مدرس: علیرضا رضوانیان

rezvanlms@gmail.com

تمرین دوم

۱) یکی از مجموعه داده‌های مشخص شده برای بخش اول تمرین ۲ (تشخیص اسپم) را انتخاب کنید، مطلوبست

الف) ابتدا بردار ویژگی را ایجاد کنید، برای این کار می‌توانید از کیسه کلمات به صورت دودویی، وزندار یا نرمال شده یا ترکیبی از ویژگی‌هایی چون طول رشته، وجود ضمیر در متن، تعداد ضمیر در متن، وجود سمبل‌های خاص در متن، تعداد سمبل‌های خاص در متن، وجود شماره تلفن در متن، وجود لینک در متن و وجود کلمه به صورت حروف بزرگ استفاده کنید. امکان ایجاد بردار ویژگی به صورت دلخواه با هر ایده پیشنهادی از جانب شما وجود دارد.

ب) با استفاده از حداقل ۲ الگوریتم (یکی از بین روش‌های Bagging و دیگری از نوع Boosting) با پیکربندی دلخواه، برای ارزیابی به صورت 3-Fold Cross-Validation علاوه بر Confusion Matrix، مقادیر F-measure, Precision, Recall, Error را گزارش و مقایسه کنید.

۲) پیش‌بینی روند تغییرات نرخ ارزهای خارجی را می‌توان توسط الگوریتم‌های یادگیری ماشین نیز انجام داد. در این تمرین، هدف پیش‌بینی کلاس تغییر قیمت دلار برا ساس داده‌های ارائه شده در خصوص نرخ‌های روزانه دلار، درهم، یورو، طلا، نفت و شاخص بورس تو سط برخی از الگوریتم‌های با نظارت یادگیری ما شین ا ست. کلاس تغییر روز بعد نیز یکی از مقادیر ۱+ (افزایشی)، ۱- (کاهشی و صفر (بدون تغییر) است. بدین منظور مراحل زیر را با دقت دنبال کنید:

الف) آماده‌سازی مجموعه داده

مجموعه داده‌ها شامل فایل‌هایی به صورت: نرخ ارز روزانه EURO/USD، نرخ ارز روزانه AED/USD ، نرخ ارز روزانه USD/IRR، شاخص روزانه کل بورس، نرخ روزانه طلا بر پایه دلار و نرخ روزانه نفت خام بر پایه دلار است. در این مرحله می‌توانید حداقل یک متغیر از هر فایل مجموعه داده (به جز متغیر تاریخ) را انتخاب کرده و از آن برای پیش‌بینی تغییرات متغیر دلار استفاده کنید. متغیری که باید مقدار روز بعد آن پیش‌بینی شود، متغیر قیمت دلار ذیل ستون (last price) در فایل USD-IRR است. حال به منظور آماده‌سازی مجموعه داده مورد نیاز برای استفاده در مراحل بعد حداقل اقدامات زیر را انجام دهید:

1. همه تاریخ‌ها را به میلادی تبدیل کرده و حداقل یک متغیر از هر دیتا ست (به جز تاریخ) را مورد ا ستفاده قرار دهید.
2. تبدیل و پاک‌سازی‌های لازم را انجام دهید تا مقادیر تاریخ و عددی به نحو صحیحی بارگذاری و قابل ا ستفاده باشند.
3. مجموعه داده‌های مختلف را براساس متغیر تاریخ روز با یکدیگر ادغام کرده و مجموعه داده اصلی را بسازید.
4. متغیرهای زیر را به مجموعه داده اصلی یکپارچه اضافه کنید:
    a. متغیر تاریخ روز بعد (هر رکورد حاوی تاریخ روز بعد خود نیز باشد)
    b. متغیر گپ (تعداد روز فاصله بین روز فعلی و روز بعدی)
    c. برچسب پیش‌بینی تغییرات هر روز نسبت به روز بعد (راهنمائي: از درصد تغییرات روز بعد نسبت به روز فعلی استفاده کرده و با مقایسه آن با یک مقدار ثابت، کلاس تغییر را پیش‌بینی کنید)

ب) آزمون مدل‌های پایه

در این بخش با ا ستفاده از حداقل سه الگوریتم از بین الگوریتم‌های SVM، naïve Bayes، Decision Tree و KNN کلاس تغییرات قیمت دلار در روز بعد را پیش‌بینی کنید. برای ارزیابی به صورت Holdout مقادیر Recall, Precision, F-measure, Accuracy را به ازای مجموعه Train و Test گزارش و مقایسه کنید.

ج) فعالیت اضافی (دلخواه و جنبه نمره اضافی)

در صورت تمایل می‌توانید با انجام کارهای اضافی زیر کار خود را تکمیل نمایید:

۱. در این تمرین شما کلاس تغییرات را پیش‌بینی کرده‌اید، با استفاده از همین داده‌ها و به کمک الگوریتم‌های رگرسیون یادگیری ماشین، می‌توانید مقدار قیمت روز بعد را نیز پیش‌بینی کنید.
۲. ویژگی‌های جدید نظیر مقادیر قیمت‌ها در ۱ یا ۲ یا n روز قبل را به مجموعه متغیرهای خود اضافه کنید. این مقادیر را می‌توانید از روی مجموعه داده‌های ارائه شده در این تمرین بسازید.
۳. مجموعه ویژگی‌ها را مورد ارزیابی قرارداده و فقط از ویژگی‌های با ارزش در الگوریتم‌های خود استفاده کنید.
۴. پارامترهای مهم در الگوریتم‌های مختلف را مورد بررسی قرار دهید و مقدار بهینه آن‌ها را تعیین کنید.
۵. از الگوریتم‌های دیگری برای پیش‌بینی استفاده کنید.

توجه: هر گونه فعالیت اضافی و موثر از جانب خودتان را در گزارش مشخص کنید تا مورد ارزیابی قرار گیرد.

بازبینی ۲۸ آبان ۱۴۰۲