

گزارش بخش اول تمرین دوم

محمد ویس مصطفی پور 97222085

پیش پردازش داده ها: همانند تمرین یک ابتدا قدم های زیر را برای پیش پردازش داده ها انجام میدهیم.

1- مدیریت null: فیچرهایی که بیشتر از 50 درصد آن ها null باشد را حذف میکنیم. سپس داده

های تهی باقی مانده را اگر داده عددی باشن با میانگین اعداد آن فیچر پر می کنیم در نهایت همه داده های عددی را نرمالایز و سپس بین صفر و یک اسکیل سازی میکنیم

2- حذف ویژگی های بدردنخور: تعدادی از ویژگی ها مانند تاریخ، شماره خانه یا ویژگی دیگری

مثل "توضیحات" نه تنها راهگشا نیستند بلکه بدتر مدل را سردرگم میکنن، پس آن ها را حذف میکنیم

3- کوچک سازی ویژگی های غیر عددی: ویژگی های کتگوریکالی را که نیاز به کوچک سازی

دارند به چند بخش اصلی و یک بخش Other تقسیم میکنیم

4- مدیریت null بخش کتگوریکال: داده های تهی کتگوریکال را با مد داده ها پر میکنیم

5- One hot encoding داده ها: داده های کتگوریکال را وان هات انکودینگ میکنیم

6- جدا کردن داده های test & train: داده ها را به دو بخش ترین (نود درصد) و تست (10

درصد) تقسیم می کنیم.

بخش اول:

طبق چیزی که از ما خواسته شد باید در این بخش حالت های مختلف 5 فولد و 10 فولد را پیاده سازی کنیم و برای هر کدام از آن ها MSE و ACCURACY را بررسی کنیم.

سوال اول: در این سوال بدون استفاده از پکیج رگرسیون را بدست می آوریم. در حالت $k=5$ مقدار

MSE برابر با 0.015 به دست می آید که معقول است و مقدار ACCURACY هم برابر 0.24

در حالی که $K=10$ باشد، مقدار MSE برابر 0.0156 بدست می آید و مقدار ACCRACY برابر 0.25 می بینیم که تفاوت چندانی احساس نمی شود

سوال دوم: در حالت $k=5$ مقدار MSE برابر با 0.014 به دست می آید و مقدار ACCURACY هم برابر 0.25

در حالی که $K=10$ باشد، مقدار MSE برابر 0.015 بدست می آید و مقدار ACCRACY برابر 0.24

سوال سوم: در حالت $k=5$ مقدار MSE برابر با 0.013 به دست می آید و مقدار ACCURACY هم برابر 0.28

در حالی که $K=10$ باشد، مقدار MSE برابر 0.013 بدست می آید و مقدار ACCRACY برابر 0.27

سوال چهارم: در حالت $k=5$ مقدار MSE برابر با 0.0117 بدست می آید و مقدار ACCURACY هم برابر 0.46

در حالی که $K=10$ باشد، مقدار MSE برابر 0.0115 بدست می آید و مقدار ACCRACY برابر 0.50

سوال پنجم: در حالت $k=5$ مقدار MSE برابر با 0.01 بدست می آید و مقدار ACCURACY هم برابر 0.49

در حالی که $K=10$ باشد، مقدار MSE برابر 0.0109 بدست می آید و مقدار ACCRACY برابر 0.52

سوال ششم: در حالت $k=5$ مقدار MSE برابر با 0.016 بدست می آید و مقدار ACCURACY هم برابر 0.21

در حالی که $K=10$ باشد، مقدار MSE برابر 0.016 بدست می آید و مقدار ACCRACY برابر 0.23

بخش دوم: رگرسیون های Ridge تقریباً همانند بودن و درصد دقت برابری (حدوداً 46 تا 50 درصد) را نشان می دادند. ولی در هنگام استفاده از رگرسیون Lasso به یکباره دقت افت کرد و مقدار MSE بالا رفت.