به نام خدا

تمرین دوم مبانی یادگیری ماشین(بخش 1)

محمدرضا ضیالاری (97222057)

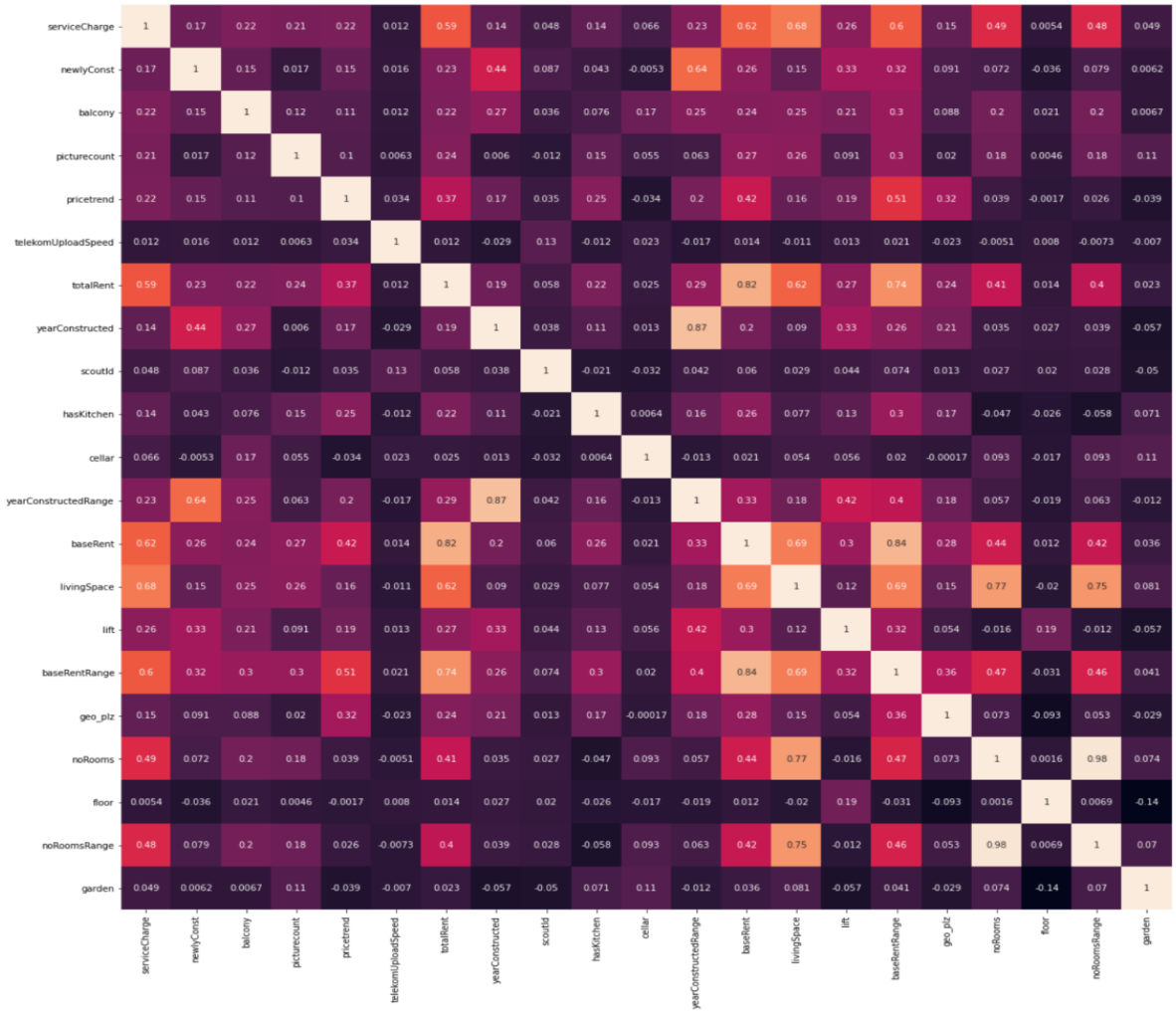
-مقدمه

در این بخش از تمرین از ما خواسته شده که بر اساس داده های تمرین 1 و معیار سنجش MSE و Accuracy ، 5fold-cross validation و 10fold-cross validation را بر روی مدل های مختلف اجرا کنیم .

ابتدا پیش پردازش بر روی داده ها انجام می دهیم .

1- ویژگی هایی که بیشتر از 30 درصد آنها تهی بود را از داده ها حذف می کنیم .
2- قسمت های تهی ویژگی های عددی را با میانگین آنها پر می کنیم و سپس آنها را نرمال سازی می کنیم و بین 0-1 اسکیل می نماییم.
3- ویژگی های به درد نخور مانند تاریخ و شماره خانه و ویژگی های کتگوریکال پراکنده مانند description را حذف می کنیم
4- ویژگی های کتگوریکال باقی مانده ای که نیاز به کوچک سازی دارند را به چند بخش اصلی و یک بخش other تقسیم میکنیم تا هنگام one hot کردن به مشکل ازدیاد ویژگی برنخوریم .
5- بخش های تهی ویژگی های کتگوریکال و بولین را با مد آن ویژگی پر می کنیم .
6- ویژگی های کتگوریکال و بولین را one hot می کنیم .
7- در نهایت با اطلاعات 253250 خانه با 53 ویژگی مواجه هستیم .
8- 90 درصد داده ها را برای ترین و 10 درصد را برای تست در نظر میگیریم .
9- X های ما همه ی ویژگی ها بجز مساحت و y های ما مساحت خانه ها می باشد .

ماتریس کرولیشن را نیز به دست می آوریم:



همانگونه که مشاهده می کنیم ویژگی های noRooms و baseRent بیشترین کورولیشن خطی را با مساحت خانه دارند و telekomUploadSpeed و scoutId کمترین کرولیشن را دارا هستند .

حالت 1:

در این حالت ما از مدل رگرسیونی خطی ساخته شده توسط خودمان استفاده می کنیم و آن را مورد سنجش قرار می دهیم . مدل ما بصورت زیر با لرنینگ ریت 0.0001 و تعداد 100 ایپاک اجرا شده است . که ورودی با بیشترین کرولیشن(noRooms) ورودی این مدل و خرووجی آن مساحت می باشد .

$$\hat{y} = w_1 * x_1 + w_2 * x_2 + \cdots + w_{61} * x_{61} + b$$

$$error^i = \frac{1}{2}({y_{train}}^i - \hat{y}^i)$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1} N(error^i)^2$$

$$\text{MSE} = \frac{1}{N}\left((error^0)^2 + (error^1)^2 + \cdots + (error^N)^2\right)$$

$$\text{MSE} = \frac{1}{N}\left((y^0_{train} - (w_1 * x_1^0 + \cdots + w_{61} * x_{61}^0 + b))^2 + \cdots \right.$$
$$\left. + ((y^N_{train} - (w_1 * x_1^N + \cdots + w_{61} * x_{61}^N + b))^2\right)$$

$$\frac{\partial \text{MSE}}{\partial w_1} = \frac{-2}{2N}\left({error^0} * {x_1}^0 + {error^1} * {x_1}^1 + \cdots + {error^N} * {x_1}^N\right) = \frac{-2}{2N}\left(\sum_{i=1}^{N} error^i * x_1^i\right)$$

$$\frac{\partial \text{MSE}}{\partial w_2} = \frac{-2}{2N}\left({error^0} * {x_2}^0 + {error^1} * {x_2}^1 + \cdots + {error^N} * {x_2}^N\right) = \frac{-2}{2N}\left(\sum_{i=1}^{N} error^i * x_2^i\right)$$

$$\vdots$$

$$\frac{\partial \text{MSE}}{\partial w_{61}} = \frac{-2}{2N}\left({error^0} * {x_{61}}^0 + {error^1} * {x_{61}}^1 + \cdots + {error^N} * {x_{61}}^N\right) = \frac{-2}{2N}\left(\sum_{i=1}^{N} error^i * x_{61}^i\right)$$

$$\frac{\partial \text{MSE}}{\partial b} = \frac{-2}{2N}\left(\sum_{i=1}^{N} error^i\right)$$

میانگین mse برای 5fold-cross validation برابر 1.37e+266 می باشد که مقدار بسیار زیادی است اما دلیل قابل قبولی برای رفع این مورد یافت نشد چرا که ایرادی در مدل پیدا نکردم و میانگین مقدار دقت برابر 60.9196 درصد می باشد .

برای حالت 10fold-cross validation نیز میانگین mse بسیار بالا بود و برابر 4.9e+265 می باشد و میانگین دقت نیز برابر مشابه حالت 5fold-cross validation برابر 60.9196 درصد می باشد .

حالت 2:

در این حالت از رگرسون خطی موجود در پکیج استفاده کردیم و 5fold-cross validation و 10fold-cross validation را روی آن اجرا کردیم . در اینجا نیز ورودی ما ویژگی با بیشترین کرولیشن (noRooms) بود . که نتایج برای 5fold-cross validation بصورت میانگین mse

برابر 0.006 و دقت برابر 66.46 درصد بود . همچنین برای 10fold-cross validation بصورت میانگین mse برابر 0.006 و میانگین دقت برابر 66.45 درصد بود.

حالت 3:

در این حالت مدل ما یک رگرسیون خطی با استفاده از پکیج ها و ورودی 2 ویژگی با بیشترین کرولیشن و 2 ویژگی با کمترین کرولیشن می باشد . نتایج برای 5fold-cross validation بصورت میانگین mse برابر 0.006 و دقت برابر 66.55 درصد بود . همچنین برای 10fold-cross validation بصورت میانگین mse برابر 0.006 و میانگین دقت برابر 66.58 درصد بود.

حالت 4 :

در این حالت از تمام ویژگی هایی که بعد از پیش پردازش آنها را داریم به عنوان ورودی استفاده شد و مدل ما یک رگرسیون خطی با استفاده از پکیج ها می باشد . نتایج برای 5fold-cross validation بصورت میانگین mse برابر 0.00295 و دقت برابر 69.9 درصد بود . همچنین برای 10fold-cross validation بصورت میانگین mse برابر 0.00294 و میانگین دقت برابر 69.91 درصد بود.

حالت 5:

در این حالت نیز از تمام ویژگی هایی که بعد از پیش پردازش آنها را داریم به عنوان ورودی استفاده شد و مدل ما یک رگرسیون Ridge با استفاده از پکیج ها می باشد . نتایج برای 5fold-cross validation بصورت میانگین mse برابر 0.00295 و دقت برابر 69.9 درصد بود . همچنین برای 10fold-cross validation بصورت میانگین mse برابر 0.00294 و میانگین دقت برابر 69.9 درصد بود.

حالت 6:

در این حالت نیز مشابه دو حالت قبل از تمام ویژگی هایی که بعد از پیش پردازش آنها را داریم به عنوان ورودی استفاده شد و مدل ما یک رگرسیون Lasso با استفاده از پکیج ها می باشد . نتایج برای 5fold-cross validation بصورت میانگین mse برابر 0.015 و دقت برابر 60.92 درصد بود . همچنین برای 10fold-cross validation نیز بصورت میانگین mse برابر 0.015 و میانگین دقت برابر 69.92 درصد بود.

برای ویژگی های یکسان و فولد های یکسان چندان تفاوتی میان رگرسون خطی و رگرسیون Ridge وجود نداشت اما بعد از استفاده از رگرسیون Lasso دقت کاهش پیدا کرد و میانگین mse نیز بالا رفت .