

تمرین سری اول درس یادگیری ماشین - رگرسیون غیر خط

در این تمرین، شما باید با استفاده از رگرسیون غیر خطی و داده‌هایی که خودتان تولید کرده‌اید، عملکرد این الگوریتم را بررسی کنید. برای این منظور، برای رسم راحت خروجی‌ها، فرض کنید که یک ویژگی ورودی و یک ویژگی خروجی دارید (مسئله در فضای دو بعدی است). جهت پیاده‌سازی رگرسیون غیر خطی، همانطور که بیان شد، شما باید توان‌های متغیرهای ورودی را نیز به معادلات خود اضافه کنید. همچنین برای این کار، می‌توانید از تابع `sklearn.preprocessing.PolynomialFeatures` استفاده کنید.

برای انجام این تمرین، شما ابتدا باید چهار مجموعه داده مختلف در دست کنید. این چهار مجموعه داده، همگی باید با یک معادله خط به خصوص (پارامترهای خط را خودتان به صورت تصادفی مشخص کنید) و با تعداد داده‌های ۵، ۲۵، ۱۰۰ و ۱۰۰۰ باشند. با این روش، می‌توان اثر تعداد نمونه‌های آموزشی بر مدل نهایی رگرسیون خطی را مطالعه کرد. همچنین به داده‌های آموزشی، مقدار دلخواهی نیز اضافه کنید.

سپس سه مدل مختلف را بر روی هر چهار مجموعه آموزش دهید. این مدل‌ها هر سه باید مدل رگرسیونی باشند و به ترتیب با استفاده از توان چند جمله‌ای ۱، ۴ و ۱۶ آموزش دیده باشند.

در نهایت در کد خود، نمودار تمام خطوط به دست آمده با استفاده از هریک از این روش‌ها بر روی همه دادگان (در مجموع ۱۲ نمودار) را رسم کنید. همچنین یک نمودار برای مطالعه دقت نهایی مدل رسم کنید. در این نمودار، محور افقی باید حجم دادگان باشد و محور عمودی، دقت نهایی به دست آمده از مدل. در این نمودار باید سه خط با سه رنگ مختلف رسم کنید که نشان دهنده هر یک از مدل‌های تولید شده است.

در نهایت با استفاده از این نمودارها و کدهای نوشته شده (و کدهای دیگری که خودتان برای بررسی بیشتر می‌نویسید) با تحلیل خود به سوالات زیر پاسخ دهید و دلایل خود را بیان کنید:

1. به نظر شما برای دادگان خطی، استفاده از رگرسیون غیر خطی مناسب است و دقت کافی دارد؟ در چه شرایطی این روش جواب مناسب نخواهد داشت؟
2. اگر هیچ دانشی در مورد دادگان خود نداشته باشیم، چطور می‌توان توان چند جمله‌ای لازم را بیابیم؟ (فرض کنید دادگان ما دارای تعداد زیادی ویژگی هستند که رسم آنها را غیر ممکن می‌سازد).
3. آیا کاهش خطای مدل در زمان آموزش همواره مطلوب است، آیا امکان دارد مدلی که دقت آموزشی کمتری دارد، در زمان استفاده بدتر عمل کنید؟ چرا؟

در پاسخ به سوالات این نکته را در نظر بگیرید:

در یک استفاده واقعی، مقداری داده جمع‌آوری می‌شود و با استفاده از آنها مدل آموزش می‌بیند و سپس برای استفاده نهایی از مدل استفاده می‌شود. در چنین شرایطی دادگانی که مدل در زمان استفاده مشاهده خواهد کرد، با دادگان آموزشی متفاوت هستند. هدف نهایی ما، یافتن دقت بیشتر در زمان استفاده از مدل است. یعنی می‌خواهیم مدل بر روی دادگانی که تا کنون مشاهده نکرده است، دقت مناسبی داشته باشد.

در نهایت کد خود را در گیت‌هاب قرار داده و آدرس آن را به همراه پاسخ خود به سوالات، برای بنده ایمیل کنید. در صورتی که در کد آپلود شده نمودارها قابل مشاهده نیستند، حتما تصاویر آنها را نیز به پیوست در ایمیل قرار دهید.