تمرین دوم یادگیری ماشین

توجه:

1. **استفاده از کتابخانه‌های از پیش آماده مجاز است.**
2. **لطفاً علاوه بر ضمیمه کردن کد، نتایج را تحلیل و در فایل گزارش خود ضمیمه کنید.**

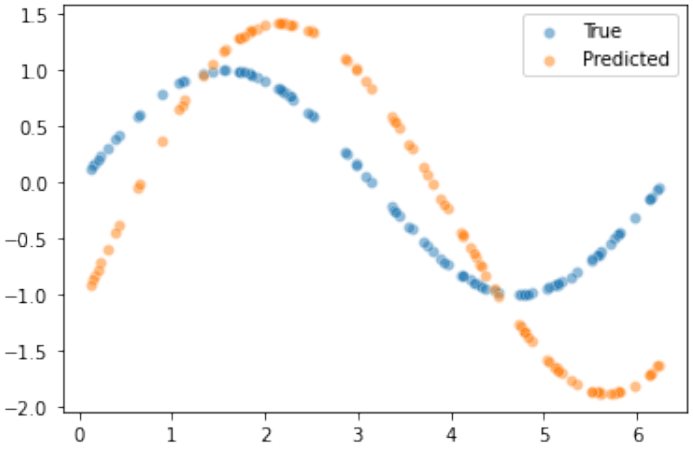
سوال اول

مجموعه داده‌ای برای مسئله رگرسیون خطی در اختیار شما قرار گرفته شده است. داده‌ها به دو بخش آموزشی و تست در پوشه مجموعه داده سوال اول تقسیم شده‌اند. ستون اول متغیر مستقل و ستون دوم متغیر وابسته است.

۱. رگرسیون چندجمله‌ای درجه ۵ را بر روی داده‌ها انجام دهید.

۲. ۵ مدل مستقل به ترتیب با ۱۰، ۲۵، ۵۰، ۱۰۰ و ۲۰۰ نمونه (با حفظ ترتیب داده‌ها) آموزش دهید. خطای MSE بر روی داده آموزش و تست را گزارش دهید. همچنین، مشابه شکل ۱ مقادیر پیش‌بینی شده و واقعی برای داده تست را نمایش دهید.

۳. بر پایه اشکال و نتایج در رابطه با Underfit و Overfit هر مدل صحبت کنید. سپس دلیل برتری برخی مدل‌ها و یا نزدیکی آن‌ها به یکدیگر توضیح دهید.



شکل ۱

سوال دوم

یک مجموعه داده برای مسئله رگرسیون خطی در اختیار شما قرار گرفته شده است. داده‌ها به دو بخش آموزشی و تست در پوشه مجموعه داده سوال دوم تقسیم شده‌اند. در این مجموعه داده، ۱۳ ستون اول متغیرهای مستقل و ستون آخر متغیر وابسته است. معیار ارزیابی را MSE در نظر بگیرید.

۱. ابتدا داده‌ها را بر اساس مجموعه داده آموزش استاندارد سازی کنید (از مجموعه داده تست استفاده نکنید).

$$X_{standard} = \frac{X - E[X]}{\sqrt{Var(X)}}$$

۲. ۳ رگرسیون چند جمله‌ای درجه ۱، درجه ۳ و درجه ۵ را با ۳ مقدار منظم سازی (λ) 0.0، 1 و 10 بررسی کنید (جمعاً ۹ مدل می‌شود). سپس خطا آموزشی و خطا ارزیابی را با استفاده از تکنیک Repeated 5 Fold Cross-Validation با تکرار ۱۰ گزارش دهید. سپس با استفاده از Boxplot نتایج را نمایش دهید (از مجموعه داده تست استفاده نکنید).

۳. نتایج را تحلیل کنید که کدام یک از مدل‌ها Overfit یا Underfit شده‌اند. همچنین با توجه به نتایج، آیا می‌توان تنظیمات بهتری ارائه داد؟ توضیح دهید.

۴. تنظیم بهنیه را تعیین کنید و مدل با تنظیم بهینه را بر روی داده تست ارزیابی کنید. نتیجه را در فایلی با نام prediction.csv ذخیره کنید. توجه کنید که هر سطر بایستی پیش‌بینی سطر متناظر با داده تست باشد.

سوال سوم

به هر کدام از پرسش‌های زیر پاسخ مناسب دهید.

۱. اگر دقت یک مدل بر روی داده آموزشی ۱۰ درصد و بر روی داده تست ۷۰ درصد باشد، چه اظهار نظری می‌توان کرد؟

۲. اگر خطا تست یک مدل پس از تعدادی تکرار افزایش پیدا کند، چه نتیجه‌ای می‌توان گرفت؟ چگونه می‌توان مشکل را حل کرد؟

۳. اگر خطا آموزشی یک مدل حتی با تغییر ابرپارامترها [1] و تعداد تکرار همچنان بالا بماند، چه اظهار نظری می‌توان کرد؟

[1] Hyperparameters