مبانی یادگیری ماشین - تکلیف سری سوم

مدرس: دکتر حامد ملک

پاییز ۱۴۰۱

ددلاین: ۱۱ آذر - ساعت ۲۳:۵۹

مسائل تحلیلی

۱. فرض کنید که مدلی طراحی شده که بر اساس عکس‌های MRI از یک شخص تشخیص می‌دهد که آیا این فرد سرطان دارد یا خیر. پس از طراحی این مدل و اعمال نتایج تست، مشخص شد که صحت[1] این مدل برابر ۹۳.۳۱ درصد است. مدل طراحی‌شده را ارزیابی کنید و بگویید که آیا این مدل برای استفاده در صنعت پزشکی مناسب است یا خیر.

۲. یک مجموعه داده داریم که تعداد کل سطر داده‌های آن ۱۰۰۰ نمونه است. یکی از مشکلاتی که برای طراحی مدل برای این تعداد داده‌های کم به وجود می‌آید چیست؟ راه‌حل شما برای پیشگیری از این مشکل و آموزش دادن مدل چیست؟

۳. در چه زمان‌هایی از روش standardization و چه زمان‌هایی از normalization استفاده می‌شود؟ آیا اصلاً تفاوتی در این که از کدام روش استفاده کنیم، وجود دارد؟ توضیح دهید.

۴. فرض کنید با استفاده از رگرسیون خطّی برای حل یک مسئله مانند پیش‌بینی قیمت خانه در شهر تهران با استفاده از ویژگی مساحت خانه قیمت آن را پیش‌بینی می‌کنیم. مدلی که به دست می‌آید، دقت نسبتاً خوبی دارد. اگر بخواهیم یک پارامتری مانند منطقه خانه را به بردار ویژگی مسئله اضافه کنیم و مدل را دوباره آموزش دهیم، آیا کار درستی کرده‌ایم؟ توضیح دهید.

---

[1] Accuracy

۵. در مسئله تشخیص قیمت ماشین، یک‌سری پارامتر مانند رنگ، کیلومتر، مدل و سال تولید را داریم. اگر در حل این مسئله برای encode کردن ویژگی رنگ از one-hot encoding استفاده کنیم، آیا تفاوتی با این دارد که به هر کدام از رنگ‌ها یک عدد طبیعی (مانند ۶ برای رنگ قرمز) نسبت دهیم؟ توضیح دهید.

مسائل کدی

بخش مهندسی ویژگی:

در این بخش از تمرین قرار است میزان فروش محصولات یک فروشگاه زنجیره‌ای را در یک ماه پیش‌بینی کنید. برای این کار ابتدا به توضیح دیتاست میپردازیم.

فایل‌های دیتاست:

- **sales_train.csv** - همان داده آموزش است. در این فایل فروش این فروشگاه از ژانویه ۲۰۱۳ تا اکتبر ۲۰۱۵ موجود است.
- **test.csv** - مجموعه‌ی تست. در این فایل در واقع باید به ازای هر جفت کالا و فروشگاه فروش آن کالا در آن فروشگاه را در ماه نهایی پیش‌بینی کنید.
- **items.csv** - اطلاعات افزوده‌ای در مورد کالاها.
- **item_categories.csv** - دسته‌ای که هر کالا در آن قرار میگیرد را نشان می‌دهد.
- **shops.csv**- اطلاعات افزوده‌ای در مورد فروشگاه‌ها.

فیلدهای دیتاست:

- **ID** - یک آی‌دی که نشان‌دهنده‌ی یک جفت فروشگاه و کالا است.
- **shop_id** - آی‌دی هر فروشگاه
- **item_id** - آی‌دی هر کالا
- **item_category_id** - آی‌دی هر دسته از کالاها
- **item_cnt_day** - تعداد فروش در روز - انتظار این است که مقدار ماهانه‌ی این ستون را پیش‌بینی کنید.
- **item_price** - قیمت فعلی یک کالا
- **date** - تاریخ
- **date_block_num** - نشان‌دهنده‌ی ماه در دیتاست است.
- **item_name** - نام یک کالا
- **shop_name** - نام یک فروشگاه
- **item_category_name** - نام هر دسته از کالاها

همانطور که در میان توضیحات دیتاست مشاهده کردید، از شما خواسته‌شده است که میزان فروش هر کالا در هر زیر فروشگاه را در ماه بعدی دیتاست پیش‌بینی کنید. برای این کار می‌توانید ابتدا میزان

item_cnt_dayهر کالا را در هر فروشگاه را در ماه موجود در بخش **test.csv** پیش‌بینی کنید و نهایتا مجموع آنها را برای ماه حساب کنید.

برای این بخش استفاده از مدل‌های آماده مجاز است و می‌توانید از کتابخانه scikit-learn استفاده کنید. ضمنا، با توجه به اینکه انتخاب مدل محدودیتی ندارد، لازم است که نحوه‌ی عملکرد مدل را به صورت کامل در گزارش تمرین ذکر کنید.

دیتاست این بخش از این لینک قابل دسترسی است.

بخش ارزیابی مدل:

هدف از این بخش، ارزیابی مدل پیاده‌سازی‌شده در قسمت قبل برای ارزیابی عملکرد مدل است.

۱. در بخش ابتدایی نیاز است تا با انتخاب معیار ارزیابی[2] مناسب، مدل خود را ارزیابی کنید. دلیل انتخاب معیار ارزیابی خود را شرح دهید و توضیح دهید چرا برای ارزیابی مدل شما معیار مناسبی است؟

۲. منحنی یادگیری[3] را رسم کنید و خروجی آن را تفسیر کنید. آیا نیاز است که مدل مجددا با پارامترهای دیگری تمرین داده‌شود؟ آیا مقادیر داده استفاده‌شده برای **train/validation/test** به درستی تنظیم شده‌بودند؟

حال که با بررسی معیارهای ارزیابی فهمیدیم که مدل تا چه حد خوب کار می‌کند، آیا همچنان می‌توانیم به او اعتماد کنیم؟ :) برای فهمیدن چگونگی تصمیم‌گیری مدل یادگیری ماشین، چندین روش وجود دارد که در این تمرین با روش SHAP آشنا می‌شویم که خوشبختانه کتاب‌خانه بسیار خوبی برای این روش وجود دارد. (توجه کنید که روش SHAP برای ارزیابی نیست و تنها می‌گوید که تصمیم‌گیری مدل چگونه و بر اساس چه فیلدهایی است)

با استفاده از کتابخانه SHAP چگونگی تصمیم‌گیری مدل خود را بررسی کنید و توضیح دهید که آیا می‌توان به مدل اعتماد کرد یا نه؟

کتاب‌خانه SHAP از این لینک قابل دسترسی است.

[2] Evalutaion metric
[3] Learning curve

نکات تمرین

- به سوالات تحلیلی به دقت و کامل پاسخ دهید.
- در صورت هرگونه تقلب نمره صفر برای شما لحاظ می‌گردد.
- استفاده از زبان غیر از پایتون مجاز نیست.
- این تمرین تحویل حضوری ندارد؛ بنابراین نوشتن مستندات بسیار مهم و بخش قابل توجهی از نمره است.

موفق باشید