دانشگاه تهران

دانشکده علوم مهندسی

یادگیری ماشین

استاد: دکتر فهیم

هدف این تمرین، آشنایی با پیاده سازی رگرسیون خطی جهت پیشبینی میزان ترافیک ایستگاه مترو است. این تمرین از سه فاز تشکیل شده است؛ در فاز اول به تجزیه و تحلیل داده ها میپردازید. سپس در فاز دوم به ساخت یک مدل رگرسیون خطی مرتبه اول به صورت دستی (بدون استفاده از مدل آماده) میپردازید، سپس در فاز نهایی با متد گرادیان کاهشی رگرسیون چند متغیره را پیاده سازی میکنید و به تخمین تعداد میزان ترافیک در ایستگاه مترو میپردازید.

بررسی مجموعه داده

در این فاز داده‌های خام را بررسی خواهید کرد. این تجزیه و تحلیل داده‌ها با نام EDA 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.

- ساختار کلی داده‌ها را با متدهای info و describe بدست بیاورید.
- آیا نیاز به normalization و یا standardizing وجود دارد یا خیر؟
- داده های کیفی را به شکل مناسب کمی کنید. از داده‌ی تاریخ چگونه میتوان استفاده کرد؟
- نمودار وابستگی ویژگی‌ها به یکدیگر را رسم کنید. کدام ویژگیها وابستگی بیشتری به ستون هدف دارند؟
- برای آموزش و در نهایت ارزیابی مدل یادگیری ماشین نیاز است که داده‌ها را به سه دسته test و validation وtrain تقسیم کنیم. نسبت این تقسیم به چه صورت است؟ چه روش‌هایی برای تقسیم و ساخت این سه دسته وجود دارد؟ آیا نیاز به دیتای ولیدیشن وجود دارد؟
- درمورد رگیولاتورها تحقیق کنید. دیتای validation چه کمکی در این زمینه میتواند بکند.
- outlier data چیست؟ و با چه روشی میتوان این داده ها را حذف کرد؟

- برای Data Exploitation سه دسته کلی Univariate Analysis و Bivariate Analysis و Multivariate Analysis تقسیم میشوند. در مورد روش ها وپلات های هر کدام توضیح دهید و چند تا را به انتخاب خود پیاده کنید.

رگرسیون خطی

در این فاز از پروژه، به ساخت یک مدل رگرسیون خطی درجه ۱ بدون استفاده از مدل آماده میپردازید .توجه کنید که در این فاز، به هیچ عنوان استفاده از کتابخانه‌های آماده (به جزnumpy و pandas) مجاز نمی‌باشد.

هدف مجموعه دیتاست داده شده، پیش بینی کردن میزان ترافیک در ساعات مختلف روز است که در ستون traffic_volume مقدار واقعی آن قرار داده شده است. بقیه‌ی اطلاعات مربوط به فیچرها در جدول زیر قرار دارند.

Variable Name	Role	Type	Description	Units	Missing Values
holiday	Feature	Categorical	US National holidays plus regional holiday, Minnesota State Fair		no
temp	Feature	Continuous	Average temp in kelvin	Kelvin	no
rain_1h	Feature	Continuous	Amount in mm of rain that occurred in the hour	mm	no
snow_1h	Feature	Continuous	Amount in mm of snow that occurred in the hour	mm	no
clouds_all	Feature	Integer	Percentage of cloud cover	%	no
weather_main	Feature	Categorical	Short textual description of the current weather		no
weather_description	Feature	Categorical	Longer textual description of the current weather		no
date_time	Feature	Date	Hour of the data collected in local CST time		no
traffic_volume	Target	Integer	Hourly I-94 ATR 301 reported westbound traffic volume		no

یک مدل رگرسیون مرتبه ۱ بسازید. 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.

پس از پیشبینی داده های آزمون، میبایست معیاری برای ارزیابی کارایی خروجی بدست آمده تعیین کنیم. از آنجایی که مدل ما linear regression است درباره متدهای RSS و MSE ,RMSE و 2R Score مطالعه کنید و هرکدام را در گزارش خود توضیح دهید.

Outlier data را حذف کنید و یک بار دیگر مدل خود را train کنید. آیا نتایج بهتری بدست می‌آید؟

در کلاس با توابع پایه (basis function) آشنا شدید. در مورد انواع آن تحقیق کنید و چند نوع آن را در گزارش خود توضیح دهید. با توجه به شکل توزیعی داده ها در بخش Data Exploitation، تابع مناسبی را بر روی فیچرهای دیتاست اعمال کنید، مدل را دوباره train کنید و سپس نتایج ارزیابی را با حالت قبل مقایسه کنید.

رگرسیون چند متغیره

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 روش هایی هست که بتوان مرحله به مرحله به وزن‌های بهینه نزدیک شویم. با استفاده از روش گرادیان کاهشی، یک مدل Multiple Regression بسازید. پس از پیاده‌سازی کامل الگوریتم و آموزش مدل خود، مولتیپل رگرشن را برای حالت های 2و 3و 4و ... متغیره امتحان کنید و نمودار RMSE بر حسب تعداد متغیرهای رگرسیون را در گزارش خود رسم کنید.

مانند حالت قلب توابع پایه مناسب را برای هر فیچر انتخاب کنید و روی داده ها اعمال کنید و سپس رگرسیون را train کنید.

مانند حالت قبل نمودار RMSE بر حسب تعداد متغیرهای رگرسیون را در گزارش خود رسم کنید. آیا با اعمال توابع پایه به نتایج بهتری دست پیدا میکنید؟