
**دانشگاه تهران**

**پردیس دانشکده های فنی**

**دانشکده مهندسی برق و کامپیوتر**


# یادگیری ماشین

## تمرین اول


**امیرحسین عباسکوهی**

**۸۱۰۱۹۷۵۳۹**


**استاد ابولقاسمی**

سوال ۶)

کد مربوط به این سوال در فایل Q6.py قرار گرفته است.

مراحل انجام شده:

در این سوال با مسئله سختی رو به رو نیستیم زیرا مدلی که قرار است بر اساس آن داده را پیش بینی کنیم صرفا یک مقایسه کننده می باشد. به همین منظور با استفاده از دستور os.walk بر روی تصاویر موجود در فایل دیتا ها حرکت میکنیم.

در فایل دیتا یکسری از تصاویر در اسم خود حرف c یا m به نشانه Chelsea و Manchester United نداشتند بنابراین به صورت دستی برای این تصاویر m یا c قرار داده شد که دیتای جدید همراه با دیگر فایل ها آپلود شده است.

حل به بررسی نتایج می پردازیم. به ازای تمام تصاویر با استفاده از کتابخانه PIL تمامی تصاویر را باز میکنیم و به ازای هر پیکسل مدار هر سه کانال red، green، و blue را جمع زده و میانگین میگیریم. سپس باید فاصله میانگین رنگ ها را با دو رنگ آبی و قرمز که نقاط (0,0,255) و (255,0,0) می باشند مقایسه کنیم. فاصله دو رنگ در واقع همان واصله بین دو نقطه در فضای سه بعدی است اما در اینجا برای سادگی کار تنها بررسی میکنیم برای یک رنگ میانگین رنگ آبی آن بیشتر است یا میانگین رنگ قرمز. بر اساس آن نتایج را به دست می آوریم.

برای ماتریس Confusion منچستر را کلاس Positive در نظر میگیریم. نتیج به دست آمده را در زیر میبینیم:

| | |
|---|---|
| ۱۸ | ۵۷ |
| ۴۶ | ۱ |

```
--  --
57  18
 1  46
--  --
Accuracy = 0.8442622950819673
Precision = 0.76
Recall = 0.9827586206896551
```

همانطور که مشاهده می شود مقدار accuracy مقدار ۸۴ درصد می باشد که با توجه به اینکه مدل تصمیم گیری بسیار ساده ای داشتیم نتیجه چندان بدی نمیباشد و نشان دهنده این است که ۸۴ درصد از تصاویر را درست دسته بنده کرده ایم.

مقدار Recall هم مقدار ۹۸ است که مقدار بسیار بالایی است و نشان دهنده این است که ۹۸ درصد تصاویر منچستر را درست پیش بینی کرده ایم.

مقدار آخر هم Precision است که مقدار ۷۶ درصد است که مقدار بدی با توجه به مدل نیست و نشان دهنده این است که ۷۶ درصد از تصاویری که ما منچستر پیش بینی کرده بودیم درست بوده است.

حال به تصاویری که اشتباه پیش بینی کردیم میپردازیم.

تنها تصویری که مربوط به منچستر بوده است و ما چلسی پیش بینی کرده ایم:

این موضوع به نوع مدل برمیگرددو همانطور که میبینیم رنگ لباس داور آبی است و تنها مورد قرمز پیراهن بازیکن قرمز است. در نتیجه میانگین رو به مقدار آبی متمایل شده است. همچنین حضور دو بازیکن مشکی پوش در اینجا هم موثر بوده است.

درباره تصاویری مربوط به تیم چلسی بوده اند و ما منچستر تشخیص داده ایم تعداد ۱۸ تصویر وجود دارد که همه آن ها را نمیتوان در اینجا نشان داد اما نمونه ای از آن ها و لیست تصاویر را در زیر میبینم:

```
FP: c13.jpg
FP: c14.jpg
FP: c19.jpg
FP: c2.jpg
FP: c24.jpg
FP: c26.jpg
FP: c30.jpg
FP: c33.jpg
FP: c35.jpg
FP: c41.jpg
FP: c42.jpg
FP: c44.jpg
FP: c47.jpg
FP: c49.jpg
FP: c54.jpg
FP: c60.jpg
FP: c65.jpg
FP: c7.jpg
FN: m9.jpg
```

در اکثر ای موارد موردی که باعث اشتباه شده است میتواند این موضوع باشد که در اکثر این تصاویر بکگراند تماشاگران قرمز است، پیراهن سفید و جوراب سفید که در زنگ سفید کانال قرمز مقدار ۲۵۵ دارد، و همینطور چمن است. دقت کنید که چمن سبز کامل نیست و مقدار کانال قرمز برای آن بیشتر است به همین دلیل رو به پیش بینی اشتباه رفته است.

البته اگر چلسی را کلاس Positive در نظر میگرفتیم این مقادیر طبع تفاوت داشت اما در کل

با توجه به confusion matrix مدل عملکرد مطلوبی داشته است(با توجه به ساده بودن).

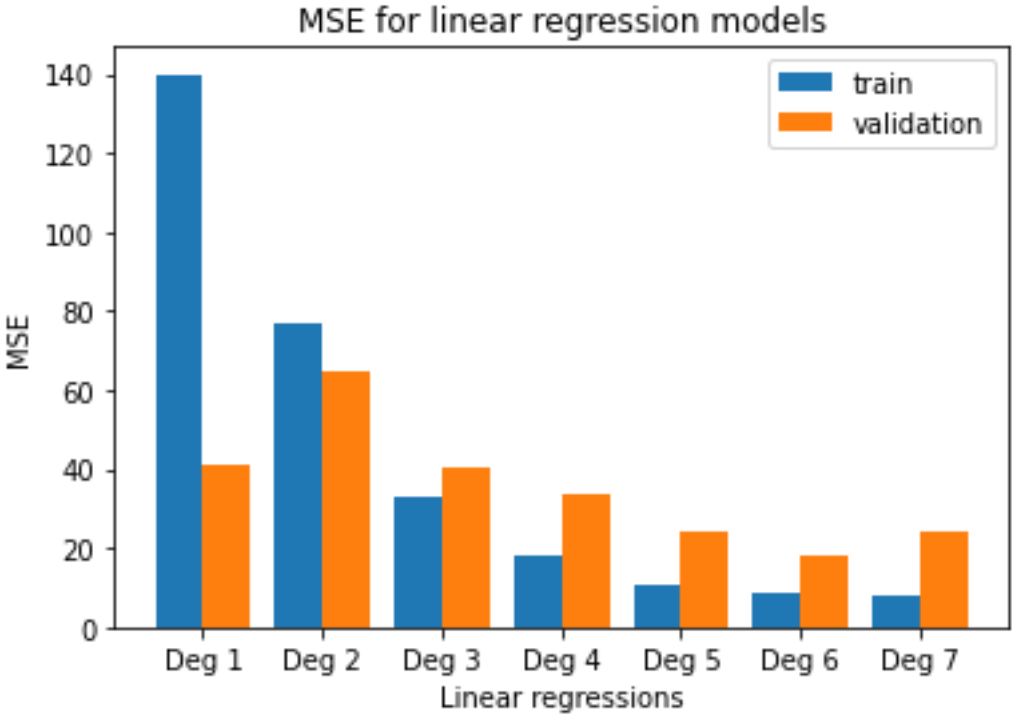
سوال ۷)

در اینجا به جزییات کد نمی پردازیم اما کلیت مدل به اینصورت است که برای هرکدام از درجات یک LinearRegression را در نظر میگیریم. از طرفی برای درجات بالاتر باید پارامتر های با درجات مختلف را اضافه کنیم که در نتیجه آن باید از PolynomialFeatures استفاده کنیم.

در داده ی فیجر Influencer به صورت عددی تبدیل شده است و از روش عدد گذاری استفاده شده است و نه روش one hot.

تابع MSE به صورت دستی پیاده سازی شده است. همینطور برای محاسبه bias, variance و MSE از mlxtend.evaluate.bias_variance_decomp استفاده میکنیم.

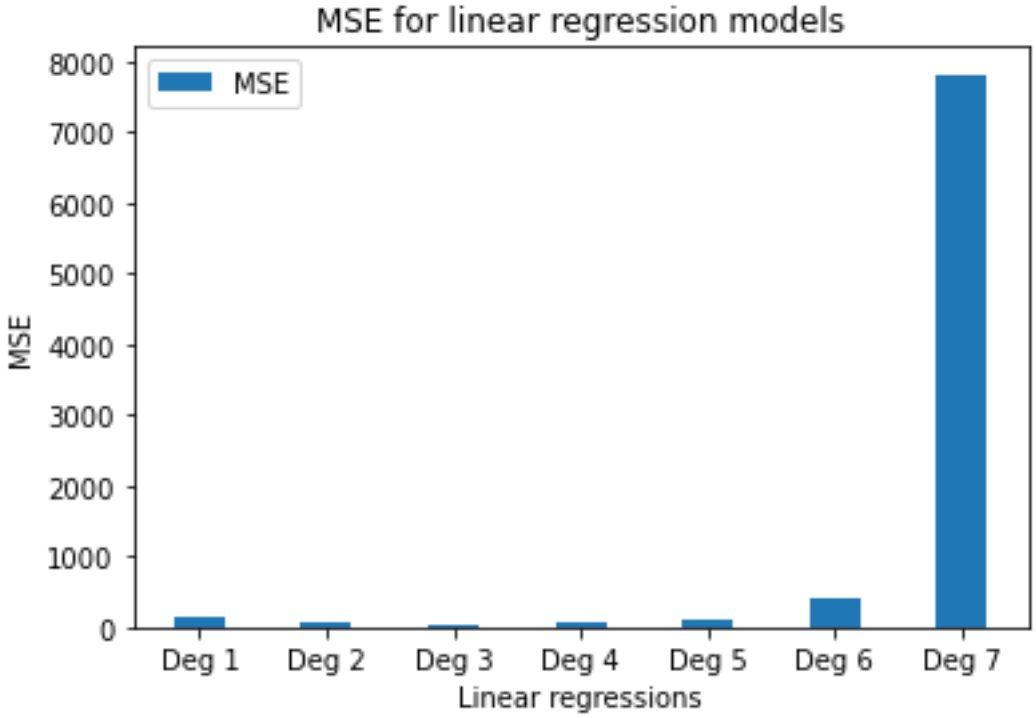
اولین تصویر مربوط به نتایج MSE برای دیتاست های train وvalidation می باشد. این نمودار با توجه shuffle کردن داده ها در زمان train-validation-test split متفاوت خواهد بود. نمونه ای از آن در اینجا آمده است:

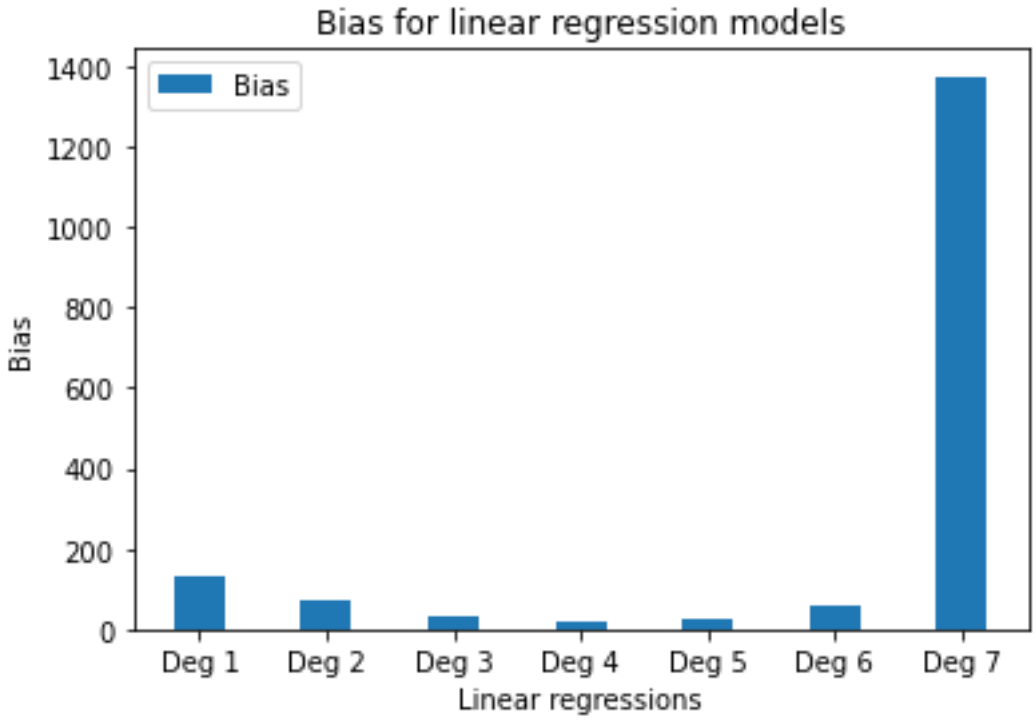


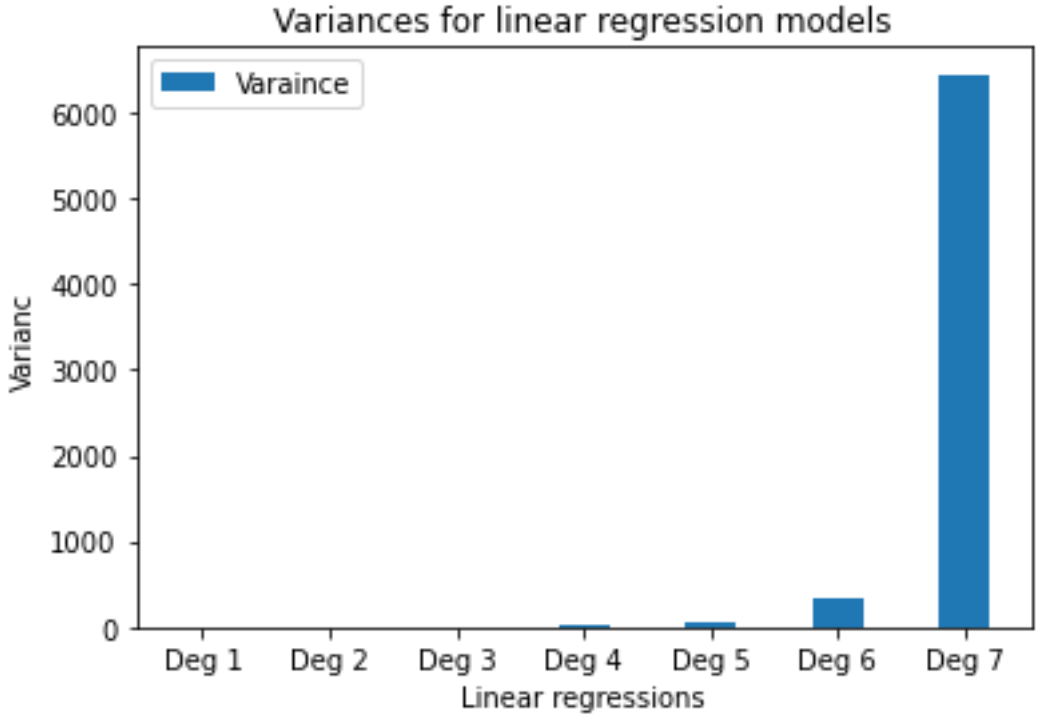

همانطور که میبینیم نمودار تقریبا نشان دهنده همان بحث bias variance trade off می باشد. در ابتدا که درجه مدل پایین است نمودار دقت خوبی ندارد اما رفته رفته روی داده train این موضوع با مدل قوی تر و پیچیده تر بهتر میشود اما از طرفی با بالارفتن درجه مدل به noise ها حساس می شود و در نتیجه به مشکل واریانس میخوریم به همین دلیل درجه ۵ یا ۶ نتیجه بهتر در کل دارد.

نتایج بعدی به ترتیب برای MSE، bias و variance برای درجه های مختلف می باشد که روی داده تست و به دست آمده است.

V


MSE for linear regression models
MSE
8000
7000
6000
5000
4000
3000
2000
1000
0
MSE
Deg 1
Deg 2
Deg 3
Deg 4
Deg 5
Deg 6
Deg 7
Linear regressions


Bias for linear regression models
Bias
1400
1200
1000
800
600
400
200
0
Bias
Deg 1
Deg 2
Deg 3
Deg 4
Deg 5
Deg 6
Deg 7
Linear regressions




همانطور که میبینیم واریانس با پیچیده شدن مدل بالاتر میرود اما برای bias این مقدار در ابتدا زیاد است اما به مرور کم می شود و در نهایت دوباره افزایش دارد. با توجه به این موارد MSE در درجه سوم نتیجه بهتری دارد و میتوان این مدل را به عنوان مدل مناسب برگزید.

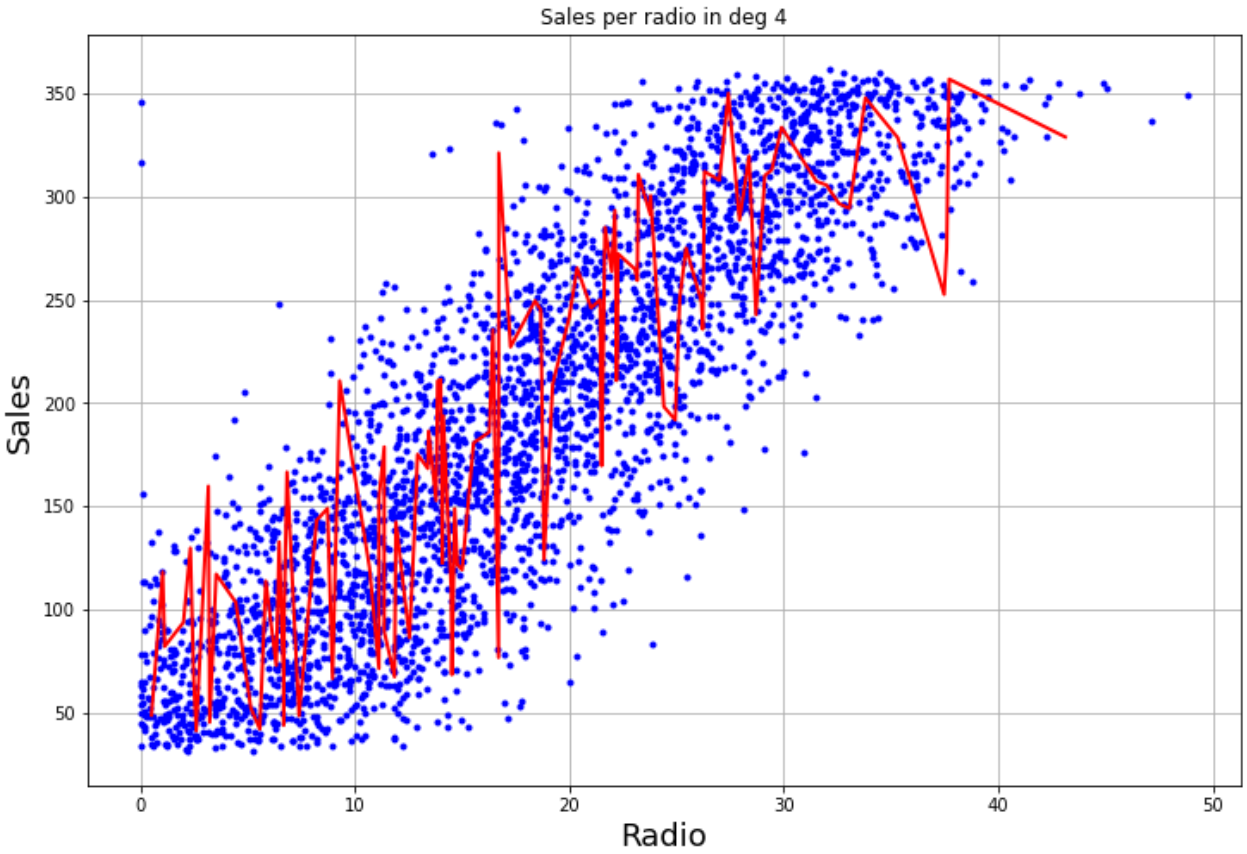
این موضوع همان bias variance trade off می باشد.

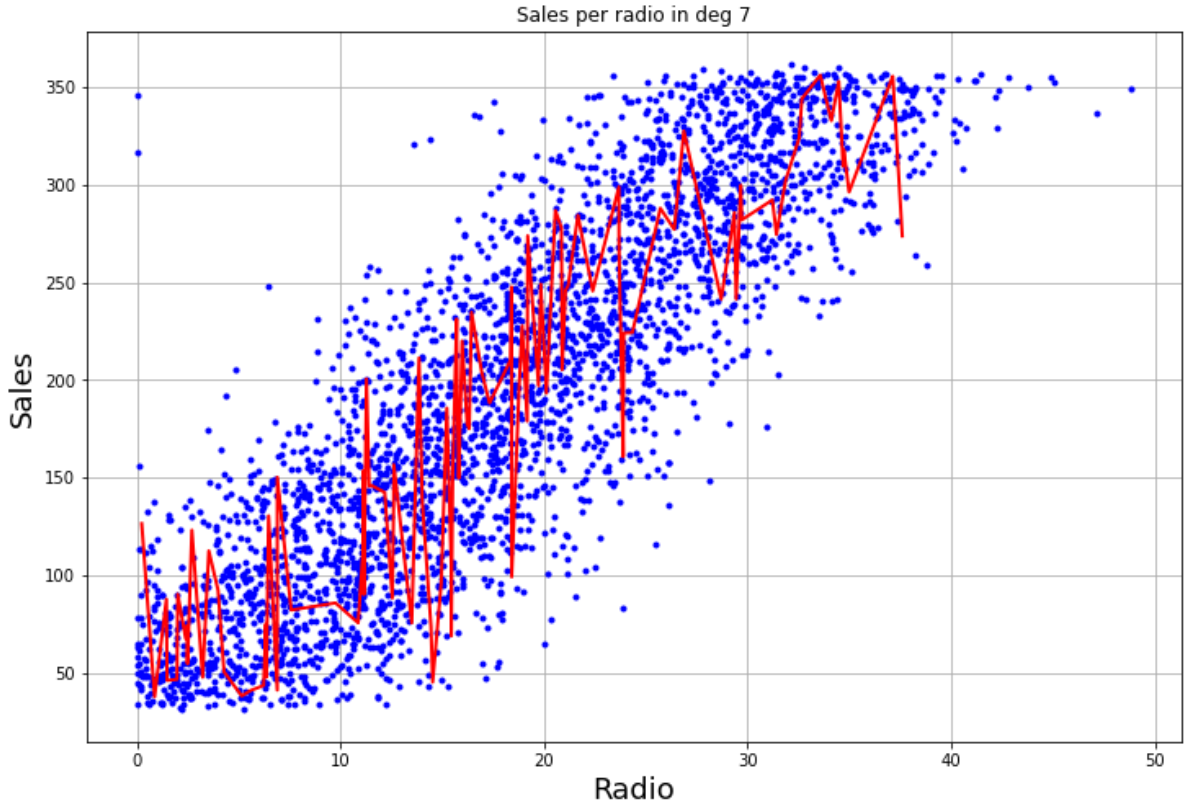
نتایج عددی نمودار ها:

| Deg | Bias | Variances | MSE |
|---|---|---|---|
| 1 | 133.79614597185937 | 1.5639686141047096 | 135.36011458596408 |
| 2 | 74.63198303264957 | 5.256333143850978 | 79.88831617650048 |
| 3 | 30.90078243747101 | 6.858908384244397 | 37.75969082171536 |
| 4 | 19.811704467669216 | 33.093793494184055 | 52.90549796185325 |
| 5 | 24.025643756821925 | 69.84981112861784 | 93.87545488543984 |
| 6 | 56.59559645147436 | 340.98438044809365 | 397.579976899568 |
| 7 | 1375.6472769777051 | 6441.479129808111 | 7817.126406785817 |

در نهایت نمودار پیش بینی را برای بر اساس فیچر Radio رسم میکنیم. نمودار های این بخش در notebook موجود می باشد اما با توجه به نمودار ها بازهم همان پیچیده تر شدن مدل و توجه بیشتر به noise ها مشاهده میشود.

برای رسم نمودار، ابتدا داده های تست را مرتب میکنیم، سپس ۱۰۰ داده به صورت رندم برمیگزینیم و بر اساس مدل پیش بینی را انجام میدهیم. در نهایت نمودار داده های Train را رسم میکنیم و مدل پیش بینی شده را رسم میکنیم. در اینجا فقط داده فیچر Radio را بر روی محور X قرار داده ایم. نمودار را برای درجه ۴ و ۷ در زیر میبینم:






البته انتظار میرفت که نمودار به صورت خطی و polynomial با درجات بالاتر باشد اما نمودار به این نتایج نرسیده است که ممکن اسن بحث اثر گذاری داده های دیگر باشد.(ما به روش درست مثل PCA کاهش ابعاد را انجام نداده ایم و صرفا از یکسری فیچر ها صرف نظر کرده ایم)