هوش مصنوعی **تمرین کامپیوتری شماره‌ی ۴**

مدرس: **دکتر فدایی** طراحان: علیرضا اربابی، اولدوز نیساری

مهلت تحویل: جمعه ۱ دی ۱۴۰۲، ساعت ۲۳:۵۹

مقدمه

هدف این تمرین، آشنایی با روش‌های یادگیری ماشین[1] جهت پیش‌بینی تعداد خریدهای کالای مشتریان یک فروشگاه است. این تمرین از سه فاز تشکیل شده است؛ در فاز اول به ساخت یک مدل Linear Regression مرتبه اول به صورت دستی(بدون استفاده از مدل آماده) می‌پردازید، سپس متد گرادیان کاهشی[2] و در فاز آخر با کمک کتابخانه Scikit-Learn اقدام به تخمین تعداد خریدها می‌پردازید.
در فاز اول و دوم لازم است که فایل نوت‌بوک قرار داده شده در سایت را دانلود کرده و بخش‌های مشخص شده را کامل نمایید. پیاده‌سازی فاز سوم نیز در ادامه آن‌ها و در همان نوت‌بوک انجام می‌شود.

آشنایی با مجموعه داده

مجموعه داده‌ای که در اختیار شما قرار دارد، شامل اطلاعات مربوط به خریدهای مشتریان یک فروشگاه به همراه ویژگی های شخصیتی آنها می‌باشد. تحلیل شخصیت مشتریان می‌تواند به فروشنده کمک کند تا با شناخت مشتریان، کالاهایی را عرضه نماید که فروش بیشتری داشته و رضایت مشتری را به همراه داشته باشد.

توضیح ستون های این مجموعه داده در جدول زیر قرار داده شده است:

نام ستون	توضیح
ID	Customer's unique identifier
Year_Birth	Customer's birth year
Education	Customer's education level
Marital_Status	Customer's marital status
Income	Customer's yearly household income
Kidhome	Number of children in customer's household
Teenhome	Number of teenagers in customer's household
Dt_Customer	Date of customer's enrollment with the company

[1] Machine Learning
[2] Gradient Descent

Recency	Number of days since customer's last purchase
Complain	if the customer complained in the last 2 years, 0 1 otherwise
MntCoffee	Amount spent on coffee in last 2 years
MntFruits	Amount spent on fruits in last 2 years
MntMeatProducts	Amount spent on meat in last 2 years
MntFishProducts	Amount spent on fish in last 2 years
MntSweetProducts	Amount spent on sweets in last 2 years
MntGoldProds	Amount spent on gold in last 2 years
NumPurchases	Number of purchased products in last 2 years
UsedCampaignOffer	If the customer has used a campaign offer, otherwise 0 1
NumWebVisitsMonth	Number of visits to company’s website in the last month

بررسی مجموعه داده

در این فاز داده‌های خام را بررسی خواهید کرد. این تجزیه و تحلیل داده‌ها با نام EDA[3] شناخته می‌شود و برای دریافت یک دید کلی نسبت مجموعه داده به کار می‌رود. مراحل زیر را انجام دهید و در هر مرحله نتیجه را تحلیل کرده و در گزارش بیاورید.

۱. ساختار کلی داده‌ها را با متدهای info و describe بدست بیاورید.
۲. برای هر ویژگی[4]، تعداد و نسبت داده‌های از دست رفته[5] را بدست بیاورید.
۳. نمودار وابستگی[6] ویژگی‌ها به یکدیگر را رسم کنید. کدام ویژگی‌ها وابستگی بیشتری به ستون هدف دارند؟
۴. برای ویژگی‌های بدست آمده در مرحله قبل نمودار تعداد مشاهدات هر مقدار منحصر به فرد را رسم کنید.
۵. ارتباط ویژگی‌ها با ستون هدف را دقیق‌تر بررسی کنید؛ از نمودارهای scatter و hexbin می‌توانید استفاده کنید.
۶. شما می‌توانید هر بررسی دیگری که به شناخت مجموعه کمک می‌کند را پیاده و تحلیل کنید.

[3] Exploratory Data Analysis
[4] Feature
[5] Missing

[6] Correlation

پیش پردازش مجموعه داده

در دنیای واقعی، اطلاعات جمع‌آوری شده به راحتی کنترل نمی‌شوند و در نتیجه مقادیر خارج از محدوده، ناممکن، از دست رفته و به طور کلی گمراه‌کننده برای آموزش مدل در مجموعه داده‌ها وجود دارند. در نتیجه قبل از ادامه پروژه باید این موارد را شناسایی و اصلاح کنیم. همچنین گاهی برای بهبود کارایی مدل و سرعت یادگیری می‌توان فرمت این داده‌ها را تغییر داد و خلاصه‌تر کرد. در نهایت این فاز مهمترین فاز یک پروژه یادگیری ماشین است؛ در غیر این صورت خروجی هم خروجی بسیار نادقیقی خواهد بود.
(به عبارتی "garbage in, garbage out")
در موارد زیر، علت انتخاب روش خود برای حل مسئله را نیز توضیح دهید:
۷. دو روش برای حل مشکل Missing Values، حذف کل ستون و پر کردن مقادیر خالی با آماره‌ها (برای مثال مد) می‌باشد. باقی روش‌ها را توضیح دهید و مقایسه کنید.
۸. بر اساس نتایج فاز قبل، کدام داده‌ها بیشترین میزان داده گم شده را دارند؟ برای تمامی ویژگی‌ها مشکل داده های گم شده را با کمک روش‌های مطرح شده حل کنید.
۹. در ویژگی‌های عددی، normalizing یا standardizing به چه منظور انجام می‌شود؟ در این پروژه نیاز به انجام این کار هست؟
۱۰. برای استفاده ویژگی‌های دسته‌ای، که معمولا بصورت یک string یا object در مجموعه داده ذخیره شده‌اند، در آموزش مدل چه پیش‌پردازش‌هایی مفید هستند؟ آیا همه داده‌های دسته این نیازمند این روش‌ها هستند؟
۱۱. آیا امکان حذف برخی ستون‌ها وجود دارد؟ چرا؟
۱۲. برای آموزش و در نهایت ارزیابی مدل یادگیری ماشین نیاز است که داده‌ها را به دو دسته test و train تقسیم کنیم. نسبت این تقسیم به چه صورت است؟ چه روش‌های برای تقسیم و ساخت این دو دسته وجود دارد؟
۱۳. گاهی علاوه بر دو دسته بالا یک دسته سومی هم وجود دارد. در مورد این دسته (validation) توضیح دهید.
۱۴. متد K-Fold Cross Validation به چه صورت انجام می‌شود؟ توضیح دهید.

آموزش، ارزیابی و تنظیم

فاز اول: Linear Regression

در این فاز از پروژه، به ساخت یک مدل linear regression درجه ۱، بدون استفاده از مدل آماده می‌پردازید. توجه کنید که در این فاز، به هیچ عنوان استفاده از کتابخانه‌های آماده(به جز numpy) مجاز نمی‌باشد.
۱۵. در ابتدای فایل نوت بوک قرار داده شده، فرمول محاسبه پارامترهای لازم برای یک مدل رگرسیون درجه ۱ قرار داده شده است. محاسبات ریاضی فوق را بررسی کنید، و علت بدست آمدن مقادیر ذکر شده را شرح دهید.
* هدف مجموعه دیتاست داده شده، پیش بینی کردن تعداد خریدهای یک مشتری از فروشگاه می‌باشد که در ستون NumPurchases مقدار واقعی آن قرار داده شده است.
۱۶. پس از تکمیل کردن بخش‌های مشخص شده در نوت بوک، یک مدل رگرسیون مرتبه ۱ ساخته می‌شود. از آنجایی که تابع رگرسیون ساخته شده از مرتبه ۱ است، تنها یک ویژگی را می‌توان به عنوان ورودی این تابع انتخاب نمود. به نظر شما کدام ویژگی نسبت به سایر ویژگی‌ها خروجی دقیق‌تری به ما می‌دهد؟ علت انتخاب خود را توضیح دهید.

۱۷. پس از انتخاب ویژگی مناسب از داده های train و پیش‌بینی داده های آزمون، می‌بایست معیاری برای ارزیابی کارایی خروجی بدست آمده تعیین کنیم. از آنجایی که مدل ما linear regression است و عملیات classification را روی آن انجام نداده‌ایم، نمی‌توان از متدهای ارزیابی کارایی مربوط به classification استفاده کرد. درباره متدهای RMSE, MSE, RSS و R2 score مطالعه کنید و هرکدام را در گزارش خود توضیح دهید.
۱۸. با استفاده از متد RMSE و R2 score، مقادیر پیش بینی شده را ارزیابی کنید. عملیات فوق را بر روی چند ویژگی دیگر نیز انجام دهید. از مقادیر بدست آمده چه استنباطی می‌کنید؟

فاز دوم: Multiple Regression

در این قسمت، رگرسیون را روی چندین ویژگی انجام می‌دهیم. در مرحله قبل، توانستیم با استفاده از دو معادله و دو مجهول به مقادیر بهینه وزن‌ها برسیم. با افزایش تعداد ویژگی‌ها، حل این دستگاه بسیار دشوار می‌شود و نیاز به روش‌ای هست که بتوان مرحله به مرحله به وزن‌های بهینه نزدیک شویم. شما در ادامه نوت‌بوکی که در اختیارتان قرار گرفته است، با استفاده از روش گرادیان کاهشی، یک مدل Multiple Regression می‌سازید. پس از پیاده‌سازی کامل الگوریتم و بخش‌های خواسته شده در نوت‌بوک طبق توضیحات داده شده، مدل را به ازای ۲، ۳ و ۵ ویژگی اجرا کنید. انتخاب ویژگی‌ها دست شماست ولی باید برای انتخاب خود دلیل داشته باشید. دقت مدل جدید را با فاز قبل مقایسه کرده و عملکرد آن را توضیح دهید.

فاز سوم: طبقه‌بندی

در این فاز از پروژه، سه مدل بر پایه Decision Trees، K-Nearest-Neighbours و Logistic Regression با استفاده از کتابخانه scikit learn پیاده‌سازی می‌کنید. سپس فراپارامترها[7] را تغییر دهید و مدل را بهینه کنید. بهینه‌سازی مدل‌ها به این منظور است که تابع هزینه کمینه شود اما overfitting رخ ندهد.
قبل از شروع مدل‌سازی، باید توجه کرد که ستون هدف فعلی در مجموعه داده، قابل استفاده برای یک مسئله طبقه‌بندی نیست؛ پس لازم است که یک ستون هدف جدید، میزان فروش، را ایجاد کنیم. نام این ستون را PurchaseRate گذاشته و به ازای هر مقدار از ستون NumPurchases، اگر از مقدار میانه (median) این ستون بیشتر بود، مقدار متناظر در ستون PurchaseRate را HIGH و در غیر این صورت، LOW قرار دهید.
پس از ایجاد ستون جدید، طبقه‌بندی را بر اساس ستون PurchaseRate انجام می‌دهیم. حال به مدل‌سازی و حل این مسئله می‌پردازیم:
۱۹. دقت هر مدل را بر اساس confusion matrix رسم شده بدست آورید و نتایج را توضیح دهید.
۲۰. برای مدل‌هایی که پارامترهای زیادی دارند با کمک تابع GridSearchCV، مقادیر بهینه برای پارامترها را بدست آورید.
۲۱. در مورد underfitting و overfitting تحقیق کنید. آیا در مدل‌های شما این پدیده‌ها رخ دادند؟
۲۲. سعی کنید برخی از پیش پردازش‌هایی که انجام دادید را تغییر دهید. تاثیر آنها بر دقت مدل‌هایتان را بررسی کنید.
توجه داشته باشید که برای مدل KNN، تغییر تعداد همسایه‌ها کافی‌ست.
۲۳. با استفاده از کتابخانه مناسب، درخت تصمیم نهایی خود را رسم کنید.

[7] Hyperparameters

روش‌های یادگیری جمعی[8]

یادگیری گروهی به این معناست که پیش‌بینی نهایی را با تجمیع نتایج حاصل از چند مدل انجام دهیم. در این فاز به پیاده‌سازی و تحلیل نتایج مدل‌های Random Forest می‌پردازیم. توجه داشته باشید که مدل استفاده در این روش نیز مخصوص طبقه‌بندی بوده و لذا ستون هدف، ستون PurchaseRate خواهد بود.

در این مدل، تعدادی Decision Tree ساخته می‌شود که هر کدام جداگانه و با ویژگی‌های متفاوت آموزش می‌بینند. سپس برای تجمیع نهایی نتایج درخت‌ها، نوعی رای‌گیری انجام می‌شود.

۲۴. در مورد حداقل دو عدد از فراپارامتر این مدل مطالعه کنید و تأثیر تغییر این فراپارامتر را روی نتایجتان را با رسم نمودار و ذکر دقیق نتایج بسنجید.

۲۵. نتایج این مدل را با مدل Decision Tree مقایسه کنید. در مورد bias و variance و ارتباط بین آن‌ها مطالعه کنید. به نظر شما از نظر هر کدام از bias و variance یک مدل، Decision Tree بهتر عمل می‌کند یا یک مدل تجمیعی Random Forest؟ آیا نتایجی که به دست آوردید، با نظرتان مطابقت دارد؟

روش‌های مبتنی بر Differential Privacy

حفظ امنیت و حریم شخصی افراد نقش بزرگی در دنیای کنونی ایفا می‌کند، و با گسترش استفاده از یادگیری ماشین در حوزه های مختلف، امنیت داده ها نیز به مسئله پررنگی تبدیل شده است.

امنیت و حریم شخصی اطلاعات به معنای حق شخصی افراد در چگونگی استفاده از داده های آن‌ها می‌باشد. به این معنا که استفاده از داده های شخصی هر فرد می‌بایست با دریافت اجازه از او استفاده شود و نباید بدون اخذ اجازه، از داده های شخصی افراد استفاده شود. در مجموعه دیتاست داده شده نیز از دیتاهای شخصی افراد استفاده شده است و فرض می‌کنیم که مجوز استفاده از داده های افراد به صورت قانونی اخذ نشده است.

۲۶. یکی از روش های مرسوم جهت گمنام کردن مجموعه داده‌ها و حفظ امنیت شخصی افراد، اضافه کردن نویز به دیتا است. به نظر شما افزودن نویز به داده های موجود در ردیف های مختلف یک مجموعه داده، چه تاثیری بر حفظ امنیت و حریم شخصی (Privacy) افراد دارد؟

۲۷. از روش های مرسوم جهت اضافه کردن نویز به دیتا، می‌توان به نویز لاپلاس و نویز نمایی اشاره کرد. تفاوت این دو روش چیست؟

۲۸. یکی از روش های فوق را انتخاب کرده و اقدام به اضافه کردن نویز به مجموعه دیتاست داده شده نمایید. سپس مراحل مربوط به فاز طبقه بندی با استفاده از کتابخانه را بر روی مجموعه داده جدید که با نویز همراه شده است انجام دهید و نتایج را با قسمت قبل مقایسه کنید.[9]

امتیازی: روش‌های مبتنی بر gradient-boosting

Gradient-boosting یکی از روش های یادگیری ماشین برای مسائل رگرسیون و طبقه‌بندی است که در لکچرهای درس با آن آشنا شده‌اید.

۲۹. با جستجو در منابع مختلف اینترنت، چگونگی کارکرد این متد را توضیح دهید. تفاوت درخت boosting را با decision tree توضیح دهید.

۳۰. XGBoost یکی از جدیدترین روش های یادگیری ماشین بر اساس متد boosting است که در سال 2016 ارائه شده است. با جستجو در منابع اینترنتی، چگونگی کارکرد این درخت را توضیح دهید.

---

[8] Ensemble Learning

[9] جهت آشنایی بیشتر با مبحث امنیت اطلاعات در یادگیری ماشین به این مقاله رجوع کنید: لینک مقاله

۳۱. حال با دانلود و نصب کتابخانه XGBoost از این لینک، اقدام به ساخت مدل با استفاده از درخت XGBoost نمایید. مانند بخش قبل، با استفاده از تابع GridSearchCV، فراپارامترهای بهینه را بدست آورید؛ سپس اقدام به ارزیابی خروجی‌های بدست آمده از مدل فوق نمایید.

نکات پایانی

- توضیحات مربوط به هر بخش از پروژه را بطور خلاصه و در عین حال مفید در گزارش خود ذکر کنید. از ابزارهای تحلیل داده مانند نمودارها استفاده کنید.
- پس از مطالعه کامل و دقیق صورت پروژه، در صورت وجود هرگونه ابهام یا سوال با طراحان پروژه در ارتباط باشید.
- نتایج، گزارش و کدهای خود را در قالب یک فایل فشرده با فرمت AI_CA4_[stdNumber].zip در سامانه ایلرن بارگذاری کنید.
- محتویات پوشه باید شامل فایل jupyter-notebook، خروجی html و فایل‌های مورد نیاز برای اجرای آن باشد. از نمایش درست خروجی‌های مورد نیاز در فایل html مطمئن شوید.
- دقت کنید که نیازی به آپلود مجموعه داده‌ها در سامانه ایلرن نیست.