هوش مصنوعی تمرین کامپیوتری شماره‌ی صفر

مدرس: **دکتر فدایی و دکتر یعقوب‌زاده** طراح: بهار افشار، مهسا مسعود

مهلت تحویل: چهارشنبه ۱۴ مهر ۱۴۰۰، ساعت ۲۳:۵۵

مقدمه

در این پروژه شما با Jupyter Notebook و برخی کتاب‌خانه‌های پایتون آشنا می‌شوید که ابزارهای کاربردی در حوزه هوش مصنوعی و یادگیری ماشین هستند. در این پروژه شما ابتدا به بررسی و Visualization داده‌ها می‌پردازید. سپس با استفاده از تحلیل‌هایی که روی داده‌ها انجام داده‌اید، یک مدل ساده‌ی رگرسیون خطی برای پیش‌بینی به دست می‌آورید.
کتاب‌خانه های مورد استفاده در این پروژه pandas، numpy و matplotlib به همراه ابزار Jupyter Notebook خواهند بود که برای آشنایی بیشتر با آنها می‌توانید لینک مربوط به هرکدام را مطالعه کنید.

معرفی مجموعه داده

فایل FuelConsumptionCo2.csv در کنار صورت پروژه قرار گرفته‌است که حاوی اطلاعاتی از ماشین‌های مختلف است که برای پیشبینی مقدار کربن‌دی‌اکسید تولیدی هر ماشین به کار می‌رود. هر رکود این دیتاست حاوی اطلاعات زیر است:

۱. سال تولید
۲. شرکت سازنده
۳. مدل
۴. کلاس خودرو
۵. سایز موتور
۶. تعداد سیلندر
۷. نوع جعبه‌دنده
۸. نوع سوخت
۹. مقدار مصرف سوخت در شهرها
۱۰. مقدار مصرف سوخت در اتوبان‌ها
۱۱. مقدار مصرف سوخت ترکیبی(شهر و اتوبان)
۱۲. مقدار مصرف سوخت ترکیبی در واحد مایل/گالن
۱۳. مقدار کربن‌دی‌اکسید خروجی(هدف)

ورودی مدل یکی از ویژگی‌هایی است که در بالا آمده‌اند و خروجی آن نیز ستون هدف(مقدار کربن‌دی‌اکسید تولیدی) است. برای تعداد کمی از نمونه‌ها مقدار ستون هدف موجود نیست. در این پروژه می خواهیم این مقادیر را با استفاده از یک مدل

رگرسیون ساده پیش‌بینی کنیم. برای ساخت این مدل از سایر نمونه‌ها، که مقدار ستون هدف برای آن‌ها مشخص است، استفاده خواهیم کرد.

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 csv را با استفاده از کتاب‌خانه pandas خوانده و محتوای آن را در یک DataFrame ذخیره کنید. سپس با استفاده از توابع tail, head و describe اطلاعات مربوط به داده‌ها را نشان داده و توضیح دهید که هر کدام از خروجی‌ها نشان‌دهنده چه اطلاعاتی هستند.

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 FUELTYPE وجود دارد که شامل مقادیر X, E, D, Z می باشد. مقادیر این ستون را به گونه ای تغییر داده که هر کدام از این مدل‌ها به یکی از اعداد بازه‌ی [0,3] نگاشته شوند.

۳. شاید متوجه شده‌باشید 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 نشان دهید. سپس مقدار سلول‌هایی را که خالی هستند با میانگین همان ستون جایگزین کنید. توجه داشته باشید که سلول‌هایی را که مقدار هدف آنها خالی است نباید جایگزین کنید. **مزایا و معایب این روش (پر کردن سلول‌ها با مقدار میانگین) را در گزارش خود ذکر نمایید.**

سطرهایی که مقدار ستون هدف آنها NaN است را از دیتافریم اصلی جدا کرده و در دیتافریم جدیدی ذخیره کنید. در مراحل بعدی از دیتافریم اصلی(و نه این دیتافریم جدید) استفاده کنید.

۴. با فراخوانی یک تابع از کتابخانه pandas، میانگین مقدار مصرف سوخت در شهرها را برای اتومبیل‌هایی که میزان کربن‌دی‌اکسید تولیدیشان از ۲۴۰ کمتر است، بدست آورید. این مقدار برای خودروهایی که CO2 تولیدیشان از ۳۰۰ بیشتر است چقدر است؟

---

[1] Categorical

[2] Numerical

[3] Label Encoding

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

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

در این پروژه تنها از ویژگی‌هایی استفاده میکنیم که مقدار آنها عددی باشد. در قسمت‌های بعد ستون‌های غیر عددی را کنار بگذارید(ستون FUELTYPE را هم کنار بگذارید).

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

۸. از آنجایی که هدف پیشبینی مقدار تولیدی کربن‌دی‌اکسید براساس ویژگی‌های ورودی است، میخواهیم رابطه‌ی هریک از این ویژگی‌ها و تاثیر آنها بر مقدار تولیدی کربن‌دی‌اکسید را در نمودار مشاهده کنیم.

الف) با استفاده از کتاب‌خانه matplotlib به ازای هر ویژگی یک plot scatter رسم کنید که قیمت خودرو را برحسب آن ویژگی نشان بدهد. این نمودارها را در گزارش خود بیاورید.

ب) ویژگی دارای بیشترین همبستگی با مقدار کربن‌دی‌اکسید تولیدی (از لحاظ خطی بودن) را انتخاب کرده و انتخاب خود را توجیه کنید.

۹. ویژگی انتخاب‌شده در قسمت قبل را در نظر بگیرید. از روی داده‌های این ستون به همراه داده‌های ستون هدف (مقدار کربن‌دی‌اکسید تولیدی)، یک دیتا فریم جدید بسازید (در ادامه با این دیتا فریم جدید کار خواهیدکرد).

شما در این مرحله باید به منظور تخمین مقدار کربن‌دی‌اکسید تولیدی، یک تخمینگر خطی بر اساس ویژگی انتخاب شده طراحی کنید. در واقع میخواهیم خطی بر داده‌های نمودار منطبق کنیم که به نحوی مقدار کربن‌دی‌اکسید تولیدی خودروها را تخمین بزند.

تابع تخمین‌گر[4]

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

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

[4] Hypothesis Function

که متغیر x همان متغیر ورودی یا ویژگی 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 متغیر هدف (مقدار کربن‌دی‌اکسید تولیدی) را تخمین بزند. 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_jها خواهد بود، اما در این پروژه به منظور سادگی فرض می‌کنیم که ورودی مدل صرفا یک عدد باشد.

۱۰. به منظور ارزیابی تابع تخمین‌گر، تابعی به نام تابع هزینه با فرمول زیر تعریف میکنیم (که به آن MSE یا Mean Squared Error گفته می‌شود، توجه کنید که در فرمول زیر y_i همان مقادیر ستون هدف می‌باشد).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - h_\theta(x))^2$$

توجه داشته باشید که خطای محاسبه‌شده روی داده با ستون هدف غیر NaN، باید کمتر از ۰.۵ باشد و در غیر این صورت از شما نمره کسر می‌گردد.

۱۱. نمودار تابع تخمین‌گر و plot scatter ویژگی منتخب را روی یک نمودار رسم کرده و آن را تحلیل کنید.

۱۲. حال برای تمام سطرهای دیتافریم جدیدی که در انتهای بخش ۳ ذخیره کرده‌اید، با استفاده از این تخمین‌گر خطی قیمت را تخمین زده و نتیجه را نشان دهید.

توضیحات Vectorization

Vectorization در واقع عمل رهایی کد از حلقه هاست. در هوش مصنوعی، شما با داده های بزرگی کار می کنید؛ 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 سریعتر انجام می‌شود. در این لینک میتوانید در مورد vectorization و broadcasting در numpy بیشتر بخوانید.

نکات پایانی

موعد آپلود پروژه تا پایان روز چهارشنبه ۱۴ مهر می‌باشد.

- 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)

- در صورتی که سوالی در مورد پروژه داشتید بهتر است در فروم درس مطرح کنید تا بقیه از آن استفاده کنند؛ در غیر این صورت از طریق ایمیل با طراحان در ارتباط باشید.
- هدف از تمرین، یادگیری شماست. لطفا تمرین را خودتان انجام دهید.

موفق باشید!