به نام خدا

تمرین سری اول درس یادگیری ماشین

پاییز ۱۴۰۴- دانشگاه گیلان

مدرس: یاسمن برشبان

سوال۱: بایاس -واریانس (۱۰ نمره)

الف) با افزایش تعداد داده های آموزشی، واریانس و بایاس مدل یاد گرفته شده چگونه تغییر میکند؟ (۵ نمره)

ب) در مساله رگرسیون، از تابع درجه ۵ برای تخمین استفاده شده است، اما خطای آموزش قابل توجهی داریم. درجه تخمین را باید چگونه تغییر دهیم؟ (۵ نمره)

سوال۲: دسته بندی خطی: (۱۵ نمره)

الف) مشکل استفاده از تابع هزینه SSE در مساله دسته بندی چیست؟ (۷ نمره)

ب) برتری تابع هزینه روش دسته بندی پرسپترون نسبت به تابع هزینه ای که تنها تعداد نمونه های غلط دسته بندی شده را در نظر می گیرد چیست؟ (۸ نمره)

سوال۳: LDA (۱۵نمره)

فرض کنید داده های مساله، نقاط زیر در فضای دو بعدی باشند. راستای بهینه را با استفاده از الگوریتم LDA محاسبه کنید.

$$C_1 = \{(1,1),(1,2),(2,1),(2,4),(3,1),(3,3)\}$$
$$C_2 = \{(2,2),(3,4),(4,2),(5,1),(5,4),(5,5)\}$$

سوال ۴: رگرسیون (۶۰ نمره- پیاده سازی)

در این سوال می‌خواهیم از رگرسیون خطی برای پیش‌بینی هزینه پزشکی افراد بر اساس ویژگی های شخصی آن ها استفاده کنیم. داده ها در دو کلاس train.csv و test.csv ذخیره شده اند. هر داده دارای ۶ ویژگی ورودی (x) است و یک خروجی هزینه پزشکی (y) دارد.

لینک داده

همان طور که در داده های سوال دیده می شود، ویژگی های جنسیت، منطقه و سیگاری بودن از نوع دسته بندی شده هستند (Categorical) و عددی نیست. برای انکود کردن این ویژگی ها روش های متفاوتی وجود دارد. در این تمرین از دو روش استفاده می کنیم:

- integer encoding: به هر دسته، یک عدد یکتا (معمولا بین ۰ تا n-1) اختصاص داده می شود.
- One hot encoding (OHE): برای n دسته در یک ویژگی، n متغییر از نوع باینری در نظر گرفته می شود و برای ووردی که مقدار i ام را دارد، متغیر i ام ۱ می شود و دیگر متغییرها ۰ می شوند.
همانند شکل زیر

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

در داده های سوال، ویژگی ها را به صورت زیر تغییر دهید:
جنسیت: مقادیر femail و male را به ترتیب تبدیل به ۰ و ۱ میکنیم

سیگاری بودن: مقادیر no و yes را به ترتیب تبدیل به ۰و۱ می کنیم.
منطقه: region در این ویژگی ۴ ستون وجود دارد. از روش OHE برای این ویژگی ها استفاده کنید. دلیلی که برای region از OHE استفاده می شود این است که در روش integer encoding، ترتیب بین مقادیر عددی ممکن است خطا ایجاد کند. در حالی که بین مناطق هیچ ترتیبی وجود ندارد اما در OHE این مشکل برطرف می شود.

قسمت الف) رگرسیون بدون منظم سازی:

رگرسیون خطی را بدون جمله منظم ساز پیاده سازی کرده، نتایج را روی داده های تست گزارش کنید. (۱۰ نمره)

ب) در ابتدا تعداد داده های تمرین را ۱۰ در نظر بگیرید و با گامهای ۱۰ تایی ، تا ۱۰۰۰ افزایش دهید. تغییرات خطای تست و آموزش را با افزایش داده آموزش بررسی کنید. برای تابع هزینه از MSE استفاده کنید. نمودار این تغییرات را در گزارش بیاورید. (۱۰ نمره)

ج) روش های batch gradient descentو stochastic gradient descent را پیاده کنید و مقدار w و خطا را در هر حالت به همراه خطا گزارش کنید. (۲۰ نمره)

د) رگرسیون با منظم ساز: ۲۰ نمره

در این قسمت به تابع هزینه SSE جمله منظم ساز L2 را با ضریب α بیفزایید. با استفاده از ۵ fold cross validation بهترین مقدار پارامتر را برای α از بین اعضای مجموعه 0.0001, 0.001, …..,1,……10000 و بیابید سپس نتیجه را بر روی داده های تست بدست آورید.

نمودار خطا را بر حسب لگاریتم α رسم کنید. نتایج نهایی و مقدار خطا را برای داه های تست و آموزش در گزارش بنویسید.

نکته:

۱- فایل گزارش را به همراه تمام کدها در یک فایل ZIP به اسم HW1 به همراه نام، نام خانوادگی و شماره دانشجویی آپلود کنید.
۲- در این تمرین می‌توانید از کتابخانه‌های آماده مانند NumPy، Matplotlibو scikit-learn برای پیاده‌سازی مدل استفاده کنید. هدف تمرین، درک رفتار الگوریتم و تحلیل نتایج مدل است، نه پیاده‌سازی از صفر.
۳- در صورت تشابه پاسخ دو دانشجو، نمره هر دو دانشجو از آن سوال صفر خواهد شد.

موفق باشید