به نام خدا

دانشکده مهندسی برق و کامپیوتر دانشگاه تهران

هوش مصنوعی، ترم پاییز ۹۹

پروژه صفر

آشنایی با هوش مصنوعی

مهلت ارسال تا دوشنبه ۷ مهر

طراحان پروژه:

مبینا شاه‌بنده

امید واهب

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 ادامه با استفاده از تحلیل‌هایی که روی داده‌ها انجام داده‌اید، مدل ساده ای برای پیش‌بینی بدست می‌آورید.

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

توضیحات مسئله

فایل advertising_dataset.csv در کنار صورت پروژه قرار گرفته است که حاوی اطلاعات مربوط به ۱۰۰۰ کاربر یک سایت است که با یک آگهی اینترنتی برخورد کرده‌اند. در هر سطر از این فایل یک رکورد از یک کاربر آمده که شامل اطلاعات زیر است:

۱. کل زمانی که کاربر در سایت سپری کرده
۲. سن کاربر
۳. میانگین درآمد منطقه کاربر
۴. میزان استفاده روزانه کاربر از اینترنت
۵. موضوع آگهی
۶. شهر محل زندگی کاربر
۷. جنسیت کاربر
۸. کشور محل زندگی کاربر
۹. زمانی که رکورد ثبت شده

و

۱۰. آیا کاربر روی آگهی کلیک کرده یا خیر (هدف)

ورودی مدل یکی از ویژگی‌هایی که در بالا آمده‌اند و خروجی آن هم ستون هدف (آیا کاربر روی آگهی کلیک کرده یا خیر) است.

برای تعداد کمی از نمونه‌ها، مقدار ستون هدف موجود نیست. در این پروژه می‌خواهیم این مقادیر را با استفاده از یک مدل آماری 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 هرکدام 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 [۱] است؛ به صورتی که هر کدام از دسته‌ها با یک عدد جایگزین شوند.
همانطور که مشاهده می‌کنید، ستونی دسته‌ای با نام Gender وجود دارد که مقادیر Male و Female در آن وجود دارد. مقادیر این ستون را برای هر سطر به گونه ای تغییر داده که در صورت ۱ بودن نشان دهنده این باشد که جنسیت کاربر مرد است و در صورت ۰ بودن، زن.

۳. شاید متوجه شده باشید که مقدار بعضی ستون‌های بعضی سطرها، NaN است که معمولاً این مشکل در داده‌ها وجود دارد. pandas مقادیری را که خالی باشند (گم شده [۲]) با NaN نشان می‌دهد. حال با استفاده از همین کتابخانه، برای هر ستون تعداد سطرهایی را که مقدار

---

[۱] Label encoding

[۲] Missing data

آن ستون برای آنها خالی است نشان دهید و مقدار سلول‌هایی را که خالی هستند با میانگین همان ستون جایگزین کنید. توجه داشته باشید که سلول‌هایی را که مقدار ستون هدف آنها خالی است نباید جایگزین کنید. یک روش دیگر برای رسیدگی به مشکل سلول‌های خالی را بیان کنید و آن را با روش قبلی مقایسه کنید.

۴. با استفاده از کتابخانه pandas نشان دهید که چه تعداد از کاربران زن و چه تعداد مرد هستند. سپس نشان دهید چه تعداد از کاربران روی آگهی کلیک کرده و چه تعداد روی آگهی کلیک نکرده‌اند.

۵. تعداد کاربرانی را که سن‌شان از شما بیشتر و دارای جنسیتی مشابه شما هستند، بدست آورده و نشان دهید.

۶. میانگین سن کاربرانی را که روی آگهی کلیک کرده‌اند و کاربرانی را که روی آگهی کلیک نکرده‌اند 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 غیر عددی را کنار بگذارید (ستون Gender را هم کنار بگذارید).

۹. یکی از راه‌های بهبود داده‌ها برای مدل‌های یادگیری ماشین، نرمال‌سازی داده‌ها [۳] است. برای هر ستون ویژگی، نرمال‌سازی را با کم کردن میانگین و تقسیم کردن بر انحراف معیار انجام داده و نتیجه را نشان دهید.

۱۰. ابتدا برای هر دو حالتی که کاربر روی آگهی کلیک کرده و روی آگهی کلیک نکرده، میانگین و انحراف معیار هر کدام از ویژگی‌ها 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 کلیک/عدم کلیک کاربر روی یک نمودار رسم شوند و خوانا باشند. این نمودارها را تحلیل

---

[۳] Data normalization

کنید و بهترین ویژگی را برای انتخاب به عنوان ورودی مدل گزارش کنید. استدلال خود را برای انتخاب این ویژگی شرح دهید.

۱۱. با استفاده از میانگین‌ها و انحراف معیارهای ویژگی انتخاب شده در قسمت قبل، برای سطرهایی که مقدار ستون هدف آنها خالی (NaN) است، کلاس متناسب (کلیک/عدم کلیک) را 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 سریع‌تر انجام می‌شود. در این لینک می‌توانید در مورد vectorization و broadcasting در numpy بیشتر بخوانید.

ملاحظات

- موعد آپلود پروژه تا پایان روز دوشنبه ۷ مهر است.
- تمامی نتایج باید در یک فایل فشرده با عنوان AI-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.
  - 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.

- فایل csv نتایج پیش‌بینی مدل (شامل اندیس‌ها و کلاس متناظر آنها).

- توجه داشته باشید که تمام بخش‌های پروژه باید قابلیت اجرای مجدد را در زمان تحویل داشته باشند و درصورت 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

ovaheb@gmail.com

موفق باشید!