هوش مصنوعی و یادگیری ماشین
تمرین دوم (رگرسیون)
فروردین ۱۴۰۴

هدف این تمرین آشنایی با روش‌های پیش‌پردازش داده‌ها و ارزیابی کارآیی مدل‌های مختلف رگرسیون است.

دادگانی که برای این کار در نظر گرفته شده است (فایل Housing.csv) شامل ۲۹۳۰ داده مربوط به فروش خانه‌های مسکونی در یکی از شهرهای آمریکا در فاصلهٔ سال‌های ۲۰۰۶ و ۲۰۱۰ است که قیمت فروش خانه‌ها را بر پایهٔ ۸۰ ویژگی آن‌ها از جمله مساحت زمین، تعداد اتاق‌ها، سال ساخت، محله، نوع سقف و کیفیت ساخت (ضعیف تا عالی) نشان می‌دهد. هدف ما یافتن بهترین مدلی است که می‌تواند با استفاده از این ویژگی‌ها قیمت خانه را پیش‌بینی کند.

۱- دادگان را به صورت dataframe خوانده و با استفاده از روش info اطلاعات کلی آن (شامل تعداد مقادیر موجود برای هر یک از ویژگی‌ها) را نمایش دهید.

۲- در صورت وجود داده‌های پرت (outlier) آن‌ها را حذف کنید و مقادیر ناموجود (missing value) را (با ذکر روش بکار گرفته شده) با مقادیر مناسب جایگزین کنید.

۳- اطلاعات آماری دادگان را بررسی کنید (مقادیر کمینه، بیشینه و انحراف از معیار را برای دادگان بدست آورید.)

۴- با استفاده از ماتریس همبستگی، ویژگی‌هایی را که بیشترین تاثیر را بر قیمت خانه دارند مشخص کنید (بخش توضیحات را ببینید).

۵- برای مشخص‌تر کردن ویژگی‌هایی (از میان ویژگی‌های انتخاب شده در بند ۴) که بیشترین تاثیر را بر قیمت خانه دارند، با استفاده از کتابخانهٔ seaborn و دستور jointplot، jointplot مربوط به این ویژگی‌ها را رسم کنید.

۶- با استفاده از دستور SelectKBest در کتابخانهٔ scikit-learn، تعداد ویژگی‌ها را به گونه ای انتخاب کنید که مدل‌های رگرسون بیشترین دقت را داشته باشند (از آنجا که هدف این تمرین ارزیابی کارآیی مدل‌های رگرسیون است باید از آزمون f_regression استفاده کنید.)

۷- دادگان را به دو بخش آموزش (training) و آزمون (test) تقسیم کنید. (random_state = 42, test_size = 0.25) (انتخاب عدد ۴۲ دلیل خاصی ندارد و صرفا کمک می‌کند که در همهٔ اجراها عدد تصادفی یکسانی تولید شود تا نتایج این اجراها قابل مقایسه باشند.)

۸- به کمک داده‌های آموزش و با استفاده از کتابخانهٔ scikit-learn مدل‌های Linear Regression، Lasso Regression، Ridge Regression و Polynomial Regression را آموزش دهید.

۹- توضیح دهید که خطای RMS و معیار R^2 score چگونه محاسبه می‌شوند. سپس مقدار آن‌ها را برای هر یک از مدل‌های بالا روی داده‌های آزمون محاسبه و گزارش کنید.

۱۰-توضیح دهید که bias-variance trade-off چیست و چگونه بر عملکرد مدل‌های یادگیری ماشین تأثیر می‌گذارد. سپس با ارایهٔ یک مثال نشان دهید که افزایش پیچیدگی مدل چگونه بر خطاهای بایاس و واریانس تأثیر می‌گذارد.

توضیحات:

ماتریس همبستگی (Correlation Matrix)

این ماتریس که درایه‌های آن ضرایب همبستگی بین متغیرها (در اینجا ویژگی‌های) مختلف است میزان و جهت همبستگی خطی بین این متغیرها را مشخص می‌کند. همبستگی مثبت به این معنی است که با افزایش یک متغیر، دیگری هم افزایش می‌یابد و همبستگی منفی به این معنی است که با افزایش یک متغیر، دیگری کاهش می‌یابد. میزان همبستگی دو متغیر با عددی در بازهٔ ۱- تا ۱+ نشان داده می‌شود. مقدار ۱+ نشان دهنده همبستگی مثبت کامل، مقدار ۱- نشان‌دهنده همبستگی

منفی کامل و مقدار صفر نشان‌دهنده عدم همبستگی دو متغیر است. همبستگی بین ویژگی‌ها از رابطهٔ زیر بدست می‌آید که در آن $\bar{x}$ و $\bar{y}$ میانگین مقادیر دو ویژگی است. x_i و y_i نیز نشان‌دهنده مقادیر دو ویژگی در داده‌های مختلف است.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

رگرسیون‌هاي Ridge وLasso

در کاربردهایی با تعداد ویژگی زیاد می‌توان از رگرسیون‌هاي ریج و لاسو براي ایجاد مدل‌هاي آماري ساده‌تر و قابل اعتمادتر استفاده کرد. رگرسیون Ridge با کنترل تأثیر هر ویژگی بر خروجی مدل و پیشگیري از تاثیر بیش از حد برخی ویژگی‌ها بر خروجی، به افزایش دقت و تعمیم‌پذیري مدل کمک می‌کند ولی همهٔ متغیرها را در مدل نگه می‌دارد

رگرسیون Lasso علاوه بر این کنترل، ویژگی‌هاي کم‌تاثیر بر خروجی را نیز حذف می‌کند و به مدل کمک می‌کند که بر مهم‌ترین ویژگی‌ها تمرکز کند. هر دو روش شامل یک جملهٔ عادي‌ساز (Regularizer) هستند که میزان ساده‌سازي مدل را کنترل می‌کند. این جمله در رگرسیون ریج، نرم ۲ ضرایب وزنی هر ویژگی و در رگرسیون لاسو، نرم ۱ این ضرایب را به تابع هزینه می‌افزاید. براي آشنایی بیشتر با این دو روش می‌توانید به لینک زیر مراجعه کنید.

https://medium.com/@devsachin0879/ridge-regression-and-lasso-regression-a-beginners-guide-b3e33c77678

لینک‌های مفید:

- راهنمایی دستور SelectKBest:
https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

- راهنمایی پیاده‌سازی Linear Regression، Ridge Regression، Lasso Regression، Polynomial Regression
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html

چند تذکر:

- تحویل گزارش این تمرین ضروری است و به تمرین بدون گزارش نمره‌ای تعلق نمی‌گیرد. حجم گزارش معیاری برای ارزیابی نخواهد بود و لزومی به توضیح جزئیات کد نیست؛ اما از آنجا که برای این تمرین از کتابخانه‌های موجود استفاده می‌کنید لطفا تمامی پارامترهای تنظیم‌شده در هر قسمت از کد را گزارش کرده و فرض‌هایی را که برای پیاده‌سازی‌ها و محاسبات خود به کار برده‌اید ذکر کنید. از ارائهٔ توضیحات کلیشه‌ای و همانند بَرداری از منابع موجود بپرهیزید.
- در فرایند ارزیابی گزارش، کدهای شما لزوما اجرا نخواهد شد. بنابراین همهٔ نتایج و تحلیل‌های خود را به‌طور کامل ارائه کنید.
- شباهت بیش از حد گزارش و کدها باعث از دست دادن نمرهٔ تمرین خواهد شد. همچنین گزارش‌هایی که در آنها از کدهای آماده استفاده شده باشد پذیرفته نخواهند شد.
- گزارش شما باید به صورت تایپ شده و با فرمت pdf ارائه شود و کدهایی که به همراه گزارش تحویل می‌دهید باید قابل اجرا باشند. در انتها تمامی فایل‌های لازم را در یک فایل zip یا rar بارگذاری و ارسال کنید.
- در صورت استفاده از گیت هاب جهت ارائهٔ گزارش و کد، نمره امتیازی تعلق میگیرد.

- پرسش‌های خود را از دستیار آموزشی مربوطه (محمد برهانی) بپرسید: (@Borhani_1996)