مدرس: **دکتر فدایی و دکتر یعقوب‌زاده** طراحان: محمدطاها فخاریان، فاطمه سیددباغی

مهلت تحویل: شنبه ۱۶ مهر ۱۴۰۱، ساعت ۲۳:۵۹

مقدمه

در این پروژه، شما با Jupyter Notebook و برخی کتابخانه‌های پایتون آشنا می‌شوید که ابزارهای مهمی در مسیر هوش مصنوعی و یادگیری ماشین هستند. در این پروژه ابتدا به بررسی و visualization داده‌ها پرداخته و در ادامه‌ی تحلیل‌هایی که روی داده‌ها انجام داده‌اید، یک مدل ساده‌ی classification برای پیش‌بینی به دست می‌آورید. کتابخانه‌های مورد استفاده در این پروژه pandas، numpy و matplotlib به همراه ابزار notebook jupyter خواهند بود، که برای آشنایی بیشتر با آنها می‌توانید لینک مربوط به هرکدام را مطالعه کنید.

معرفی مجموعه داده

فایل train.csv در کنار صورت پروژه قرار گرفته است؛ که برای پیش‌بینی اینکه آیا مسافر حاضر در کشتی زنده می‌ماند یا خیر، استفاده می‌شود. در هر سطر از این فایل یک رکورد مربوط به یک کاربر آمده است که اطلاعات زیر را نشان می‌دهد:

- وضعیت زنده ماندن یا نماندن
- نوع بلیط
- جنسیت
- سن
- تعداد خواهر، برادر یا همسران هر شخص در کشتی
- تعداد والدین یا فرزندان هر شخص در کشتی
- شماره بلیط
- کرایه بلیط مسافر
- شماره کابین
- بندری که در آن مسافر سوار شده است (C = Cherbourg, Q = Queenstown, S = Southampton)

فایل test.csv نیز در کنار صورت پروژه قرار داده شده‌است اما ستون Survived برای داده‌های آن وجود ندارد. در این پروژه می‌خواهیم این مقادیر را با استفاده از یک مدل آماری ساده پیش‌بینی کنیم. برای ساخت این مدل از سایر نمونه‌ها (train.csv) استفاده می‌کنیم.

روش حل مسئله:

توجه داشته باشید که در تمامی مراحل داده‌کاوی، شما باید هر عملی را با Vectorization انجام دهید. **استفاده از حلقه مجاز نمی‌باشد.** توضیحات مربوط به vectorization در انتها آمده است.

۱. ابتدا فایل train.csv را با استفاده از کتابخانه pandas خوانده و محتوای آن را در یک dataframe ذخیره کنید. سپس با استفاده از متدهای head, tail, describe و info از کتابخانه pandas، ساختار کلی داده‌ها را بررسی کرده و توضیح دهید که هر کدام از خروجی‌ها، چه اطلاعاتی را نشان می‌دهد.

۲. حال با استفاده از تابع info کتابخانه pandas نوع هر کدام از ستون‌های داده را نشان دهید. بعضی ستون ها از نوع دسته‌ای[1] و بعضی دیگر از نوع عددی[2] هستند. برای پردازش ستون‌های غیر عددی، یکی از راه‌های ممکن برچسب‌گذاری[3] است؛ به صورتی که هر کدام از دسته‌ها با یک عدد جایگزین شوند.

برای مثال در این مجموعه داده، ستونی دسته‌ای با نام Sex وجود دارد که شامل مقادیر male و female می‌باشد. مقادیر این ستون را به گونه‌ای تغییر داده که هر کدام از این مدل‌ها به یکی از اعداد بازه‌ی [0,1] نگاشته شوند.

۳. شاید متوجه شده باشید که مقدار بعضی از ستون‌های بعضی سطرها، NaN است که معمولاً این مشکل در داده‌ها وجود دارد. pandas مقادیری که خالی باشند را با NaN نشان می‌دهد. حال با استفاده از همین کتاب‌خانه و با فراخوانی یک تابع، برای هر ستون تعداد سطرهایی را که مقدار آن ستون برای آنها خالی است را نشان دهید. سپس مقدار سلول‌هایی را که خالی هستند را با روش مناسب، مانند میانگین همان ستون، جایگزین کنید. توجه داشته باشید که ستون‌هایی که مقادیر اکثر سلول‌های آن‌ها NaN هستند را می‌توان به جای پر کردن، به طور کامل حذف کرد. **مزایا و معایب روش پر کردن سلول‌ها با مقدار میانگین را در گزارش خود ذکر نمایید.**

۴. در این مجموعه داده، ستون‌هایی وجود دارند که برای هر سلول، مقدار منحصربه‌فردی دارند؛ از این رو، حضور این ستون‌ها اطلاعات بیشتری برای پیش‌بینی زنده ماندن یا نماندن مسافران در اختیار ما قرار نمی‌دهند و در ادامه کار، بهتر است این ستون‌ها را از داده حذف کرد.

---

[1] Categorical

[2] Numerical

[3] Label Encoding

۵. با فراخوانی یک تابع از کتابخانه pandas نشان دهید چه تعداد از مسافران زن و چه تعداد مرد هستند. سپس نشان دهید چه تعداد از مردان در بندر Southampton سوار شده‌اند.

۶. تعداد مسافران بالای ۳۵ سال بدون هیچ همراهی را نشان دهید که نوع بلیت آن‌ها ۳ می‌باشد.

۷. با فراخوانی یک تابع از کتابخانه pandas، میانگین کرایه بلیت مسافرانی که در بندر Queenstown سوار شده‌اند را نشان دهید.

۸. قسمت قبل را بار دیگر بدون استفاده از vectorization (با استفاده از حلقه) انجام دهید. زمان اجرای دو روش را ثبت و مقایسه کرده، در گزارش خود بیاورید.

۹. با استفاده از تابع hist کتابخانه pandas، شکل توزیع هر ستون از داده را روی نمودار نشان دهید.

۱۰. یکی از راه‌های بهبود داده‌ها برای مدل‌های یادگیری ماشین، نرمال‌سازی داده‌هاست. برای تمام ستون‌ها، نرمال سازی را با کم کردن میانگین و تقسیم کردن بر انحراف‌معیار انجام داده و نتیجه را نشان دهید.

۱۱. ابتدا برای هر دو حالتی که مسافر زنده مانده است یا نه، میانگین و انحراف معیار را بدست آورده و ذخیره کنید. سپس با استفاده از scipy.stats تابع چگالی احتمال (PDF) توزیع نرمال ویژگی مربوطه با میانگین و انحراف معیاری که بدست آوردید را رسم کنید. توجه کنید که باید هر دو منحنی مربوط به حالات زنده‌مانده/نمانده روی یک نمودار با رنگ متفاوت رسم شوند و خوانا باشند. این نمودارها را تحلیل کنید و بهترین ویژگی(ها) را برای انتخاب به عنوان ورودی مدل گزارش کنید. استدلال خود را برای انتخاب این ویژگی شرح دهید.

۱۲. با استفاده از میانگین‌ها و انحراف معیارهای ویژگی انتخاب شده در قسمت قبل، برای سطرهای فایل test.csv، کلاس متناسب (زنده ماندن یا نماندن) پیش‌بینی کرده و همراه اندیس متناظر نشان داده و در یک فایل csv ذخیره کنید.

توضیحات Vectorization

Vectorization در واقع عمل، رهایی کد از حلقه‌هاست. در هوش مصنوعی، شما با داده‌های بزرگی کار می‌کنید؛ در نتیجه اینکه کد شما بتواند روی این داده‌ها سریع عمل کند بسیار مهم است. با استفاده از vectorization، محاسبات روی مجموعه‌های بزرگی از داده‌ها به صورت موازی و در نتیجه بسیار سریع تر انجام می‌شود. در این لینک میتوانید در مورد vectorization و broadcasting در numpy بیشتر بخوانید.

نکات پایانی

۱. دقت کنید که هدف پروژه تحلیل نتایج است؛ بنابراین از ابزارهای تحلیل داده مانند نمودارها استفاده کنید و توضیحات مربوط به هر بخش از پروژه را به طور خلاصه و در عین حال مفید، در گزارش خود ذکر کنید. اگر در جایی ذکر شده مقایسه‌ای انجام دهید، حتما نتایج را دقیق ذکر کنید و سپس آن‌ها را تحلیل و مقایسه کنید.

۲. نتایج و گزارش خود را در یک فایل فشرده با عنوان AI_CA0_<#SID>.zip تحویل دهید. محتویات پوشه باید شامل موارد زیر باشد:

- فایل jupyter-notebook، خروجی html و فایل‌های مورد نیاز برای اجرای آن باشد. توضیح و نمایش خروجی‌های خواسته شده بخشی از نمره این تمرین را تشکیل می‌دهد. از نمایش درست خروجی‌های مورد نیاز در فایل html مطمئن شوید.
- در صورتی که از jupyter-notebook استفاده نمی‌کنید، کدهای تمام قسمت‌هایی از تمرین که پیاده‌سازی نموده‌اید، در یک پوشه به نام Code قرار دهید و گزارش پروژه با فرمت PDF شامل شرح تمامی کارهای انجام‌شده، نتایج به دست‌آمده و تحلیل‌ها و بررسی‌های خواسته‌شده در صورت پروژه را هم در کنار آن پوشه قرار دهید.
- فایل csv نتایج پیش‌بینی مدل (شامل اندیس‌ها و کلاس متناظر آنها).

۳. در صورتی که سوالی در مورد پروژه داشتید بهتر است در فروم درس یا در گروه تلگرام مطرح کنید تا بقیه از آن استفاده کنند؛ در غیر این صورت از طریق ایمیل با طراحان در ارتباط باشید.

۴. هدف از تمرین، یادگیری شماست. لطفا تمرین را خودتان انجام دهید.

موفق باشید!