

سوال: «به نظر می‌رسد منحنی یادگیری رگرسیون خطی مشکلی دارد، می‌توانید حدس بزنید چرا؟»
دلیل اصلی این مشکل این است که مسئله پرسیده شده و ماهیت جواب آن به همدیگر نمی‌خورد و با همدیگر تطابق ندارند!
ماهیت مدل رگرسیون است که در واقع برای پیش بینی مقادیر پیوسته خطی به کار گرفته می‌شود.

اما مسئله‌ی ما یک مسئله‌ی کلاسیفیکیشن دودویی است (گره = ۰، سگ = ۱). یعنی جواب یا ۰ است یا ۱ و هیچ پیوستگی در کار نیست.

تابع هزینه در رگرسیون خطی تلاش می‌کند تا خطای MSE را به حداقل برساند.

اما مدل‌های کلاسیفیکیشن مثل دسیژن تری برای بهینه‌سازی طبقه بندی یا دقت طراحی شده‌اند.

معیار ارزیابی در تابع `plot_learning_curve`، دقت است.

بنابراین، ما در حال ارزیابی مدلی هستیم که برای مینیمم کردن خطای مربعات تری شده، اما آن را بر اساس دقت طبقه‌بندی می‌سنجیم.

مشکلی که در نمودار یادگیری رگرسیون خطی مشاهده می‌شود، ناشی از عدم تطابق اساسی بین هدف مدل

که بهینه‌سازی MSE است و هدف تسک که بهینه‌سازی Accuracy است، هست. اگرچه ما با اعمال یک

آستانه ($\text{threshold} > 0.5$) خروجی پیوسته‌ی آن را تبدیل به پیش‌بینی لیبل ۱ و ۰ کرده ایم، اما این کار،

ماهیت اصلی و فرایند یادگیری مدل را تغییر نمی‌دهد. به همین دلیل، منحنی یادگیری آن در مقایسه با

مدل‌هایی که ذاتاً برای طبقه‌بندی طراحی شده‌اند (مانند Random Forest)، ضعیف و نامناسب به نظر

می‌رسد.