مقدمه

هدف اصلی این پروژه آشنایی با برخی کتابخانه‌های پایتون است که به عنوان ابزاری قدرتمند، در مسیر یادگیری مفاهیم هوش مصنوعی و یادگیری ماشین به شما کمک زیادی خواهند کرد. در این پروژه عملی، در ابتدا به اندازه کافی با داده‌ها کار کرده و در ادامه از آن‌ها برای پیش‌بینی یک مدل ساده از رگرسیون خطی استفاده خواهید کرد.

کتابخانه‌های مورد استفاده در این پروژه numpy و pandas و matplotlib به همراه ابزار jupyter notebook خواهند بود، که برای آشنایی بیشتر با آن‌ها می‌توانید لینک مربوط به هرکدام را مطالعه کنید.

تعریف مسأله

در این پروژه شما قرار است ابتدا با ابزارهای معرفی شده، داده‌ها را کاوش کرده و در نهایت یک مدل ساده برای پیش‌بینی احتمال پذیرش دانشجویان فارغ‌التحصیل (graduate)، در دانشگاه‌های آمریکا برای مقطع کارشناسی ارشد طراحی کنید.

این مدل مشخصه‌های زیر را به عنوان ورودی گرفته:

۱. شماره سریال دانشجو
۲. نمره GRE (حداکثر ۳۴۰)
۳. نمره TOEFL (حداکثر ۱۲۰)
۴. رتبه دانشگاه مبدا (حداکثر ۵)
۵. امتیاز SOP (Statement of Purpose) (حداکثر ۵)
۶. امتیاز LOR (Letter of Recommendation) (حداکثر ۵)
۷. CGPA (حداکثر ۱۰)
۸. داشتن یا نداشتن تجربه research (عدد باینری ۱ یا ۰)

و

۹. شانس پذیرش (عدد حقیقی بین ۰ و ۱)

را خروجی می‌دهد.

در حالت کلی این پیش‌بینی برای هر دانشجوی مورد آزمون (test) بر اساس اطلاعات و نتیجه پذیرش سایر دانشجویانی که قبلا به عنوان داده train به مدل داده شده اند انجام می‌شود.

قسمت اول. بارگیری و کاوش داده‌ها

فایل AdmissionPredict.csv در کنار پروژه قرار گرفته که حاوی اطلاعات حدود ۴۰۰ دانشجوی فارغ‌التحصیل است.

I. محتوای این فایل را با کتابخانه pandas بخوانید و روی dataframe خروجی، توابع head (یا tail) و describe و info را فراخوانی کرده، نتیجه را مشاهده کنید.

I I. شاید متوجه شده باشید که در داده‌هایی که در اختیار دارید، نقص هایی وجود دارد. با استفاده از توابع کتابخانه pandas، تعداد مقادیر NaN را در هر ستون از دیتافریم بدست آورید. (pandas مقادیر گمشده یا missing را با NaN نمایش می‌دهد)

I I I. سلول هایی که دچار نقص شده‌اند را با مقدار میانگین همان ستون پر کنید تا به دیتاست کاملی برسید. نوع داده‌ای که جایگزین می‌کنید باید مطابق نوع داده‌ی همان ستون باشد (به عنوان مثال در ستونی با مقادیر صحیح مقدار صحیح میانگین را جایگزین کنید). توجه کنید که ستون "Chance of Admit" متغیر هدف و برای پیش‌بینی بوده و نقص‌های آن نباید جایگزین شود.

قسمت دوم. همبستگی و ارتباط مشخصه‌ها با متغیر هدف

از آنجا که هدف پیش‌بینی شانس پذیرش دانشجو بر اساس مشخصه‌های ورودی است، خوب است که رابطه هریک از این مشخصه ها و تاثیر آن‌ها بر شانس پذیرش را در نمودار به چشم ببینیم.

I. بنابراین با استفاده از کتابخانه matplotlib به ازای هر مشخصه، یک scatterplot که شانس پذیرش برحسب مشخصه مورد نظر را نشان می‌دهد، رسم کنید. این ۸ نمودار را در گزارش خود ضمیمه کنید.

I I. همچنین مشخصه‌ای که به نظر شما بیشترین هم‌بستگی (از لحاظ خطی بودن) با شانس پذیرش دارد را انتخاب کرده، روی انتخاب خود استدلال کنید.

قسمت سوم. کار با داده‌ها

در این قسمت می‌خواهیم با استفاده از توابع numpy و pandas (و بدون استفاده از حلقه) تسک های زیر را انجام دهیم:

I. با این فرض که یک دانشگاه خاص فقط به دانشجویانی که نمره CGPA حداقل ۹ و نمره TOEFL حداقل ۱۱۰ داشته باشند، پذیرش می‌دهد، دانشجویان با شرایط مطلوب را بر این اساس فیلتر کنید و تعداد آن‌ها را گزارش کنید.

II. به ازای هر رتبه دانشگاه، میانگین نمره GRE دانشجویان فارغ‌التحصیل از دانشگاه‌های با این رتبه را به دست آورید و میانگین‌های مربوط به این ۵ رتبه (از ۱ تا ۵) را در گزارش خود ذکر کنید.

قسمت چهارم. رگرسیون خطی تک متغیره

I. مشخصه انتخاب شده در قسمت دوم را در نظر بگیرید. از روی داده‌های این ستون به همراه داده‌های ستون هدف (شانس پذیرش)، یک دیتافریم جدید بسازید. (در ادامه با این دیتافریم جدید کار خواهید کرد)

شما در این مرحله باید به منظور تخمین شانس پذیرش، یک تخمین‌گر خطی بر اساس مشخصه انتخاب شده طراحی کنید. در واقع می‌خواهیم خطی بر داده‌های نمودار منطبق کنیم که به نحوی شانس پذیرش را تخمین بزند.

- تابع تخمین‌گر (Hypothesis Function):

در این قسمت تابع تخمین‌گر را به صورت زیر تعریف می‌کنیم:

$$h_\theta(x) = \theta_0 + \theta_1 x$$

که متغیر x همان متغیر ورودی یا مشخصه انتخاب شده است. می‌خواهیم پارامترهای θ_0 (عرض از مبدا) و θ_1 (شیب) را به گونه‌ای انتخاب کنیم که تابع خطی $h_\theta(x)$ با دقت قابل قبولی، y را تخمین بزند. (متغیر y همان متغیر خروجی یا target (در اینجا شانس پذیرش) می‌باشد)

در حالت کلی ورودی مدل می‌تواند بیش از یک عدد باشد و در واقع یک بردار باشد، که در این صورت θ نیز برداری از θ_j ها خواهد بود، اما در این پروژه به منظور سادگی فرض می‌کنیم که ورودی مدل صرفا یک عدد باشد.

- تابع هزینه (Cost Function):

به منظور ارزیابی تابع تخمین‌گر، تابعی به نام هزینه مانند زیر تعریف می کنیم (که به آن MSE یا Mean Squared Error می‌گویند):

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

در فرمول بالا، m تعداد داده‌های train، $x^{(i)}$ مقدار مشخصه سطر i ام داده‌ها و $y^{(i)}$ مقدار هدف سطر i ام داده‌ها را نشان می‌دهند.

هدف یافتن θ_0 و θ_1 ای است که $J(\theta_0, \theta_1)$ را کمینه کند.

II. در این قسمت می‌بایست به روش دلخواه خود، θ_0 و θ_1 ای که مقدار تابع هزینه $J(\theta_0, \theta_1)$ را کمینه می‌کند، به دست آورید و در گزارش خود ذکر کنید. به این صورت که تابع تخمین‌گر بدست آمده یعنی $h_\theta(x)$ را روی نمودار شانس پذیرش بر حسب مشخصه انتخاب شده رسم کرده و مطمئن شوید به خوبی روی نقاط scatterplot منطبق می‌شود. (دقت کنید که حداکثر مقدار قابل قبول $J(\theta_0, \theta_1)$ برای این قسمت ۰.۱ است)

III. در نهایت دانشجویانی که مقدار شانس پذیرش آن‌ها NaN است را یافته، خروجی مدل خود یعنی شانس پذیرش را به ازای هر یک از این دانشجوها محاسبه و نتیجه را به همراه شناسه دانشجو در گزارش خود بیاورید.

ملاحظات

- موعد تحویل غیرحضوری تا پایان روز جمعه ۲۵ بهمن می‌باشد.
- در تمامی مراحل (در صورت امکان) تسک ها را به صورت برداری (vectorized) و بدون استفاده از حلقه انجام دهید، در غیر این صورت بخشی از امتیاز آن را از دست خواهید داد.
- تمامی نتایج باید در یک فایل فشرده با عنوان AI-CA0-<#STID>.zip تحویل داده شود. این فایل باید شامل موارد زیر باشد:
  - یک پوشه به نام Code شامل کدهای تمام قسمت‌هایی از تمرین که پیادهسازی نموده‌اید.
  - گزارش پروژه با فرمت PDF و شامل شرح تمامی کارهای انجام شده، نتایج به دست آمده و تحلیل‌ها و بررسی‌های خواسته شده در صورت پروژه.
  - درصورتی که از Jupyter Notebook استفاده می‌کنید نیازی به ارسال جداگانه کدها و گزارش نیست و هردو را می‌توانید در یک فایل Notebook ارائه دهید. حتما خروجی html فایل Notebook خود را نیز همراه فایل Notebook ارسال کنید.