مبانی یادگیری ماشین

نیم‌سال دوم ۰۲-۰۱

دانشکده‌ی مهندسی کامپیوتر

مدرس: دکتر حامد ملک

تمرین سری دوم

موعد تحویل: ۱۴۰۱/۱۲/۲۰

مسائل تئوری

مسئله‌ی ۱.

چرا در رگرسیون لجستیک از تابع هزینه آنتروپی متقاطع[1] به جای خطای میانگین مربعات[2] استفاده می‌شود؟

مسئله‌ی ۲.

تا کنون مسائل طبقه‌بندی دودویی[3] را آموخته‌اید و می‌دانید که کلاس نهایی داده بر اساس احتمال بیشتر یک کلاس انتخاب می‌شد. اکنون در این سوال از شما خواسته می‌شود در رابطه با مسئله طبقه‌بندی چندکلاسه [4] تحقیق کنید و در رابطه با تکنیک یک در مقابل همه [5] بنویسید.

مسئله‌ی ۳.

یک دیتاست چند متغیره را درنظر بگیرید، فرض کنید مشاهده کرده‌ایم که یکی از ضرایب محاسبه‌شده در عملیات رگرسیون خطی مقدار خیلی بزرگ منفی نسبت به باقی متغیرها پیدا کرده است کدام یک از گزاره های زیر صحیح است؟ توضیح دهید.

- این ویژگی تاثیر زیادی روی مدل دارد و باید حفظ شود.
- این ویژگی تاثیر زیادی روی مدل ندارد و باید در نظر گرفته نشود.
- نمی توان بدون در دست داشتن اطلاعات بیشتر درمورد این ویژگی نظر داد.

مسئله‌ی ۴.

با ارائه دلیل صحیح یا غلط بودن هر یک از گزاره‌های زیر را ثابت کنید.

- اگر بایاس زیاد است اضافه کردن تعداد داده های آموزش کمک زیادی به کم کردن بایاس نمی کند.

[1] Cross Entropy
[2] Mean squared error (MSE)
[3] Binary Classification
[4] Multiclass Classification
[5] One-vs-rest

- کم کردن خطای مدل روی داده های آموزش منجر به کاهش خطای مدل روی داده های تست می شود.
- افزایش پیچیدگی مدل رگرسیون همواره منجر به کاهش خطای مدل روی داده ی آموزش و افزایش خطای مدل روی داده ی تست می شود.

مسائل کدی

پروژه‌ی ۱.

مجموعه داده `Car_prices_classification.csv` شامل اطلاعات ۱۱۸ هزار ماشین است که مجموعا به پنج دسته قیمت تقسیم شده اند. در این مسئله مراحل زیر را انجام دهید:

الف) عملیات پیش‌پردازش[۶] را با توجه به هدف مسئله انجام دهید.

ب) داده‌ها را با نسبت مناسب به داده آموزشی و داده تست تقسیم کنید. (دلیل انتخاب درصد نسبت داده آموزشی و تست را گزارش کنید.)

پ) مدل رگرسیون لجستیک را با توجه به مباحث تدریس شده در کلاس درس از پایه پیاده سازی کنید (چرخ را دوباره اختراع کنید:)) و درصد دقت مدل را روی داده تست گزارش کنید.

ت) مدل رگرسیون لجستیک را به کمک کتابخانه scikit-learn پیاده‌سازی کنید و مجددا درصد دقت مدل را روی داده تست گزارش کنید.

ث) در پایان نمودار کاهش خطا را ترسیم نمایید. همچنین نمودار افزایش دقت[۷] و دقت نهایی را گزارش کنید.

پروژه‌ی ۲.

الگوریتم پِرْسِپترون [۸] را برای یک شبکه عصبی با تعداد m نورون پیاده‌سازی کنید. این شبکه را با استفاده از مجموعه داده Iris [۹] تعلیم دهید. این داده به صورت کلی دارای سه کلاس است. مسئله را با یکسان کردن برچسب کلاس برای نمونه‌های دو کلاس versicolor و setosa به یک مسئله دو کلاس تبدیل کنید. از سوی دیگر برای تسهیل ترسیم، تنها دو ویژگی طول کاسبرگ و عرض کاسبرگ از چهار ویژگی را در نظر بگیرید و نمونه‌ها را رسم کنید. برای هر کلاس یک رنگ لحاظ کنید، مرز تصمیم نورون تعلیم دیده را به همراه بردار وزن به‌دست آمده را در کنار نمونه‌ها ترسیم کنید. نمودار تغییرات خطا را رسم کرده و مقدار خطای نهایی را ارائه کنید.

نکات تمرین

- در این تمرین هدف این است که تا حد امکان از کتاب‌خانه‌های Numpy، Pandas و Matplotlib استفاده کنید. عدم استفاده از این کتاب‌خانه‌ها و استفاده از pure python منجر به کسر نمره خواهد شد.
- در صورت هرگونه تقلب نمره صفر برای شما لحاظ می‌گردد.

---

[6] Preprocessing
[7] Accuracy
[8] Prerceptron
[۹] قابل دسترسی با استفاده از `sklearn.datasets.load_iris`

- استفاده از زبان غیر از پایتون مجاز نیست.
- تمرین تحویل حضوری خواهد داشت.
- پیوند عمومی نوت‌بوک خود را در سامانه کوئرا بارگذاری نمایید.