

به نام خدا

تمرین سری دوم درس یادگیری ماشین

پاییز ۱۴۰۴ - دانشگاه گیلان

مدرس: یاسمن برشبان

سوال ۱: تابع زیان (۱۵ نمره)

الف) مزیت‌های استفاده از تابع هزینه‌ی Logistic Regression چیست؟ چرا این تابع نسبت به توابع ساده‌تر مناسب‌تر است؟

ب) مزیت‌های استفاده از Hinge Loss در روش‌های مبتنی بر SVM چیست؟ چرا SVM در بسیاری از شرایط عملکرد بهتری از مدل‌های خطی دیگر دارد؟ چه ویژگی‌هایی این تابع هزینه را برای دسته‌بندی قدرتمند می‌کند؟

سوال ۲: SVM: (۱۵ نمره)

الف) توضیح دهید در رابطه Soft Margin هر یک از حالات زیر به چه معنا است؟

$$(\xi_i = 0, \xi_i = 1, 0 < \xi_i < 1, \xi_i > 1)$$

ب) تاثیر پارامتر C در مساله Soft Margin به چه صورت است؟

ج) چرا کرنل در فرم دوگان SVM بسیار کارا عمل می‌کند؟

سوال ۳: logistic regression (۳۵ نمره - پیاده سازی)

هدف ما در این مساله مکان یابی تلفن همراه با استفاده از قدرت سیگنال های Wi-Fi دریافتی از چند Router در داخل یک ساختمان است. فرض کنید در یک ساختمان، هفت عدد router وجود دارد که در هر نقطه، از هر router، سیگنال با قدرت مختلفی به تلفن همراه می‌رسد. این ساختمان به چهار ناحیه تقسیم شده و در نقاط

مختلف هر ناحیه قدرت سیگنال دریافتی اندازه گیری شده است. نتایج این آزمایش ها در فایل wifi_data.txt ذخیره شده که در اختیار شما قرار داده شده است. در این فایل در هر سطر، مقدار قدرت سیگنال Wifi دریافتی از روتر های مختلف و عدد آخر هر سطر نیز شماره آن ناحیه است.

لینک داده

الف) داده های مربوط به ناحیه های ۱ و ۳ را در نظر بگیرید و برای آن یک دسته بند Logistic Regression پیاده سازی کنید. به این منظور ابتدا داده های مربوط به ناحیه ۱ و ۳ را در نظر بگیرید و ترتیب آن ها را به صورت تصادفی تغییر دهید. سپس 80% اول داده ها را به عنوان داده آموزش و 20% باقیمانده را به عنوان داده های تست در نظر بگیرید. یک دسته بند Logistic regression بر روی این داده ها پیاده سازی کنید و دقت دسته بندی و همچنین ماتریس confusion را در گزارش خود ارائه کنید.

ب) قسمت قبل را با استفاده از جمله منظم ساز $\|W\|^2$ در تابع هزینه و با ضرایب 0.5، 1 و 2 برای منظم ساز تکرار کنید. نتایج حاصل را با نتیجه قسمت قبل مقایسه کنید.

ج) دسته بند Logistic Regression را برای دسته بندی همه ناحیه ها (چهار دسته) پیاده سازی کنید. دقت دسته بند و نیز ماتریس confusion را در گزارش خود ارائه کنید.

سوال ۴ SVM (پیاده سازی ۳۵ نمره)

در این تمرین قصد داریم نحوه عملکرد دسته بند SVM را بررسی کنیم. برای این منظور از مجموعه دادگان MNIST که مجموعه ای از اعداد دست نویس انگلیسی است استفاده میکنیم.

الف) ابتدا مجموعه داده را لود کنید و تعدادی نمونه نمایش دهید.

ب) در این قسمت میخواهیم مجموعه اعداد ۳ و ۴ را دسته بندی کنیم. این دو کلاس را از بین مجموعه دادگان جدا کرده و سعی کنید دسته بندی را با استفاده از Hard Margin SVM انجام دهید. درصد خطای دسته بند را بر روی داده های آموزش و تست گزارش کنید.

ج) حال با استفاده از Soft Margin SVM سعی کنید دسته بند منعطف تری را آموزش دهید. خطا را روی داده های تست و آموزش گزارش کنید. پارامتر C مربوط به دسته بند را 0.1 در نظر بگیرید.

د) در قسمت قبل با استفاده از 5 fold cross validation مقدار مناسبی برای پارامتر C بیابید.

ه) تاثیر استفاده از کرنل گاوسی را در این مساله بررسی کنید. با استفاده از 5 fold cross validation مقدار اولیه مناسبی برای پارامتر کرنل گاوسی پیشنهاد دهید. خطا را بر روی داده های تست و آموزش برای بهترین پارامتر ذکر کنید.

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 می توانید از کتابخانه های آماده مانند NumPy، Matplotlib و scikit-learn برای پیاده سازی مدل استفاده کنید. هدف تمرین، درک رفتار الگوریتم و تحلیل نتایج مدل است، نه پیاده سازی از صفر.

۳- در صورت تشابه پاسخ دو دانشجو، نمره هر دو دانشجو از آن سوال صفر خواهد شد.

موفق باشید