درس یادگیری ماشین

مدرس: دکتر سامان هراتی‌زاده

نیم‌سال اول ۱۴۰۱-۱۴۰۰

تمرین شماره‌ی شش

الگوریتم‌های *Regression*

مهلت ارسال تمرین : ۱۴۰۰/۹/۱۸

تحویل حضوری: اعلام خواهد شد

پیشبینی قیمت خانه

هدف از این تمرین، پیاده‌سازي برنامه‌اي برای تخمین داده‌ها بر مبناي الگوریتم رگرسیون خطی است.

داده‌ها

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 خانه‌ها در ستون median_house_value می باشد.

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

- ابتدا به مقادیر ناموجود(missing value) می‌پردازیم، روش های مختلفی برای این داده‌ها موجود است، مثل حذف آن ویژگی، حذف نمونه های دارای مقادیر ناموجود و... در این تمرین مقادیر ناموجود را با میانگین باقی داده‌ها جایگزین کنید.
- همانطور که می‌دانید الگوریتم رگرسیون با مقادیر عددی کار می‌کند، 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 می‌شود 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 می‌توانید عبارت "Dummy Variable Trap" را جست و جو کنید.)
- ویژگی‌های عددی در بازه های مختلفی قرار دارند، مثلا سن ساختمان ها حدودا در بازه ۱-۵۰ سال هستند و جمعیت مناطق حدودا در بازه ۱-۳۵۰۰۰، بازه ویژگی‌ها را به بازه ۰ تا ۱ تبدیل کنید. (این کار علاوه بر قابل مقایسه کردن ویژگی ها به اجرای بهتر الگوریتم GD[1] کمک می‌کند.)

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 داده‌های اعتبارسنجی رصد می‌کنیم. تا هنگامی که هزینه در داده های اعتبارسنجی همراه با داده‌های 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 می‌گیرد که تنها در داده‌های آموزش وجود دارد و در باقی داده‌ها بی‌ارزش است. (دقت کنید که هزینه در داده‌های اعتبارسنجی ممکن است نوساناتی داشته باشد و منظور تغییر روند است.)

ت) کاربرد دیگر داده‌های 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 داده‌های اعتبار سنجی بررسی می‌کنیم و بهترین پارامتر ها را انتخاب میکنیم.

ث) الگوریتم رگرسیون خطی را با رگولاریزیشن و early stopping پیاده‌سازی 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 داده‌های 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 داده‌ها دقت کنید، اطلاعاتی مختلفی موجود است اما آیا این اطلاعات دید کافی به ما برای پیشبینی قیمت خانه می‌دهند؟ به طور معمول برای قیمت گذاری یک خانه از چه اطلاعاتی استفاده می‌شود؟

ح) می‌خواهیم از داده‌ها 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)

- میانگین جمعیت در هر خانه $\text{population\_}per\_household = \frac{population}{households}$
- میانگین تعداد اتاق ها در هر خانه $rooms\_per\_household = \frac{total\_rooms}{households}$

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

تایید اعتبار اسکناس

هدف از اين تمرين، پيادهسازي برنامهاي براي دستهبندي دادهها بر مبناي الگوريتم رگرسيون لاجستيک است.

دادهها

داده پیوست شده مربوط به تائید اعتبار (جعلی یا اصلی بودن) است .این داده ها داراي ٤ ويژگی و برچسب کلاس است. ۱۳۷۲ نمونه موجود است. است .تمام ۱ ويژگی عددي هستند. مقدار ۱ براي 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 درنظر 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 درنظر 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 می‌توانید از زبان پایتون و کتابخانه‌های pandas، NumPy، SciPy، math و یا کتابخانه های مشابه استفاده کنید.

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

- استفاده از گیتهاب و کتابخانه‌های مرتبط با الگوریتم مجاز نیست ، در صورت استفاده نمره‌های تعلق نخواهد گرفت.

- به کدهای مشابه نمره‌ای تعلق نمی‌گیرد.

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

- در صورت تاخیر تا ٢٤ ساعت ٢٠% تا ۴۸ ساعت ۴۵% و بیش از ۴۸ ساعت تمامی نمره تمرین را از دست خواهید داد.

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