تکلیف شماره ۳ یادگیری ماشین: یادگیری لاجیستیک رگرسیون- پیاده‌سازی عملی

در این بخش با دادگان داده شده باید الگوریتم لاجیستیک رگرسیون را برای طبقه‌بندی بین ۲ کلاس پیاده‌سازی کنید و آن را با نایبو بیز مقایسه کنید. برای پیاده‌سازی می‌توانید از نرم‌افزار MATLAB یا پایتون استفاده کنید.

دادگان داده شده از مجموعه Reuters-21578 که برای طبقه‌بندی متن است استخراج شده است. در مجموعه اصلی ۱۰۰ کلاس متنی وجود دارد، ولی ما برای سادگی فقط ۲ کلاس آن را جدا کرده‌ایم. هدف آن یک طبقه‌بندی باینری برای پیش‌بینی نوع کلاس متن است. در این دادگان از bag of words بعنوان ویژگی استفاده شده است، یعنی تعداد دفعاتی که هر کلمه در متن استفاده شده است. در فولدر مربوط، ۲ مجموعه train و test با مشخصات زیر وجود دارند:

train.csv: Training data. Each row represents a document, each column represents features (word counts). There are 4527 documents and 5180 words.
train labels.txt: labels for the training data (0 or 1)
test.csv: Test data, 1806 documents and 5180 words
test labels.txt: labels for the test data (0 or 1)
word indices: words corresponding to the feature indices.

حال مراحل زیر را انجام دهید:

قسمت الف)

۱- الگوریتم لاجیستیک رگرسیون را با الگوریتم gradient ascent پیاده کنید. نرخ یادگیری مناسب و نقطه توقف آموزش را با استفاده از یک مجموعه validation که بصورت تصادفی به اندازه حداکثر ۲۰ درصد از دادگان آموزشی انتخاب شده است، پیدا کنید. سپس منحنی میزان صحت روی دادگان آموزشی و تست را بکشید. همچنین مدت زمان آموزش را گزارش کنید.

۲- با همان داده‌های مرحله (۱)، یک الگوریتم NB اجرا کنید و نتایج را با (۱) مقایسه کنید. برای جلوگیری از احتمال صفر، توزیع پیشین را یک توزیع بتا با پارامترهای یکسان درنظر بگیرید. (راهنمایی: حتما از لگاریتم استفاده کنید.) همچنین مدت زمان آموزش را گزارش کنید.

۳- الگوریتم لاجیستیک رگرسیون مرحله (۱) را با فرم رگولاریزاسیون با انتخاب پارامتر مناسب اجرا کنید، و آن را با هر دو حالت قبلی مقایسه کنید. همچنین مدت زمان آموزش را گزارش کنید.

۴- برای هر دو حالت NB و لاجیستیک رگرسیون (بهترین مقدار حاصل)، منحنی آموزش و تست را براساس نمونه‌های مختلف آموزش بکشید (میزان صحت طبقه‌بندی براساس تعداد نمونه‌های آموزشی). برای این کار میزان داده‌های آموزشی را از مینیمم ۵۰ شروع کنید، و گام افزایش را هم ۵۰ بگیرید.

قسمت ب)

۱- آموزش لاجیستیک رگرسیون مرحله ۱-الف را بصورت k-fold cross validation با k=3 انجام دهید. سپس روی کل دادگان تست آن را تست کنید و درنهایت متوسط صحت طبقه‌بندی را روی دادگان آموزش

و تست گزارش کنید. این عمل تقسیم تصادفی داده‌ها برای عمل آموزش را ۳ بار تکرار کنید و در هر بار مقادیر صحت را مشخص کنید، و درنهایت هم میزان متوسط صحت را بدست آورید. تمام نتایج را با NB و مرحله ۱-الف مقایسه کنید.

۲- آموزش لاجیستیک رگرسیون مرحله ۱-الف را بصورت k-fold cross validation با k=4 انجام دهید. سپس روی کل دادگان تست آن را تست کنید و درنهایت متوسط صحت طبقه‌بندی را روی دادگان آموزش و تست گزارش کنید. این عمل تقسیم تصادفی داده‌ها برای عمل آموزش را ۳ بار تکرار کنید و در هر بار مقادیر صحت را مشخص کنید، و درنهایت هم میزان متوسط صحت را بدست آورید. تمام نتایج را با NB و مرحله ۱-الف و مرحله قبل مقایسه کنید.

۳- دو مرحله قبل را این بار برای فرم رگولاریزه شده لاجیستیک رگرسیون انجام دهید. تمام نتایج را با NB و مرحله ۱-الف و ۳-الف و مراحل قبلی قسمت (ب) مقایسه کنید.

گزارش: برنامه‌های نوشته شده به همراه نتایج و شکل‌ها و آنالیزهای خواسته شده در هر مرحله را ضمیمه کنید. همچنین یک بحث و بررسی در مورد نتایج و اثر overfitting به دادگان آموزشی و مقایسه بین NB و لاجیستیک رگرسیون در سایز مختلف دادگان آموزشی انجام دهید (سرعت و دقت). گزارش شامل تحلیل و مقایسات و توضیحات کافی در مورد نحوه پیاده‌سازی‌ها، لیست برنامه‌ها و پارامترهای بکار رفته می‌باشد. لطفا از نوشتن گزارش‌های طولانی اجتناب کنید.

توجه: کل تمرین و گزارش باید بصورت انفرادی نوشته شود، و به برنامه‌ها و گزارش‌های کپی شده نمره‌ای تعلق نمی‌گیرد.

نمره مثبت اضافی: تحلیل کامل نتایج، مقایسه با لاجیستیک رگرسیون در حالتی که پارامترهایش با NB بدست می‌آیند، نمایش منحنی لاجیستیک رگرسیون در فرم رگولاریزه شده با مقادیر مختلف پارامتر رگولاریزاسیون و دلیل انتخاب بهترین مقدار از روی منحنی.

موعد تحویل: ۱۰ مرداد ۱۳۹۹
روش تحویل: آپلود در courses

موفق باشید
سیدین