۱. 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 و Scikit-Learn خواهند بود.

۲. تعریف مسئله

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 فارغ التحصیل (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.

۳. پرسش ها

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

۳,۱. قسمت اول) بارگیری و کاوش داده ها

۱. محتوای این فایل را با کتابخانه pandas بخوانید و روی dataframe خروجی، توابع describe و info را فراخوانی کرده، نتیجه را مشاهده کنید.

۲. 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 missingرا با NaN نمایش می‌دهد.)

۳. سلول‌هایی که دچار نقص شده‌اند را با ۲ روش متفاوت پر کنید تا به مجموعه دادگان 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) در مورد مزایا و معایب هر کدام از این روش ها در حالت کلی توضیحی ارائه دهید.

تذکر: توجه کنید که ستون "Chance of Admit" متغیر هدف و برای پیشبینی بوده و بنابراین سطرهای دارای نقص در این ستون را حذف کنید.

۳,۲. 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.

۱. 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

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

۳,۳. قسمت سوم) رگرسیون

سوال یک

یکی از مجموعه دادگان بدست آمده در سوال دوم قسمت یک را در نظر بگیرید. مجموعه دادگان را به ۲ قسمت تقسیم کرده به طوری که ۸۰ درصد را به آموزش و باقی داده ها را به عنوان داده‌ی اعتبارسنجی در نظر بگیرید. حال 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.

شما در این مرحله باید به منظور تخمین شانس پذیرش، یک تخمین‌گر خطی بر اساس مشخصه انتخاب شده طراحی کنید.

ملاحظات

- در تمامی مراحل (در صورت امکان) تسک‌ها 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.
- در فرآیند پیاده سازی رگرسیون در صورت نیاز در هر مرحله می‌توانید از یکی از دستورات scale کتابخانه‌ی scikit استفاده کنید و در نهایت نتیجه‌ی بدست آمده بدون استفاده از scale و با استفاده از آن را حداقل برای بار اولی که استفاده می‌کنید، با یکدیگر مقایسه کنید. (در صورت علاقه می‌توانید دستورات مختلفscale را نیز با هم مقایسه کنید.)

۱. در این قسمت می‌بایست پارامترهای تخمین‌گر خطی، ابتدا با استفاده از معادله نرمال به دست آورید و در گزارش خود ذکر کنید. به این صورت که تابع تخمین‌گر بدست آمده را روی نمودار شانس پذیرش بر حسب مشخصه انتخاب شده رسم کرده و مطمئن شوید به خوبی روی نقاط scatterplot منطبق می‌شود.

$$\theta = (X^T X)^{-1} X^T \vec{y}$$

شکل ۱- معادله نرمال

سپس همین مرحله را یکبار دیگر با استفاده از دستورات کتابخانه‌ی Scikit-Learn پیاده سازی کنید تا از نتایج بدست آمده از پیاده سازی خود اطمینان حاصل کنید. هم چنین مقدار MSE را برای دادگان آموزش و اعتبارسنجی گزارش کنید.

۲. این بار با استفاده از نمودارهایی که در قسمت پیش رسم کرده‌اید، سه مشخصه را انتخاب کنید و خواسته های گفته شده در مرحله قبل را با استفاده از کتابخانه‌ی Scikit-Learn بدست آورید. نتایج بدست آمده را با نتایج بخش قبلی مقایسه کنید

۳. این بار، تمام مشخصه ها(به جز مشخصه‌ی سریال) رو به عنوان مشخصه های تخمین‌گر در نظر بگیرید و خواسته‌های بخش قبل را با استفاده از کتابخانه‌ی Scikit-Learn بدست آورید. نتایج بدست آمده در این بخش را با بخش های قبلی مقایسه کنید.

۴. نتایج بدست آمده در سوال پیش را این بار دیتافریمی که مقادیر گمشده‌ی آن با روش دوم پر شده است، باز تولید کند و نتیجه بدست آمده را با بخش قبلی مقایسه کنید و آن را تحلیل کنید.

سوال دو

۱. در این سوال می‌خواهیم، over-fitting و under-fitting را برای یک سری داده بررسی کنیم. ابتدا، به تعداد دلخواه نمونه sin تولید کرده، (دامنه این سیگنال بین منفی ۵ تا مثبت ۵ باشد.) سپس این داده‌ها را با نویز گوسی میانگین یک و واریانس دلخواه جمع کنید. حال سعی کنید که تابع درجه یک، سه، هفت، یازده، شانزده، بیست برای این داده‌ها برازش کنید. مقادیر MSE برای هر یک از موارد گزارش دهید. نمودار برازش شده بر این داده‌ها را رسم کنید. مشاهده خود را از نتایج به دست آمده شرح دهید.

ملاحظات

- در تمامی مراحل (در صورت امکان) تسک‌ها 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.
- جهت تولید نمونه سینوسی از sin(x * pi * 2) و نویز از random در numpy استفاده کنید و هم چنین برای رگرسیون خطی می‌توانید از کتابخانه‌ی Scikit-Learn استفاده کنید.
- در فرآیند پیاده سازی رگرسیون در صورت نیاز در هر مرحله می‌توانید از یکی از دستورات scale کتابخانه‌ی scikit استفاده کنید و در نهایت نتیجه‌ی بدست آمده بدون استفاده از scale و با استفاده از آن را حداقل برای بار اولی که استفاده می‌کنید، با یکدیگر مقایسه کنید. (در صورت علاقه می‌توانید دستورات مختلف scale را نیز با هم مقایسه کنید.)