مینی‌پروژهٔ شمارهٔ یک

۱ پیش بینی بقا در کشتی تایتانیک[۱] با استفاده از رگرسیون لجستیک[۲]

مقدمه[۳]

کشتی تایتانیک در اولین سفر خود در تاریخ ۱۵ آوریل ۱۹۱۲ با برخورد به یک کوه یخی غرق شد. این حادثه، یکی از مرگبارترین فجایع دریایی تاریخ محسوب می‌شود که جان بیش از ۱۵۰۰ نفر را گرفت. بررسی داده‌های مسافران این کشتی نشان می‌دهد که احتمال زنده ماندن افراد به عوامل مختلفی مانند سن، جنسیت و کلاس سفر بستگی دارد. تحلیل این داده‌ها می‌تواند به ما کمک کند تا الگوهای مؤثر در بقا را شناسایی کنیم. یکی از روش‌های پرکاربرد در این زمینه، رگرسیون لجستیک است که به دلیل قابلیت آن در دسته‌بندی داده‌های دودویی[۴]، روش مناسبی برای پیش‌بینی بقا یا فوت مسافران محسوب می‌شود. در این پروژه، از همین روش برای تخمین احتمال زنده ماندن هر مسافر بر اساس ویژگی‌های او استفاده خواهیم کرد.

بخش اول: آشنایی با مجموعه داده

مجموعه داده تایتانیک[۵] یکی از معروف‌ترین مجموعه داده‌ها در حوزه یادگیری ماشین[۶] است که اطلاعات مربوط به مسافران کشتی تایتانیک را شامل می‌شود. در این پروژه، قصد داریم با تحلیل ویژگی‌های مسافران، روابط بین متغیرها را بررسی کنیم و مدلی برای پیش‌بینی بقای مسافران بسازیم. این مجموعه داده شامل ویژگی‌های مختلفی از مسافران تایتانیک است که در عکس زیر قابل مشاهده است:

survival	Survival (0 = No; 1 = Yes)
pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

برای استفاده از این مجموعه داده‌ها، روش‌های مختلفی وجود دارد:

- می‌توانید آن را از طریق این لینک دریافت کنید،
- یا با استفاده از API Kaggle، مجموعه داده را مستقیما در پروژه خود آپلود و استفاده نمایید.

[۱] Titanic
[۲] logistic regression
[۳] برای آشنایی بیشتر با logistic regression می‌توانید به این لینک مراجعه کنید.
[۴] binary
[۵] Titanic Dataset
[۶] machine learning

۱.۱ بررسی مجموعه داده:

وقتی یک پروژه یادگیری را شروع می‌کنیم، داده‌هایی که در ابتدا با آنها شروع می‌کنیم، داده‌های خام[۱] هستند. لذا نیاز داریم که آن‌ها را تجزیه و تحلیل کنیم و یک دید کلی نسبت به داده‌ها به دست آوریم و با ویژگی‌های آن‌ها آشنا شویم. به فاز اولیه تجزیه و تحلیل داده‌ها اصطلاحا EDA[۲] می‌گویند. برای اجرا این فاز گام های زیر را انجام دهید:

- ساختار کلی داده‌ها را بدست آورید. (برای این کار می‌توانید از متدهای `()info.` و `()describe.` استفاده کنید.) مبهم
- نمودار همبستگی[۳] را رسم کنید و بررسی کنید که کدام ویژگی‌ها ارتباط قوی‌تری با متغیر هدف (`Survived`) دارند.
- نمودارهای پراکندگی[۴] و هگزبین[۵] را برای بررسی رابطه بین ویژگی‌های عددی و متغیر هدف رسم کنید. توضیح دهید که این نمودارها چه اطلاعاتی را نمایش می‌دهند و چگونه می‌توان از آن‌ها برای تحلیل داده استفاده کرد.
- با استفاده از کتابخانه `Plotly` نمودار پراکندگی برای نمایش توزیع بازماندگان بر اساس سن و کرایه پرداختی رسم کنید. (محور افقی، سن و محور عمودی کرایه پرداختی است.) آیا افرادی که کرایه بیشتری پرداخته‌اند، شانس بقای بیشتری داشته‌اند؟ با استفاده از (()countplot. – Seaborn) بررسی کنید که چند درصد از مردان و چند درصد از زنان زنده ماندند. توزیع بازماندگان را بر اساس جنسیت مقایسه کنید (مثلاً چند درصد از زنان و چند درصد از مردان زنده ماندند).

۲.۱ تحلیل آماری مجموعه داده:

- تعداد اعضای خانواده‌ای که همراه هر مسافر سفر کرده‌اند را محاسبه کنید. بررسی کنید که آیا افراد دارای خانواده بزرگ‌تر یا کوچک‌تر، شانس بقای بیشتری داشته‌اند؟
- بررسی کنید که آیا یک مسافر به‌تنهایی سفر کرده است یا همراه با خانواده. تحلیل کنید که آیا تنهایی تأثیر منفی یا مثبتی بر نرخ بقا داشته است.
- مسافران را به گروه‌های سنی مختلف (مانند کودک، نوجوان، جوان، بزرگسال و سالمند) تقسیم کنید. بررسی کنید که آیا گروه‌های سنی مختلف، شانس بقای متفاوتی داشته‌اند.

بخش دوم: پیش پردازش مجموعه داده

مهم‌ترین فاز هر پروژه یادگیری ماشین، فاز پیش پردازش است. در این فاز فرمت داده‌ها را تغییر می‌دهیم، آن‌ها را اصلاح و خلاصه می کنیم، تا بتوانیم برای آموزش یک مدل یادگیری ماشین از آن استفاده کنیم. چرا که در دنیای واقعی، اطلاعات جمع آوری شده به راحتی کنترل نمی‌شوند و در نتیجه مقادیر خارج از محدوده، ناممکن، از دست رفته و به طور کلی گمراه کننده برای آموزش مدل در مجموعه داده‌ها وجود دارند. این بخش باعث می‌شود مدل کارا تری بتوانیم داشته باشیم و سرعت یادگیری بالاتر می رود.

- ممکن است برخی ستون های جدول دارای داده های از دست رفته باشند، تعداد و نسبت این داده‌ها را به دست بیاورید. روش های پر کردن داده‌های از دست رفته[۶] را توضیح دهید و حداقل سه روش را پیاده سازی کنید. دلیل استفاده از هر روش را مختصرا توضیح دهید.
- آیا امکان حذف برخی ستون ها وجود دارد؟ چرا؟ در صورتی که این امکان وجود دارد با ذکر دلیل ستون های لازم را حذف کنید.

---

[۱] raw data
[۲] exploratory data analysis
[۳] correlation matrix
[۴] scatter
[۵] hexbin
[۶] missing value

- کدام ویژگی‌ها را عددی و کدام‌ها را دسته‌ای می‌گویند؟ تفاوت این دو نوع از ویژگی‌ها در چیست؟ ویژگی‌های عددی[۱] و دسته‌ای[۲] را در این مجموعه دادگان مشخص کنید. برای ویژگی های دسته‌ای، که معمولا بصورت یک `string` یا `object` در مجموعه داده ذخیره شده‌اند، در آموزش مدل چه پیش‌پردازش هایی مفید است؟ این موارد را در دیتاست اعمال کنید. در ویژگی های عددی نرمالایز کردن[۳] و استانداردسازی[۴] به چه منظور انجام می‌شود؟ تفاوت این دو روش در چیست؟ آیا در این پروژه نیاز به انجام این کار است؟

بخش سوم: انتخاب ویژگی[۵]، آموزش[۶] و ارزیابی[۷]

برای بهبود عملکرد مدل، ویژگی‌های جدیدی از داده‌های موجود استخراج می‌کنیم که می‌توانند به تشخیص الگوهای مؤثر بر بقا کمک کنند. همچنین در این مرحله، باید ویژگی‌های مهم را برای مدل یادگیری ماشین انتخاب کنیم تا دقت مدل افزایش یابد و از ویژگی‌های غیر ضروری اجتناب شود. پس از آموزش مدل‌های رگرسیون لجستیک، ضرایب مدل تحلیل شده و نسبت شانس[۸]برای هر ویژگی محاسبه می‌شود. این تحلیل به ما کمک می‌کند تا بفهمیم کدام ویژگی‌ها بیشترین تاثیر را بر احتمال بقا دارند.

- در مورد روش‌های مختلف انتخاب ویژگی تحقیق کنید. به طور خاص روش رگرسیون لاسو[۹] و حذف ویژگی‌ها به صورت بازگشتی[۱۰] را پیاده سازی کنید(از توابع آماده موجود در کتابخانه‌ها می‌توانید استفاده کنید.) و ویژگی‌های انتخاب شده توسط دو روش را مقایسه کنید. همچنین ویژگی‌های مورد نظر را انتخاب کنید(این ویژگی‌ها می‌تواند اجتماع ویژگی‌های دو روش باشند)

حال میخواهیم به کمک ویژگی‌های بدست آمده مدل‌های یادگیری ماشین برای پیش‌بینی بقای مسافران کشتی تایتانیک بسازیم. دو نوع رگرسیون لجستیک دودویی و رگرسیون لجستیک چندکلاسه را بررسی خواهیم کرد.

- مدل رگرسیون لجستیک دودویی را با ویژگی‌های منتخب آموزش دهید. بررسی کنید که دقت مدل روی داده‌های آموزشی و تستی چقدر است؟ ماتریس درهم‌ریختگی[۱۱]را رسم کنید و تحلیل کنید که مدل کدام کلاس‌ها را بیشتر اشتباه پیش‌بینی کرده است؟ نمودار ROC[۱۲] و مقدار AUC[۱۳] را رسم کنید تا عملکرد مدل را ارزیابی کنید.
- دو مدل رگرسیون لجستیک چندکلاسه[۱۴] و یکی در مقابل همه[۱۵] را پیاده‌سازی کنید.(از توابع آماده موجود در کتابخانه `sklearn` میتوانید استفاده کنید.) برای این کار مسئله را اینگونه در نظر بگیرید:
  - $0 \rightarrow$ شانس بقا کم Low Chance
  - $1 \rightarrow$ شانس بقا متوسط Medium Chance
  - $2 \rightarrow$ شانس بقا زیاد High Chance

  داده‌های بقا را بر اساس این سه کلاس دسته‌بندی کنید و بررسی کنید که چه تعداد نمونه در هر کلاس وجود دارد؟ دقت مدل را روی داده‌های آزمایش[۱۶] بررسی کنید.

---

[۱] numerical
[۲] categorical
[۳] normalization
[۴] standardization
[۵] feature selection
[۶] train
[۷] evaluation
[۸] odds ratio
[۹] lasso regression
[۱۰] recursive feature elimination
[۱۱] confusion matrix
[۱۲] receiver operating curve
[۱۳] area under the ROC
[۱۴] multinomial logistic regression
[۱۵] one-vs-rest logistic regression
[۱۶] test

- ضرایب رگرسیون لجستیک دودویی و چندکلاسه را محاسبه کنید. لیستی از ویژگی‌های مدل همراه با ضرایب تخمین‌زده‌شده ارائه دهید. ویژگی‌هایی که ضرایب مثبت دارند، چگونه بر افزایش احتمال بقا تأثیر می‌گذارند؟ ویژگی‌هایی که ضرایب منفی دارند، چگونه احتمال بقا را کاهش می‌دهند؟ نسبت شانس را برای هر ویژگی محاسبه کنید. مقدار نسبت شانس را برای هر ویژگی تفسیر کنید.

۲ پرسش دوم

۱.۲ دریافت دادگان

به این پیوند مراجعه کرده و دادگان[۱] قرار داده شده شامل ۲۰۰۰ داده را در فرمت csv. دریافت کنید.

۱.۱.۲

با استفاده از کتابخانه `pandas` در پایتون دادگان دریافت شده را به یک قاب داده[۲] تبدیل کنید و با متد `()head` آن را در ترمینال چاپ کنید.

۲.۱.۲

در تبدیل دادگان به قاب‌داده چه تغییری رخ داده است؟

۳.۱.۲

با استفاده از متد `numpy.reshape()` شاخص ستون‌های قاب‌داده را به یک آرایه ستونی تبدیل کنید.

۴.۱.۲

سپس با استفاده از کتابخانه `matplotlib` هر ردیف از داده‌ها را با رنگ‌های مختلف رسم کنید.

۵.۱.۲

به نظر شما برای پیشبینی این داده‌ها، چه مشکلی می‌تواند وجود داشته باشد؟

برای داده‌های ردیف اول قاب‌داده به سوالات ۲ تا ۵ پاسخ دهید:

۲.۲ پاکسازی و پیش‌پردازش داده‌ها

۱.۲.۲

در مورد روش‌های پاکسازی داده‌ها[۳] تحقیق کنید.

- اهمیت و نقش پاکسازی در پیش‌پردازش داده‌ها[۴] را توضیح دهید.
- داده‌های خوانده شده را حداقل به دو روش پاکسازی کنید. در پاکسازی داده‌های پرت می‌توانید آن‌ها را حذف و یا با میانگین همسایه‌ها جایگزین کنید.
- داده‌های پاکسازی شده را همراه با داده‌های اصلی رسم و نتیجه را مقایسه و گزارش کنید.

---

[۱] dataset
[۲] dataframe
[۳] data cleaning
[۴] data preprocessing

۳.۲ آموزش مدل‌های رگرسیون

۱.۳.۲

رگرسیون خطی from scratch

- مدل، تابع هزینه و الگوریتم یادگیری را کدنویسی کنید تا داده‌ها را پیشبینی کند.
- در مورد الگوریتم یادگیری خود توضیح دهید و آن را با حداقل دو تابع هزینه[۱] متفاوت امتحان کنید.
- نتایج را رسم و با یکدیگر مقایسه کنید.

۲.۳.۲

آموزش رگرسیون خطی

- با استفاده از `LinearRegression()` درکتابخانه `sklearn`، مدل‌های یادگیری برای داده‌های پاکسازی شده و داده‌های اصلی آموزش دهید.
- نتایج مدل‌های آموزش داده شده را با رنگ‌های مختلف در فضای دو بعدی رسم کنید.
- پارامترهای هر مدل را در خروجی چاپ و با یکدیگر مقایسه کنید.
- به نظر شما کدام مدل بهتر آموزش دیده است؟
- آستانه پاکسازی را در داده‌های پاکسازی شده تغییر دهید و نتیجه مدل را رسم کنید.
- چه تغییری در مدل‌ها مشاهده می‌کنید؟ مقایسه کنید.

۳.۳.۲

آموزش رگرسیون مقاوم

- در مورد رگرسیون مقاوم[۲] و تفاوت آن با رگرسیون خطی معمولی تحقیق کنید و توضیح دهید.
- با استفاده از رگرسیون‌های مقاوم درکتابخانه `sklearn` برای داده‌های پاکسازی نشده حداقل دو رگرسیون مقاوم را آموزش دهید.
- ضمن توضیح الگوریتم هرکدام از مدل‌های آموزش داده شده، آن‌ها را با رنگ‌های مختلف در فضای دو بعدی رسم و با هم مقایسه کنید.
- پارامترهای هر مدل را در خروجی چاپ کنید و با مدل‌هایی که در سوال بالا آموزش دادید مقایسه کنید.
- به نظر شما کدام مدل‌ها بهتر هستند؟
- آیا رگرسیون‌های مقاوم به پاکسازی داده‌ها نیاز دارند؟ چرا؟

۴.۲ داده‌جدید

۱.۴.۲

با داده‌های ردیف دوم و سوم قاب‌داده،

- بار دیگر به سوالات ۲ تا ۵ پاسخ دهید.
- آیا داده‌های ردیف سوم را می توان با رگرسیون تک متغیره به خوبی تقریب زد؟
- چه راهکاری پیشنهاد می‌کنید؟

---

[۱] cost function

[۲] robust regression

۲.۴.۲

فرض کنید داده‌های ردیف اول و دوم قاب‌داده مربوط به یک دسته داده می‌باشند،

- با استفاده از کتابخانه `mpl_toolkits.mplot3d`، آن‌ها را در یک فضای سه بعدی رسم کنید.
- آیا می‌توان این داده‌ها را با رگرسیون تک متغیره تقریب زد؟
- برای داده‌های ردیف اول و دوم قاب‌داده بار دیگر به سوالات ۲ تا ۵ پاسخ دهید.

امتیازی

در مورد مفاهیم تابع هزینه و تابع اتلاف[۱] تحقیق کنید.

- تفاوت تابع اتلاف و تابع هزینه را بیان کنید.
- مدل‌های شما چگونه اختلاف بین پیشبینی و مقدار واقعی را کمینه می‌کنند؟
- بررسی کنید برای این داده‌ها کدام توابع هزینه مناسب‌تر هستند.

[۱] loss function

در انجام این تمرین حتماً به نکات زیر توجه کنید:

- موعد تحویل این تمرین، ساعت ۱۸:۰۰ روز پنجشنبه ۱۴ فروردین‌ماه ۱۴۰۴ 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 آی‌دی‌های `MJAHMADEE` و `AliBagheriNejad` 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 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)
- 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.
- رعایت نکات بالا به حرفه‌ای‌ترشدن شما کمک خواهد کرد و اهمیتی معادل مطالب درسی فراگرفته‌شده دارد؛ بنابراین، درصورت عدم رعایت هریک از این نکات، از نمره تمرین شما **کسر خواهد شد**.
- آی‌دی پرسش هرگونه ابهام درخصوص سوال اول
- آی‌دی پرسش هرگونه ابهام درخصوص سوال دوم