مبانی یادگیری ماشین - تکلیف سری سوم

مدرس: دکتر ریاحی

پاییز ۱۴۰۲

مهلت: ۱۱ آذر ساعت ۲۳:۵۹

مسائل تحلیلی

۱. ضرورت وجود هرکدام از دسته‌ها در تقسیم‌بندی Training/Validation/Test را توضیح دهید.

۲. اگر از توزیع داده‌های یک dataset مطلع نباشید استفاده از کدام یک از تکنیک‌های standardization و normalization را ترجیح می‌دهید؟ اگر از توزیع کلی داده‌ها مطلع باشید چه معیارهایی را برای انتخاب در نظر می‌گیرید؟

۳. شکل کلی دو نمودار epoch-loss و dataset size-accuracy برای دو بخش training و validation با ذکر دلیل رسم کنید.

۴. با توجه به سوال قبلی چگونه می‌توان با استفاده از این نمودارها overfit مدل جلوگیری کرد؟

مسائل کدی

در این مسئله شما باید با استفاده مجموعه دادگان آموزشی داده شده یک مدل مناسب برای تشخیص دیابت طراحی کنید. با توجه به اینکه هدف اصلی تمرین feature engineering است نیازی به پیاده‌سازی مدل توسط شما نیست و می‌توانید از مدل‌های موجود در کتابخانه‌هایی مانند sklearn استفاده کنید. مجموعه‌دادگان را می‌توانید از طریق این لینک دریافت کنید. همچنین به نکات زیر دقت داشته باشید:

- علت انتخاب مدل و معیار ارزیابی را توضیح دهید.

- مجموعه‌دادگان مورد نظر را باید به شیوه مناسب به بخش‌های آموزشی، اعتبارسنجی و آزمون تقسیم بندی کنید.
- روش‌های مناسب مهندسی ویژگی را روی دادگان و مدل اعمال کنید.
- پس از آموزش مدل دقت آن را محاسبه کنید.
- در پایان نیز ماتریس درهم ریختگی را برای بخش آزمون رسم کنید.

نکات تمرین

- در صورت هرگونه تقلب نمره صفر برای شما لحاظ می‌گردد.
- استفاده از زبان غیر از پایتون مجاز نیست.

پیروز و سربلند باشید