



تمرین شماره ۲

- قبل از شروع تمرین، فایل مربوط به قوانین حل و تحویل تمرین‌ها را مطالعه کنید.
- سؤالات و مشکلات خود را درباره این تمرین می‌توانید در گروه تلگرامی درس یا با طراحان این تمرین مطرح کنید.
- نوشتن گزارش کامل و تفسیر نتایج اجباری است. جزئیاتی مانند روش‌های مورد استفاده، تاثیر هر روش در نتیجه نهایی و بهبود حاصل شده به همراه ارائه معیارهای ارزیابی خواسته شده در گزارش ضروری است. با هر تغییر و هر بهبود، تغییر مقادیر معیارهای ارزیابی نیز ذکر شود. ضمناً برای گزارش سؤالات عملی، نیاز به فایل دیگری نیست و در همان محیط کدزنی، در قالب بلاک‌های مارک‌داون و نوشتار، توضیحات مورد نیاز را قرار دهید.
- پاسخ سؤالات تئوری را در یک فایل پی‌دی‌اف^۱ قرار دهید.
- برای سؤالات عملی می‌توانید از کتابخانه‌هایی نظیر pandas، numpy و matplotlib برای کار با داده‌گان و تولید نمودارها استفاده کنید. همچنین برای تعیین معیارهای ارزیابی گفته شده، می‌توانید از کتابخانه‌ی sklearn نیز استفاده نمایید.
- برای دریافت داده‌گان سؤالات عملی، از راهنمای دریافت داده‌گان که در سامانه درس‌افزار قرار داده شده است کمک بگیرید و داده‌گان را به صورت دستی دانلود نکنید.
- مهلت ارسال پاسخ‌ها: جمعه ۲ آذر ساعت ۲۳:۵۹
- طراحان این تمرین: [امیررضا افتخاری](#) – [آرمان غفاریا](#)

۳۵ نمره

سؤالات تئوری

سؤال ۱ (۱۰ نمره)

چگونه میتوان از مدل‌های دسته‌بندی دودویی^۲ برای مسائل دسته‌بندی چندگانه استفاده کرد؟ می‌توانید پیاده‌سازی‌ها و تفاوت‌های نظری بین این استراتژی‌ها را توضیح دهید؟ این روش‌ها چگونه با سناریوهایی که در آن مرزهای تفاوت کلاس‌ها به طور واضح تعریف نشده‌اند (برای مثال هم‌پوشانی^۳ دارند)، برخورد می‌کنند؟

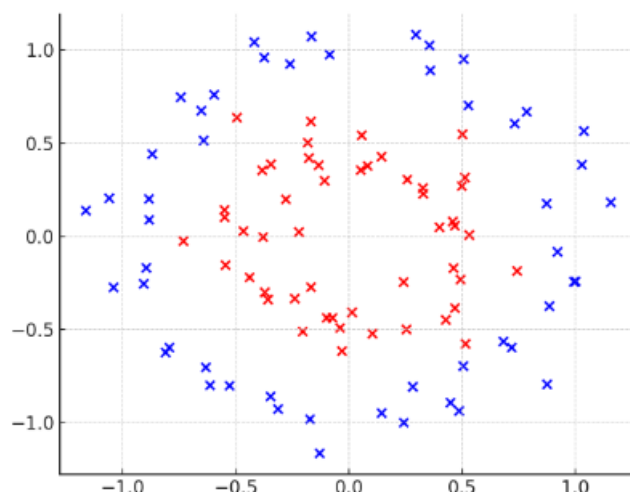
^۱ PDF

^۲ Binary Classification

^۳ Overlap

سؤال ۲ (۵ نمره)

داده‌های زیر را در نظر بگیرید. به نظرتان می‌توان از الگوریتم‌های طبقه‌بندی خطی برای طبقه‌بندی آنها استفاده کرد؟ اگر پاسخ مثبت است چگونه انجام این کار را توضیح دهید و اگر پاسخ منفی است، علت را بگویید و سپس سعی کنید با اعمال تغییراتی در این نوع الگوریتم‌ها، راهکاری برای طبقه‌بندی این داده‌ها، ارائه دهید.



سؤال ۳ (۱۰ نمره)

در زمینه‌ی طبقه‌بندی خطی، توضیح دهید که چگونه انتخاب تابع خسارت^۴ بر عملکرد مدل در مواجهه شدن با داده‌هایی که برایش جدید است، تأثیر می‌گذارد؟ مثال‌هایی از توابع خسارت مختلف ارائه دهید و تأثیر آن‌ها را بر رفتار همگرایی یک مدل طبقه‌بندی خطی، به ویژه در سناریوهایی شامل داده‌های نویزی یا پرت توضیح دهید.

سؤال ۴ (۱۰ نمره)

در زمینه رگرسیون خطی، چگونه استفاده از نگاشت‌های ویژگی چندجمله‌ای^۵، ساختار بنیادین مسئله را هنگام استفاده از تکنیک‌های مختلف حل مانند گرادیان نزولی و راه‌حل مستقیم تغییر می‌دهد؟ بررسی کنید که چگونه این تبدیلات ممکن است توانایی عمومی‌سازی^۶ مدل را تحت تأثیر قرار دهند به ویژه زمانی که بر روی مجموعه داده‌ها با الگوهای غیرخطی اعمال می‌شوند؟

^۴ Loss Function

^۵ Polynomial Feature Maps

^۶ Generalization

سؤال ۱ (۳۵ نمره)

با توجه به اینکه شرکت های بیمه اکثرا با مشکل تخمین و تعیین حق بیمه مناسب برای هر فرد مواجه هستند ، دقت در تعیین حق بیمه می تواند به مسائل مالی شرکت بیمه کمک زیادی کند.

در این سوال یک مجموعه داده [data1.csv](#) در اختیارتان قرار گرفته که مجموعه اطلاعات بیمه شدگان درمانی است. این اطلاعات شامل سن، جنسیت، شاخص توده بدنی^۷، تعداد فرزندان، عادات سیگار کشیدن و منطقه سکونت است. همچنین در ستون آخر آن هزینه های پزشکی قرار داده شده است.

میخواهیم مدلی طراحی کنیم که به کمک این داده ها بتواند هزینه های پزشکی را پیش بینی کند و باعث شود حق بیمه هر فرد، دقیق تر مشخص شود.

الف) داده ها را به دو بخش آموزش و تست تقسیم کنید و نسبت تعداد داده های تست به آموزش را ذکر کنید و علت این انتخاب خود را نیز بنویسید (راهنمایی: برای انتخاب درصد مناسب، از تابع خسارت استفاده کنید). سپس پیش از پیاده سازی مدل با توجه به هدف سوال، پیش پردازش های لازم را انجام دهید. (۵ نمره)

ب) یک مدل رگرسیون خطی را به کمک مدل های آماده موجود، پیاده سازی کنید و خطا را با معیارهای MAE، MSE و RMSE گزارش کنید. تحلیل کوتاهی درباره معنای هر یک از این معیارها ارائه دهید و یکی از معیارها را که فکر می کنید برای این مسئله مناسب تر باشد انتخاب کنید. (۱۵ نمره)

ج) در این قسمت، کدی پیاده سازی کنید که خط جداکننده ی رگرسیون را در مرحله های مختلف و مقدار خطا^۸ (از معیاری که در قسمت ب انتخاب کرده اید استفاده کنید) در آن مرحله را در طی فرایند آموزش نشان دهد و رابطه ی بین تغییرات خط رگرسیون و مقدار خطا را در هر تکرار تحلیل کنید (برای مثال مقایسه ای از سرعت کاهش خطا نسبت به تغییرات خط رگرسیون). (۱۰ نمره)

د) روند تغییرات وزن های مدل خود را در طول فرایند آموزش، به صورت عددی نمایش دهید و همچنین نشان دهید که به تدریج این وزن ها به یک مقدار ثابت نزدیک می شوند. معنای این نزدیک شدن وزن ها به یک عدد ثابت چیست؟ (۵ نمره)

⁷ Body Mass Index (BMI)

⁸ Loss Value

سؤال ۲ (۳۰ نمره)

یکی از هتل‌های معروف تهران، قصد دارد تا نوشیدنی‌هایی با کیفیت بالا جهت سرو برای مسافران خارجی، تهیه کند. جهت تحقق این امر، اطلاعات نوشیدنی‌های مختلف را به همراه امتیاز کیفیت آنها، در مجموعه دادگان [data2.csv](#) جمع‌آوری کرده است. همانطور که در این دادگان مشاهده می‌کنید، بازه‌ی امتیاز کیفیت هر نوشیدنی در ستون آخر، از صفر الی ده می‌باشد و اطلاعات هر نوشیدنی مانند اسیدیته ثابت و فرار، اسید سیتریک و... در باقی ستون‌ها قرار دارند. اکنون قصد داریم مدلی طراحی کنیم تا بتواند در ازای گرفتن اطلاعات هر نوشیدنی، امتیاز کیفیت آن را برای ما تعیین کند (موارد خواسته‌شده در این سوال اعم از تابع خسارت، تابع فعالساز و مدل گفته شده را از پایه پیاده‌سازی کنید. البته همانطور که در توضیحات تمرین گفته شده، برای معیارهای ارزیابی می‌توانید از کتابخانه‌ی sklearn استفاده کنید).

الف) داده‌ها را به دو بخش آموزش و تست تقسیم کنید و میزان نسبت هر کدام را نیز در گزارش قرار دهید. تابع خسارت را **Cross-Entropy Loss** در نظر بگیرید. سپس پیش‌پردازش‌های لازم را با توجه به نوع داده‌ها انجام دهید. ضمناً اگر داده‌ی بسیار پرت و یا از دست رفته^۹ در هر یک از ستون‌ها وجود دارد، رویکرد مناسبی برای آنها در پیش بگیرید. (۵ نمره)

ب) اکنون یک مدل رگرسیون لجستیک خطی را پیاده‌سازی کنید. تابع فعالساز را نیز **Softmax** در نظر بگیرید. سپس به کمک دادگان تست، مدل ساخته شده‌ی خود را مورد ارزیابی قرار دهید. معیارهای **Accuracy**، **F1-Score** و **Confusion Matrix** را بررسی و در گزارش بیاورید. به نظرتان مدل شما در چه کلاس‌هایی ضعیف عمل خواهد کرد؟ چرا؟ (۱۵ نمره)

ج) آیا تابع **Softmax** توانسته است نسبت احتمال هر کدام از کلاس‌ها را تا حد خوبی درست تعیین کند؟ به نظرتان چه تغییری میتوان در آن اعمال کرد تا نتایج بدست آمده را بهتر نمود؟ سعی کنید این تابع را در جهت بهبود نتایج، مورد تغییر قرار دهید و توضیحات تغییرات را به همراه مقایسه‌ی نتایج جدید بدست آمده با نتایج قبلی، در گزارش قرار دهید. (۱۰ نمره)

سالم و موفق باشید.

⁹ Miss