



درس یادگیری ماشین

مدرس: دکتر سامان هراتی زاده

نیم سال اول ۱۴۰۱-۱۴۰۰

الگوریتم های Regression

تحويل حضوری: اعلام خواهد شد

تمرین شماره ی دو

مهلت ارسال تمرین : ۱۴۰۰/۹/۱۸

پیشینی قیمت خانه

هدف از این تمرین، پیاده سازی برنامه ای برای تخمین داده ها بر مبنای الگوریتم رگرسیون خطی است.

داده ها

داده پیوست شده مربوط به میانگین قیمت خانه در مناطق مختلف کالیفرنیا می باشد. این داده شامل ۸ ویژگی عددی و یک ویژگی اسمی است. تعداد نمونه ها ۲۰۶۴۰ عدد است. و قیمت خانه ها در ستون median_house_value می باشد.

الف) برای پیاده سازی الگوریتم رگرسیون ابتدا باید داده ها را آماده کنیم.

- ابتدا به مقادیر ناموجود (missing value) می پردازیم، روش های مختلفی برای این داده ها موجود است، مثل حذف آن ویژگی، حذف نمونه های دارای مقادیر ناموجود و... در این تمرین مقادیر ناموجود را با میانگین باقی داده ها جایگزین کنید.
- همانطور که می دانید الگوریتم رگرسیون با مقادیر عددی کار می کند، برای تبدیل مقادیر اسمی به عددی از روش - Dummy variable Indicator استفاده کنید. در این روش به تعداد مقادیر مختلف در آن ویژگی متغیر جدید اضافه می شود و مقدار باینری ۰ یا ۱ تعلق یا عدم تعلق به آن دسته را نشان می دهد. در این تمرین ویژگی اسمی شامل ۵ دسته می باشد. یعنی ابتدا به ازای هر نمونه یک بردار به طول ۵ تولید که تنها یک عدد یک در آن وجود دارد. در نهایت مقادیر مربوط به دسته "INLAND" را حذف کنید. (برای اطلاعات بیشتر در مورد حذف مقادیر مربوط به یک دسته می توانید عبارت "Dummy Variable Trap" را جست و جو کنید.)
- ویژگی های عددی در بازه های مختلفی قرار دارند، مثلاً سن ساختمان ها حدوداً در بازه ۱-۵۰ سال هستند و جمعیت مناطق حدوداً در بازه ۱-۳۵۰۰۰، بازه ویژگی ها را به بازه ۰ تا ۱ تبدیل کنید. (این کار علاوه بر قابل مقایسه کردن ویژگی ها به اجرای بهتر الگوریتم^۱ GD کمک می کند.)

ب) داده ها را به صورت تصادفی به سه قسمت به ترتیب ۶۰٪ - ۲۰٪ - ۲۰٪ برای آموزش (train)، اعتبارسنجی (validation) و تست (test) تقسیم کنید.

¹ Gradient Descent

پ) یکی از روش های جلوگیری از بیش برازش (overfit) شدن استفاده از داده اعتبارسنجی و روش early stopping است. برای این کار در حین آموزش مقدار تابع هزینه را در داده های اعتبارسنجی رصد می کنیم. تا هنگامی که هزینه در داده های اعتبارسنجی همراه با داده های آموزش در حال کاهش باشد مشکلی نیست اما هنگامی که هزینه در آموزش در حال کاهش باشد اما در اعتبارسنجی روند افزایشی پیدا کند یعنی مدل اطلاعاتی را یاد می گیرد که تنها در داده های آموزش وجود دارد و در باقی داده ها بی ارزش است. (دقت کنید که هزینه در داده های اعتبارسنجی ممکن است نوساناتی داشته باشد و منظور تغییر روند است.)

ت) کاربرد دیگر داده های اعتبار سنجی تنظیم پارامتر ها است، مدل پارامتر هایی دارد که باید توسط ما تنظیم شوند، مثل نرخ یادگیری و ضریب رگولاریزیشن. برای یافتن پارامتر های بهینه نتایج مدل را با مقادیر مختلف پارامتر ها در داده های اعتبار سنجی بررسی می کنیم و بهترین پارامتر ها را انتخاب می کنیم.

ث) الگوریتم رگرسیون خطی را با رگولاریزیشن و early stopping پیاده سازی نمایید.

- موارد خواسته شده را برای ۵ ضریب یادگیری ۰/۱ و ۰/۳ و ۰/۵ و ۰/۷ و ۱ انجام دهید.
- موارد خواسته شده را برای ۴ ضریب رگولاریزیشن ۰ و ۰/۱ و ۱ و ۱۰ انجام دهید.
- برای هر بار اجرای الگوریتم نموداری از مقادیر هزینه در داده های آموزش و اعتبارسنجی را در یک نمودار رسم کنید. (جمعا ۲۰ نمودار)

ج) در نهایت بهترین پارامتر ها را انتخاب کنید و مقدار هزینه در داده های تست را با این پارامتر ها گزارش کنید.

چ) هدف ما پیشبینی میانگین قیمت خانه در مناطق مختلف بود، به داده ها دقت کنید، اطلاعاتی مختلفی موجود است اما آیا این اطلاعات دید کافی به ما برای پیشبینی قیمت خانه می دهند؟ به طور معمول برای قیمت گذاری یک خانه از چه اطلاعاتی استفاده می شود؟

ح) می خواهیم از داده ها اطلاعات بیشتری بیرون بکشیم، به قبل از هم بازه کردن اطلاعات بازگردید، دو ویژگی زیر را محاسبه کنید و برای هر نمونه اضافه کنید. سپس داده ها را هم بازه کنید. (قسمت آخر مورد الف)

$$\text{population_per_household} = \frac{\text{population}}{\text{households}}$$

• میانگین جمعیت در هر خانه

$$\text{rooms_per_household} = \frac{\text{total_rooms}}{\text{households}}$$

• میانگین تعداد اتاق ها در هر خانه

نمودار روند آموزش (تابع هزینه برای داده های آموزش و اعتبارسنجی) را با بهترین پارامترهای قبلی (نرخ آموزش و ضریب رگولاریزیشن) رسم کنید و مقدار تابع هزینه را برای داده های تست اعلام کنید.

تائید اعتبار اسکناس

هدف از این تمرین، پیاده‌سازی برنامه‌ای برای دسته‌بندی داده‌ها بر مبنای الگوریتم رگرسیون لاجستیک است.

داده‌ها

داده پیوست شده مربوط به تائید اعتبار (جعلی یا اصلی بودن) است. این داده‌ها دارای ۴ ویژگی و برچسب کلاس است. ۱۳۷۲ نمونه موجود است. تمام ۱ ویژگی عددی هستند. مقدار ۱ برای برچسب کلاس نشان‌دهنده جعلی بودن اسکناس و مقدار ۰ نشان‌دهنده اصلی بودن اسکناس است.

در پیاده‌سازی‌های خود موارد زیر را لحاظ کنید:

- با استفاده از الگوریتم رگرسیون لاجستیک مدلی بسازید که برچسب کلاس را پیش‌بینی کند.
- میزان دقت خود را روی داده‌های آزمون با استفاده از معیارهای F-score, accuracy, precision, recall گزارش کنید.
- از 5-fold-crossvalidation استفاده کنید. (نتایج حاصل از ۵ بار اجرا و میانگین آن‌ها ذکر شود)
- با تغییر ضریب یادگیری در بازه $(0,1]$ ، تاثیر ضریب یادگیری بر سرعت همگرایی را بررسی کنید. برای این کار نموداری بکشید که تعداد تکرار الگوریتم را بر حسب ضریب یادگیری نمایش دهد. (توجه داشته باشید که در این حالت برای تمامی ضرایب یادگیری شرط توقف را یک دلتای ثابت بایستی در نظر بگیرید.)
- با تغییر ضریب یادگیری در بازه $(0,1]$ ، تاثیر ضریب یادگیری بر F-score, accuracy, precision, recall را بررسی کنید. برای این کار نموداری بکشید که میانگین موارد خواسته شده را بر حسب ضریب یادگیری نمایش دهد. (توجه داشته باشید که در این حالت برای تمامی ضرایب یادگیری شرط توقف را یک دلتای ثابت بایستی در نظر بگیرید.)
- موارد خواسته شده را برای ۵ ضریب یادگیری $۰/۱$ و $۰/۳$ و $۰/۵$ و $۰/۷$ و $۰/۹$ انجام دهید. (ضریب رگولاریزیشن را مقدار ثابت ۱ قرار دهید.)

گزارش شما باید شامل توضیح الگوریتم و شرح مختصری از قسمت های مختلف کد شما باشد. دقت داشته باشید که بخشی از نمره شما به گزارش شما تعلق می گیرد، بنابراین زمان کافی به آن اختصاص دهید.

- برای پیاده سازی می توانید از زبان پایتون و کتابخانه های math، SciPy، NumPy، pandas و یا کتابخانه های مشابه استفاده کنید.

- کد خود را کامنت گذاری کنید. برای کامنت گذاری مناسب و تمیز بودن کد امتیاز مثبت در نظر گرفته خواهد شد.

- استفاده از گیتهاب و کتابخانه های مرتبط با الگوریتم مجاز نیست، در صورت استفاده نمره های تعلق نخواهد گرفت.

- به کدهای مشابه نمره ای تعلق نمی گیرد.

- برای تحویل فایل های مربوطه را در یک فایل زیپ با نام ML_StudentID_FullName_HW# قرار دهید.

- در صورت تاخیر تا ۲۴ ساعت ۲۰٪ تا ۴۸ ساعت ۴۵٪ و بیش از ۴۸ ساعت تمامی نمره تمرین را از دست خواهید داد.

- در صورت وجود هرگونه سوال یا ابهام می توانید با دستیاران آموزشی از طریق ایمیل یا گروه WhatsApp در ارتباط باشید.