هوش مصنوعی

تمرین کامپیوتری شماره ۴

مدرس: **دکتر فدایی**

طراحان: علیرضا اربابی، نگار مرادی، محمدطاها فخاریان

مهلت تحویل: جمعه ۲۲ اردیبهشت ۱۴۰۲، ساعت ۲۳:۵۵

مقدمه

هدف این تمرین، آشنایی با روش‌های یادگیری ماشین[1] جهت پیش‌بینی قیمت خانه‌های یک منطقه است. این تمرین از دو فاز تشکیل شده است؛ در فاز اول به ساخت یک مدل Linear Regression به صورت دستی (**بدون استفاده از کتابخانه**) می‌پردازید و در فاز دوم با کمک کتابخانه **Scikit-Learn** اقدام به تخمین سطح قیمت خانه‌ها می‌پردازید.

در فاز اول لازم است که فایل نوت‌بوک قرار داده شده در سایت را دانلود کرده و بخش‌های مشخص شده را کامل نمایید. پیاده‌سازی فاز دوم نیز در ادامه فاز اول و در همان نوت‌بوک انجام می‌شود.

معرفی مجموعه داده

مجموعه داده‌ای که در اختیار شما قرار دارد، شامل اطلاعات مربوط به خانه‌ها و قیمت آنها در یکی از شهرهای ایالت واشنگتن آمریکا در میان سال‌های 2014 و 2015 می‌باشد. با کمک این مجموعه داده بر اساس ویژگی‌های متفاوتی که در ادامه توضیح داده خواهند شد، قیمت خانه‌ها در این منطقه را بررسی خواهید کرد.

بررسی مجموعه داده

در این فاز داده‌های خام را بررسی خواهید کرد. این تجزیه و تحلیل داده‌ها با نام EDA[2] 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[3]، تعداد و نسبت داده‌های از دست رفته[4] را بدست بیاورید.

---

[1] Machine Learning

[2] Exploratory Data Analysis

[3] Feature

[4] Missing

۳. نمودار وابستگی[5] 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 price 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.

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:

۱. 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.

۲. 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.

۳. در ویژگی‌های عددی[6]، normalizing یا standardizing به چه منظور انجام می‌شود؟ در این پروژه نیاز به انجام این کار هست؟

---

[5] Correlation

[6] Numerical

۴. برای استفاده ویژگی‌های دسته‌ای[7]، 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؟

۵. آیا امکان حذف برخی ستون‌ها وجود دارد؟ چرا؟

۶. 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؟

۷. گاهی علاوه بر دو دسته بالا یک دسته سومی هم وجود دارد. در مورد این دسته (validation) 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 کتابخانه می‌پردازید. توجه کنید که در این فاز، **به هیچ عنوان استفاده از کتابخانه های آماده (به جز math) مجاز نمی‌باشد.**

۱. در ابتدای فایل نوت بوک قرار داده شده، فرمول محاسبه پارامترهای α و β برای یک مدل رگرسیون درجه ۱ قرار داده شده است. محاسبات ریاضی فوق را بررسی کنید، و علت بدست آمدن مقادیر ذکر شده برای متغیرهای α و β را شرح دهید.

۲. 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.

۳. 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

[7] Categorical

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.

۴. با استفاده از متد RMSE و R2 score، مقادیر predict شده را ارزیابی کنید. عملیات فوق را بر روی ویژگی‌های yr_built , bathrooms و zipcode نیز انجام دهید. چه استنباطی از مقادیر بدست آمده دارید؟

فاز دوم: 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 فراپارامتر[8]ها 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 -سطح قیمت- را ایجاد کنیم. نام این ستون را price_level گذاشته و به ازای هر مقدار از ستون price، اگر از مقدار میانه (median) این ستون بیشتر بود، مقدار متناظر در ستون price_level را HIGH و در غیر این صورت، LOW قرار دهید.

پس از ایجاد ستون جدید، طبقه‌بندی را بر اساس ستون price_level انجام می‌دهیم. حال به مدل‌سازی و حل این مسئله می‌پردازیم:

۱. دقت هر مدل را بر اساس confusion matrix رسم شده بدست آورید و نتایج را توضیح دهید.

۲. برای مدل‌هایی که پارامترهای زیادی دارند با کمک تابع GridSearchCV[9]، مقادیر بهینه برای پارامترها را بدست آورید.

۳. در مورد underfitting و overfitting تحقیق کنید. آیا در مدل‌های شما این پدیده‌ها رخ دادند؟

---

[8] Hyperparameter

[9] https://medium.datadriveninvestor.com/an-introduction-to-grid-search-ff57adcc0998

۴. 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.

روش‌های یادگیری جمعی[10]

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 price_level 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.

۱. 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.

۲. 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؟

[10] Ensemble Learning

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.

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

۲. XGBoost[11] 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.

۳. 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.

[11] https://scholar.google.com/citations?view_op=view_citation&hl=en&user=DpLFv4gAAAAJ&citation_for_view=DpLFv4gAAAAJ:2KloaMYe4IUC

نکات پایانی

۱. دقت کنید که هدف پروژه تحلیل نتایج است بنابراین از ابزارهای تحلیل داده مانند نمودارها استفاده کنید و توضیحات مربوط به هر بخش از پروژه را به طور خلاصه و در عین حال مفید در گزارش خود ذکر کنید. اگر در جایی ذکر شده مقایسه‌ای انجام دهید، حتما نتایج را دقیق ذکر کنید و سپس آنها را تحلیل و مقایسه کنید.

۲. در همه‌ی قسمت‌ها(به جز پیاده‌سازی فاز اول)، مجازید از متدهای کتابخانه‌ی Scikit-Learn، Seaborn، Matplotlib و Pandas استفاده کنید ولی باید اطلاعات لازم در مورد هر کاری که انجام می‌دهید را داشته باشید؛ در هنگام تحویل ممکن است در مورد هرکدام از شما سوال پرسیده شود.

۳. نتایج و گزارش خود را در یک فایل فشرده با عنوان AI_CA4_<SID>.zip تحویل دهید. محتویات پوشه باید شامل فایل 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 خروجی های مورد نیاز در فایل html مطمئن شوید.