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

مینی‌پروژهٔ شمارهٔ یک

در انجام این مینی‌پروژه حتماً به نکات زیر توجه کنید:

- موعد تحویل این مینی‌پروژه، ساعت ۱۸:۰۰ روز جمعه ۲ آذرماه 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 یک

۱ پرسش اول

برای حل این سؤال از این مجموعه داده [۱] استفاده می‌کنیم. فایل داده را دانلود کرده و آن را در محیط پایتون بارگذاری کنید. به سوالات زیر پاسخ دهید.

[۱] dataset

۱.۱

- درباره این مجموعه داده به صورت خلاصه توضیح دهید.
- ویژگی‌های [۲] موجود در این مجموعه داده را نام ببرید.
- چه تعداد نمونه [۳] در این مجموعه داده موجود است؟

۲.۱

با استفاده از تابع `sns.pairplot` پخش [۴] داده را نمایش دهید. (در صورت زیاد بودن تعداد ویژگی‌ها، به دلخواه چهار یا پنج ویژگی را انتخاب کرده و پخش آن‌ها را نمایش دهید)

۳.۱

همبستگی [۵] موجود میان ویژگی‌های مختلف را به صورت نقشه حرارتی [۶] نشان دهید. (برای حداقل دو ویژگی طبقه‌بندی‌شده [۷] و دو ویژگی پیوسته [۸])

۴.۱

آیا در میان داده‌های موجود، داده Nan وجود دارد؟ در صورت وجود Nan در هر یک از نمونه‌ها، آن را حذف کنید.

۵.۱

ویژگی Atrrition Flag دارای چند کلاس است.

- نام کلاس‌های موجود در این ویژگی چیست؟
- پخش داده موجود در این ویژگی را به صورت یک pie plot نمایش دهید.
- ویژگی Atrrition Flag که می‌خواهیم مدلی برای پیش‌بینی آن بسازیم، دارای عدم تعادل [۹] است. تحقیق کنید که آیا این عدم تعادل در عملکرد مدل نهایی تأثیر دارد یا نه. توضیح دهید.
- چه راهکارهایی برای اصلاح این مشکل وجود دارد؟ تحقیق کنید.
- اگر بخواهیم از یک الگوریتم برای متعادل کردن مجموعه داده‌استفاده کنیم، باید این کار را قبل از تقسیم‌بندی داده به بخش‌های آموزش و آزمون انجام دهیم یا پس از آن؟ توضیح دهید.

۶.۱

داده‌های موجود در ویژگی Atrrition Flag را به عنوان خروجی انتخاب کرده و بقیه ویژگی‌ها را به عنوان داده ورودی درنظر بگیرید. داده‌ها را با نسبت دلخواه به سه بخش آموزش، اعتبارسنجی و آزمون تقسیم کنید.سپس یک الگوریتم طبقه‌بندی [۱۰] از کتابخانه `scikit learn` انتخاب کنید. سپس مدل را به اشکال زیر آموزش دهید:(دقت کنید که در تمامی مراحلی که مدل خود را آموزش می‌دهید باید گزارش طبقه‌بندی [۱۱] و ماتریس درهم‌ریختگی [۱۲] را با استفاده از توابع موجود در کتابخانه `sklearn` و یا هر کتابخانه دیگری، محاسبه کرده و نتایج را برای داده آموزش و اعتبار سنجی گزارش نمایید.)

---

[۲] features
[۳] sample
[۴] distribution
[۵] correlation
[۶] heatmap
[۷] categorical
[۸] continuous
[۹] unbalancing
[۱۰] classification
[۱۱] classification report
[۱۲] confusion matrix

۱. بدون متعادل کردن داده‌ها، مدل خود را آموزش دهید.

۲. یک الگوریتم متعادل سازی مجموعه داده را معرفی کرده و پس از متعادل کردن داده، مدل خود را آموزش دهید. (**راهنمایی:** اگر داده خود را متعادل کردید و مدل تنها یک کلاس را پیش‌بینی می‌کرد، بعد از متعادل کردن داده، آن را بُر بزنید [۱۳] .)

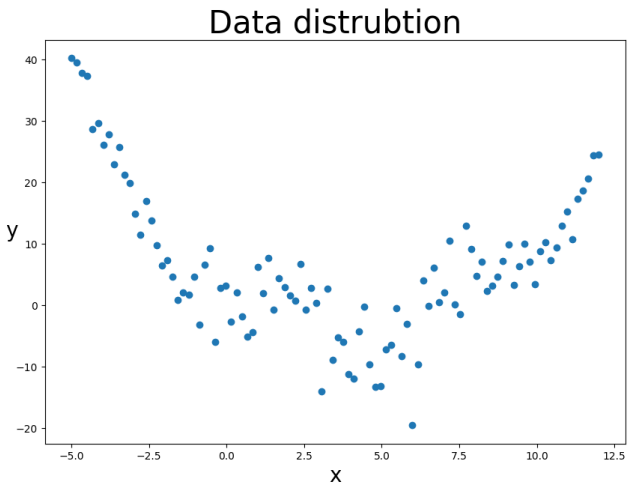
عملکرد مدل‌هایی که آموزش دادید را با هم مقایسه کرده و تحلیل کنید.

امتیازی

بخش دوم پرسش اول را تکرار کنید با این تفاوت که این بار پخش داده را با توجه به کلاس‌های مختلف موجود در ویژگی Attrition_Flag نمایش دهید.

۲ پرسش دوم

داده شکل ۱ را درنظر بگیرید. (داده نشان داده شده در این لینک موجود است.)



شکل ۱: پخش داده

- می‌توانید با استفاده از دستور `np.load()` داده‌ها را بخوانید.
- در تمامی مراحلی که مدل آموزش می‌دهید، نمودار خطا [۱۴] برای داده‌های آموزش [۱۵] و آزمون [۱۶] را نمایش دهید.

۱.۲

مجموعه داده را به بخش‌های آموزش و آزمون تقسیم کنید و داده مربوط به هر یک از مجموعه داده ها را بر روی یک نمودار نمایش دهید. مشخص کنید که کدام داده برای چه مجموعه داده‌ای است.

---

[۱۳] shuffle
[۱۴] error graph
[۱۵] train
[۱۶] test

۲.۲

سه معیار برای سنجش عملکرد مدل‌های رگرسیون معرفی کنید و هر یک را توضیح دهید. در مراحلی که مدل خود را آموزش می‌دهید، از این معیارها برای سنجش عملکرد مدل‌های خود استفاده نمایید. (برای محاسبه عملکرد مدل، استفاده از توابع آماده بلامانع است)

۳.۲

یک مدل رگرسیون خطی [۱۷] درجه اول (بدون استفاده از توابع آماده) روی داده مورد نظر آموزش دهید. به نظر شما آیا یک مدل خطی درجه اول می‌تواند به خوبی داده مورد نظر را تخمین بزند؟ توضیح دهید.

۴.۲

در این بخش، تعداد دور حلقه آموزش (Iteration) را ثابت در نظر بگیرید. در ابتدا برای آموزش مدل از تنها یک داده آموزش استفاده کرده، مدل را آموزش داده و سپس مقادیر خطا برای داده آموزش و آزمون را ذخیره نمایید. در مرحله بعد یک داده به داده آموزش اضافه کرده و روند قبلی را تکرار کنید تا این که در مرحله آخر با استفاده از تمامی داده‌های آموزش مدل را آموزش دهید. نمودار خطا برای داده آزمون و آموزش را بر حسب تعداد داده آموزش رسم کنید. توضیح دهید با افزایش داده آموزش چه اتفاقی برای خطاهای آزمون و آموزش می‌افتد.

۵.۲

با توجه به نتایج بخش قبل به سوال زیر پاسخ دهید.
برای انجام فعالیتی، خطای انسان برابر ۱ است. یک مدل یادگیری ماشین برای انجام همین فعالیت آموزش داده شده است که خطا آموزش آن برابر ۱۰ است. اگر برای آموزش این مدل از داده بیشتری استفاده کنیم، آیا می‌توانیم خطا مدل را به اندازه خطا انسان کاهش دهیم؟ توضیح دهید.

۶.۲

به مدل رگرسیون خطی که در بخش قبل آموزش دادید، مرحله به مرحله یک جمله با درجه دلخواه اضافه کنید. (مثلا در مرحله اول x^2 را به مدل اضافه کنید). این کار را حداقل برای ۵ جمله تکرار کنید.

- نمودار خطا بر حسب تعداد جملات چندجمله‌ای را نمایش دهید.
- آیا با افزایش تعداد جمله‌های مدل، خطای آزمون همواره کاهش میاید؟ توضیح دهید.

۷.۲

از میان الگوریتم‌های رگرسیون موجود در کتابخانه `scikit learn`، به دلخواه ۳ الگوریتم را انتخاب کرده و به صورت خلاصه آن‌ها را توضیح دهید. سپس از این سه الگوریتم برای آموزش مدل استفاده کرده و نتایج آن‌ها را با هم مقایسه کنید.

امتیازی

درباره reguralization تحقیق کنید و مدل چند جمله‌ای خود را با استفاده از reguralization دوباره آموزش دهید. (بدون استفاده از توابع آماده)

[۱۷] linear regression