به نام او ...

تمرین دوم درس داده کاوی

در این تمرین دو مجموعه داده برای حل دو مسئله‌ی رگرسیون در اختیار شما قرار داده می‌شود. هر دو بخش تمرین به صورت رقابتی بوده و نتایج بهتر نمره‌ی بیشتری در پی خواهند داشت.

بخش اول (۵۰ نمره)

مجموعه داده‌ی این قسمت از نمره های تعدادی دانشجو در ۷ درس مختلف تشکیل شده است. هدف آموزش مدلی با استفاده از ۶ نمره ی اول برای پیش بینی نمره ی هفتم می باشد. مراحل زیر را برای این بخش طی نمایید.

۱- ابتدا تعریف کوتاهی برای عبارات زیر ارائه کنید.

- K-fold Cross-Validation
- MSE, MAE, RMSE
- Covariance, Correlation
- Regression toward the mean
- L1 norm, Lasso Regression
- L2 norm, Ridge Regression

۲- نمودار هر یک از ۶ ویژگی نسبت به ویژگی هدف را رسم کرده و با در نظر گرفتن مفهوم Correlation در مورد هر یک از نمودارها به صورت شهودی بحث کنید.

۳- به جای مقادیر نامشخص مجموعه داده، مقدار ۰ قرار داده شده است. راهکاری برای پر کردن مقادیر نامشخص ارائه دهید. هرچه‌قدر این راه حل هوشمندانه‌تر باشد نمره‌ی بیشتری تعلق خواهد گرفت.

۴- از روش‌های زیر و یک روش دلخواه دیگر برای آموزش مدل و پیش‌بینی ویژگی هدف استفاده کنید (نیازی به پیاده سازی روش‌ها به طور مستقیم نیست و می‌توانید از کتابخانه‌هایی مانند Scikit استفاده کنید). خطای هرکدام را با 10-fold Cross-Validation با در نظر گرفتن معیار RMSE گزارش کنید. توجه داشته باشید که تغییر هایپر پارامترهای هر روش می‌تواند تاثیر قابل توجهی در خطا داشته باشد.

- Linear Regression
- Lasso
- Gradient Boosting

برای روش دلخواه دیگر می‌توانید از رگرسیون با استفاده از توابع چندجمله‌ای، رگرسیون با توابع سینوسی و کسینوسی، یا رگرسیون با بسط‌های نمایی استفاده کنید. همچنین می‌توانید از Regularization برای تعیین تنظیم پیچیدگی مدل استفاده نمایید. مقادیر ویژگی هفتم را برای مجموعه داده ی آزمایشی پیش‌بینی کرده و به

همراه ۶ ویژگی دیگر در یک فایل csv به اسم P1_submission.csv ذخیره کنید. کد نهایی باید با خواندن فایل‌های train.csv و test.csv از پوشه ی ۱، فایل خروجی را تولید کند. از هر روش دلخواهی برای پیش بینی مقادیر ویژگی هدف می‌توانید استفاده کنید. در نظر داشته باشید که خطای کم‌تر معادل با نمره‌ی بیش‌تر خواهد بود. بنابراین از ایده‌های خود برای کاهش خطا استفاده کنید. در نظر داشته باشید که می توانید زیرمجموعه‌ای از ویژگی‌ها را برای آموزش مدل استفاده کرده یا ویژگی جدیدی از روی ویژگی های موجود بسازید.

بخش دوم (۵۰ نمره)

در این قسمت یکی از مسابقه های سایت Kaggle در مورد پیش بینی قیمت خانه در نظر گرفته شده است. از این لینک می توانید به صفحه ی مسابقه دسترسی داشته باشید. محدودیت خاصی برای این قسمت اعمال نمی شود. البته انتظار می رود حداقل های زیر در حل مسئله بررسی شود.

- حل مشکل وجود مقادیر نامشخص
- استفاده از 10-fold cross-validation با معیار RMSE برای گزارش خطای مدل های مختلف
- استفاده از حداقل ۳ مدل مختلف برای یافتن بهترین روش
- بررسی زیرمجموعه های مختلف از ویژگی ها برای آموزش مدل

کد نهایی باید با خواندن فایل‌های train.csv و test.csv از پوشه‌ی ۲، مدل انتخاب شده‌ی نهایی را آموزش داده و پیش‌بینی قیمت خانه‌های مجموعه داده‌ی test.csv را در فایل P2_submission.csv با توجه به ساختار فایل sample_submission.csv موجود در پوشه ی ۲ تولید کند.

گزارش:

گزارش بایستی در قالب فایل PDF باشد. لطفا فایل Word نفرستید. در گزارش تحلیل خود را در رابطه با تمام کارهایی که انجام داده اید بیان نمایید. بخش مهمی از کار مربوط به فرآیند رسیدن به نتیجه ی نهایی می باشد. بنابراین گزارش بخش قابل توجهی از نمره را به خود اختصاص می دهد.

فایل گزارش خود را به شکل «Report2_StdNum.pdf» نامگذاری کنید. (مانند Report2_9131081.pdf)

کد:

کد اجرایی می تواند در محیط R یا Python تهیه شود. عدم وجود کد معادل نمره‌ی صفر خواهد بود.

فایل کد خود را به شکل «DM2_P1_StdNum» برای بخش اول و «DM2_P2_StdNum» برای بخش دوم نامگذاری کنید. علاوه بر این مدل نهایی مورد استفاده برای پیش بینی ویژگی هدف را در کدهای جداگانه به نام‌های «DM2_P1_Submit_StdNum» و «DM2_P1_Submit_StdNum» قرار دهید.

بارگذاری:

تمام فایل های مورد نظر را در قالب یک فایل فشرده در سایت درس بارگذاری نمایید.

فایل فشرده را به شکل «DM2_StdNum» نامگذاری کنید. (مانند DM2_9131081)

مهلت ارسال تمرین ساعت ۲۳:۵۵ دقیقه‌ی روز دوشنبه مورخ ۱۶ اسفند می‌باشد.

به ازای هر روز تاخیر در ارسال تمرین، برای دو روز اول **۵** و روز های بعد **۱۰** درصد از نمره‌ی آن از دست خواهد رفت.

هر گونه سوال در مورد تمرین را می‌توانید از طریق ایمیل AUT.DM2017@gmail.com بپرسید.

موفق باشید