در یک تصویر در کنار هم قرار دهید.

شکل ۲: نمونهٔ داده‌ها.

۱. دو تابع پایتونی در سلول‌های دوم و سوم این دفترچه‌کد نوشته شده‌اند. اولین تابع تصویر را در ورودی خود دریافت و به‌صورت نمایش باینری درمی‌آورد و دومین تابع با افزودن نویز به داده‌ها، داده‌های جدید نویزی تولید می‌کند. در مورد نحوهٔ عملکرد هریک از این توابع توضیح دهید. هم‌چنین، می‌توانید این دستورات را به‌صورتی بهتر و کارآمدتر بازنویسی کنید.

۲. یک شبکهٔ عصبی (همینگ یا هاپفیلد) طراحی کنید که با اِعمال ورودی دارای میزان مشخصی نویز برای هر یک از داده‌ها، خروجی متناسب با آن دادهٔ نویزی را بیابد. میزان نویز را تا حدی که شبکهٔ شما ناموفق عمل کند افزایش دهید و نتایج را مقایسه و تحلیل کنید.

۳. با الهام‌گرفتن از تابع نوشته‌شده برای تولید داده‌های نویزی، یک تابع بنویسید که از داده‌های ورودی، خروجی‌های دارای Missing Point تولید کند. سپس عملکرد شبکهٔ خود را با مقدار مشخصی Missing Point آزمایش و تحلیل کنید. اگر میزان Missing Point از چه حدی بیش‌تر شود عملکرد شبکهٔ طراحی‌شدهٔ شما دچار اختلال می‌شود؟ راه‌حل چیست؟ (راهنمایی: نمونه‌دادهٔ دارای Missing Point در شکل ۳ نشان داده شده است.)

شکل ۳: نمونهٔ دادهٔ دارای Missing Point.

۴ پرسش چهار

در ان سوال به ساخت و آموزش دو مدل رگرسیون با یک لایه پنهان برای پیش‌بینی قیمت خانه‌ها با استفاده از دیتاست California Housing می‌پردازیم. هدف مقایسه عملکرد دو مدل مختلف است: یکی با لایه RBF و دیگری با لایه‌های کاملاً متصل (Dense). این دیتاست شامل ویژگی‌هایی مانند درآمد میانه، سن خانه‌ها و تعداد اتاق‌ها در هر خانه است و متغیر هدف قیمت میانه خانه‌ها است.

۱. یک شبکه عصبی پیاده‌سازی کنید که در آن لایه‌ی RBF به عنوان لایه پنهان قرار باشد.
۲. از Mean Squared Error (MSE) به عنوان تابع از دست دادن و از Adam به عنوان بهینه‌ساز برای آموزش استفاده کنید.
۳.
۴. یک شبکه عصبی ساده با استفاده از لایه‌های کاملاً متصل (Dense) پیاده‌سازی کنید.
۵. از Mean Squared Error (MSE) به عنوان تابع از دست دادن و از Adam به عنوان بهینه‌ساز برای آموزش استفاده کنید.
۶. پس از آموزش هر دو مدل، آنها را روی مجموعه تست ارزیابی کرده و میزان از loss آنها را نمایش دهید.
۷. عملکرد دو مدل را مقایسه کنید. کدام مدل بهتر عمل کرده‌است؟ پاسخ خود را توجیه کنید.

راهنمایی:

- از دیتاست California Housing موجود در sklearn.datasets استفاده کنید.
- از StandardScaler برای نرمال‌سازی ویژگی‌های ورودی استفاده کنید.
- دیتاست را به مجموعه‌های آموزشی و آزمایشی تقسیم کنید.