هوش مصنوعی تمرین کامپیوتری شماره‌ی ۰

مدرس: دکتر فدایی و دکتر یعقوب‌زاده طراح: مبینا شاه بنده، سارا توکلی

مهلت تحویل: سه شنبه ۱۲ اسفند ۱۳۹۹، ساعت ۲۳:۵۵

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 رگرسیون خطی برای پیش‌بینی به دست می‌آورید.

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

توضیحات مسئله

فایل car_price_dataset.csv در کنار صورت پروژه قرار گرفته‌است که حاوی اطلاعات مربوط به خودروهایی است که در آمریکا به فروش می‌روند. در هر سطر از این فایل یک رکورد از یک خودرو آمده که شامل اطلاعات زیر است:

۱. شناسه
۲. نام
۳. طول
۴. عرض
۵. ارتفاع

۶. وزن

۷. تعداد سیلندر

۸. حجم موتور

۹. قدرت اسب بخار

۱۰. مایل بر هر گالن در شهر

۱۱. مایل بر هر گالن در اتوبان

و

۱۲. قیمت (هدف)

ورودی مدل یکی از ویژگی‌هایی که در بالا آمده‌اند و خروجی آن هم ستون هدف (قیمت خودرو) است.

برای تعداد کمی از نمونه‌ها، مقدار ستون هدف موجود نیست. در این پروژه می‌خواهیم این مقادیر را با استفاده از یک مدل رگرسیون خطی ساده پیش‌بینی کنیم. برای ساخت این مدل، از سایر نمونه‌ها (که مقدار ستون هدف برای آنها مشخص است) استفاده می‌کنیم.

روش حل مسئله

توجه داشته باشید که در تمامی مراحل داده‌کاوی، **شما باید عملیات خواسته شده را با vectorization انجام دهید و استفاده از حلقه مجاز نیست.** 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 head, tail و describe اطلاعات مربوط به داده را نشان داده و توضیح دهید که هر کدام از خروجی‌ها نشان دهنده چه اطلاعاتی هستند.

۲. حال با استفاده از تابع info کتابخانه pandas نوع هر کدام از ستون‌های داده را نشان دهید. بعضی ستون‌ها از نوع دسته‌ای (categorical) هستند و بعضی دیگر از نوع عددی. برای پردازش ستون‌های غیرعددی، یکی از راه‌های ممکن برچسب (لیبل) گذاری[1] است؛ به صورتی که هر کدام از دسته‌ها با یک عدد جایگزین شوند.

[1] Label encoding

به عنوان مثال در این مجموعه داده، ستونی دسته‌ای با نام fueltype وجود دارد که مقادیر gas و diesel در آن وجود دارد. مقادیر این ستون را برای هر سطر به گونه‌ای تغییر داده که در صورت 0 بودن نشان‌دهنده این باشد که gas است و در صورت 1 بودن، diesel.

۳. شاید متوجه شده‌باشید که مقدار بعضی ستون‌های بعضی سطرها، NaN است که معمولاً این مشکل در داده‌ها وجود دارد. pandas مقادیری را که خالی باشند (گم شده[2]) با NaN نشان می‌دهد. حال با استفاده از همین کتابخانه، برای هر ستون تعداد سطرهایی را که مقدار آن ستون برای آنها خالی است نشان دهید و مقدار سلول‌هایی را که خالی هستند با میانگین همان ستون جایگزین کنید. توجه داشته‌باشید که سلول‌هایی را که مقدار ستون هدف آنها خالی است نباید جایگزین کنید. مزایا و معایب استفاده از این روش (پرکردن سلول‌های خالی با میانگین) را در گزارش خود بیاورید.

سطرهایی را که مقدار ستون هدف (price) آنها NaN است از دیتافریم اصلی جدا کرده و در دیتافریم جدیدی ذخیره کنید. دقت کنید که شناسه خودروها نیز در این دیتافریم جدید وجود داشته باشد. در مراحل بعدی از دیتافریم اصلی (و نه این دیتافریم جدید) استفاده کنید.

۴. با استفاده از کتابخانه pandas نشان دهید که چه تعداد خودرو برای هر تعداد سیلندر در این مجموعه داده وجود دارد.

۵. تعداد خودروهای بنزینی (gas) را که قدرت اسب بخار بیشتر از ۱۰۰ و مایل بر گالن شهری کمتر از ۱۵ دارند، بدست آورده و گزارش کنید.

۶. میانگین قیمت خودروهای بنزینی (gas) و خودروهای گازوئیلی (diesel) با فراخوانی یک تابع کتابخانه pandas نشان دهید.

۷. 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.

۸. 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 قسمت های بعد ستون‌های غیرعددی را کنار بگذارید (ستون fueltype را هم کنار بگذارید).

[2] Missing data

۹. یکی از راه‌های بهبود داده‌ها برای مدل‌های یادگیری ماشین، نرمال‌سازی داده‌ها [3] است. برای ستون قیمت، نرمال‌سازی را با کم کردن میانگین و تقسیم کردن بر انحراف‌معیار انجام داده و نتیجه را نشان دهید.

۱۰. از آنجایی که هدف پیش‌بینی قیمت خودرو براساس ویژگی های ورودی است، می‌خواهیم رابطه هریک از این ویژگی ها و تاثیر آن ها بر قیمت خودرو را در نمودار مشاهده کنیم.

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

ب) ویژگی دارای بیشترین همبستگی [4] با قیمت خودرو (از لحاظ خطی بودن) را انتخاب کرده و انتخاب خود را توجیه کنید.

۱۱. ویژگی انتخاب شده در قسمت قبل را در نظر بگیرید. از روی داده‌های این ستون به همراه داده‌های ستون هدف (قیمت)، یک دیتافریم جدید بسازید (در ادامه با این دیتافریم جدید کار خواهیدکرد).

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

تابع تخمینگر (Hypothesis Function)

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

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

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 (قیمت خودرو) 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_j ها خواهد بود، اما در این پروژه به منظور سادگی فرض می‌کنیم که ورودی مدل صرفا یک عدد باشد.

---

[3] Data normalization

[4] Correlation

۱۲. 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).

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

توجه داشته باشید که خطای محاسبه شده باید کمتر از 0.5 باشد و در غیر این صورت از شما نمره کسر می‌گردد.

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

۱۴. 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 (به ازای هر شناسه خودرو، قیمت متناظر را نشان دهید).

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 سریع‌تر انجام می‌شود. در این لینک می‌توانید در مورد vectorization و broadcasting در numpy بیشتر بخوانید.

ملاحظات

- موعد آپلود پروژه تا پایان روز سه شنبه ۱۲ اسفند است.
- تمامی نتایج باید در یک فایل فشرده با عنوان CA0-<#STID>.zip تحویل داده شود. این فایل باید شامل موارد زیر باشد:
  - یک فایل Notebook شامل کدها و گزارش در کنار هم (متن‌ها را می‌توانید با استفاده از Markdown بنویسید). حتما خروجی html فایل Notebook خود را نیز همراه فایل Notebook ارسال کنید. **نام فایل Notebook را به صورت CA0-<#STID>.ipynb قرار دهید.**

- 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 انجام شده، نتایج به دست آمده و تحلیل‌ها و بررسی‌های خواسته شده در صورت پروژه را هم در کنار آن پوشه قرار دهید.

- توجه داشته باشید که تمام بخش‌های پروژه باید قابلیت اجرای مجدد را در زمان تحویل داشته باشند و در صورت عدم حضور در تحویل، نمره‌ای دریافت نخواهید کرد.
- هیچ‌گونه شباهتی در انجام این پروژه بین افراد مختلف پذیرفته نمی‌شود. در صورت کشف هرگونه تقلب برای همه‌ی افراد متقلب نمره 100- در نظر گرفته می‌شود.
- استفاده از مراجع با ارجاع به آنها بلامانع است. اما در صورتی که گزارش شما ترجمه عینی از آنها باشد، یا از گزارش افراد دیگر استفاده کرده باشید، کار شما تقلب محسوب می‌شود.
- 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 به طراحان پروژه ایمیل بزنید:

shbmobina@gmail.com

saratvk1377@gmail.com

موفق باشید!