به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

گزارش پروژه اول (رگرسیون خطی)

درس یادگیری ماشین آماری

مهدی طاهراحمدی ۹۲۳۱۰۴۲

استاد: دکتر نیک آبادی

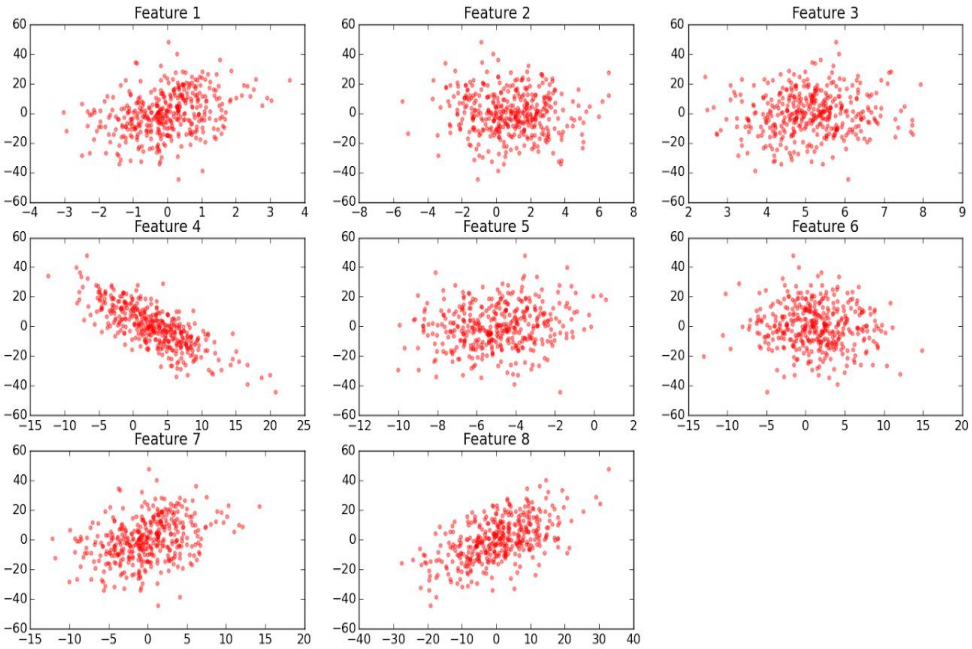
پاییز ۹۵

بخش اول)

مجموعه داده های مصنوعی:

الف:

نمودار Scatter مرتبط با هر ویژگی به شکل زیر است:



ویژگی ۱:

محدوده نسبتا کم است (تقریبا بین -3 تا 4). داده پرت زیاد . همبستگی مثبت خیلی کم.

ویژگی ۲:

محدوده دو برابر ویژگی اول. داده پرت بیشتر دارد. همبستگی خیلی کم در قیاس با ویژگی اول.همبستگی با داده هدف تقریبا منفی است.

ویژگی ۳:

همبستگی خیلی ضعیف.داده پرت کمتر نسبت به ویژگی قبل.

ویژگی ۴:

همبستگی منفی قوی.

ویژگی ۵:

مانند ویژگی های ۱ و ۲ و۳ همبستگی خیلی ضعیف دارد.

ویژگی ۶:

همبستگی تقریبا صفر است. داده پرت زیاد دارد.

ویژگی ۷:

همبستگی مثبت ضعیف.
ویژگی ۸:
همبستگی مثبت قوی. بعد از ویژگی 4 بیشترین همبستگی را دارد.

به طور خلاصه:

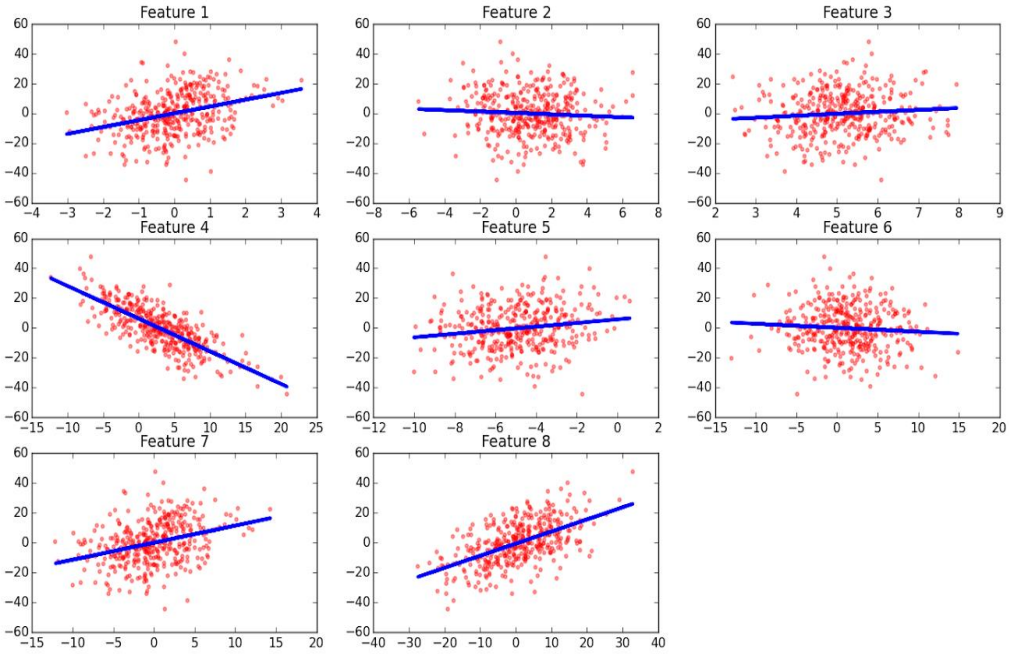
	Strength	Slope	Outliers	Scale
F1 vs L	Weak	+	High	(-4, 4)
F2 vs L	Very Weak	-	Moderate	(-8, 8)
F3 vs L	Very Weak	+	High	(2, 9)
F4 vs L	Very Strong	-	Very Low	(-15, 25)
F5 vs L	Weak	+	Moderate	(-10, 0)
F6 vs L	Very Weak	-	Moderate	(-15, 20)
F7 vs L	Moderate	+	Moderate	(-15, 20)
F8 vs L	Strong	+	Low	(-30, 30)

ب:
با استفاده از روابط موجود در کتاب:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}, \qquad (14.5)$$

$$\widehat{\beta}_0 = \overline{Y}_n - \widehat{\beta}_1 \overline{X}_n. \qquad (14.6)$$

پارامتر های linear regression از روش least squares estimate بدست آمدند.
نمودار های حاصل به شکل زیر است:



همچنین ارور استاندارد و تخمین واریانس از روابط زیر بدست آمدند:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_i \widehat{\epsilon}_i^2. \qquad (14.10)$$

$$\widehat{\mathsf{se}}\,(\widehat{\beta}_0) = \frac{\widehat{\sigma}}{s_X\sqrt{n}}\sqrt{\frac{\sum_{i=1}^n X_i^2}{n}} \qquad (14.12)$$

$$\widehat{\mathsf{se}}\,(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{s_X\sqrt{n}}. \qquad (14.13)$$

$$where\ s_X^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

معیار های R2 , RSS هم از روابط زیر محاسبه شده اند:

In the form of an equation:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y})^2$$

Total sum of squares

Regression (explained) sum of squares

Residual (unexplained) sum of squares

Regression (explained) sum of squares

$$r^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

Total sum of squares

گزارش مقادیر بدست آمده:

Test data	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\widehat{se}(\widehat{\beta_0})$	$\widehat{se}(\widehat{\beta_1})$	$\widehat{\sigma}^2$	RSS	R^2	$RSS\ Test$	$R^2\ Test$
F1	-0.07	4.58	261.1	252.6	181.96	72420.7	0.11	14463.9	0.18
F2	0.17	-0.48	157.2	73.5	203.57	81023.3	0.004	17831.5	-0.006
F3	-6.93	1.31	1524.25	292.33	202.79	80711.4	0.008	16988.5	0.04
F4	6.21	2.19	43.07	7.45	85.13	33883.2	0.583	8751.9	0.5
F5	5.71	1.2	386.74	72.88	198.82	79134.0	0.027	16991.6	0.04
F6	0.11	-0.26	72.23	16.74	203.23	80888.5	0.006	17770.2	-0.002
F7	-0.07	1.14	65.29	15.78	181.96	72420.8	0.110	14463.95	0.18
F8	-0.72	0.81	23.97	2.4	139.8	55667.3	0.315	14081.01	0.2

ج:

بنا بر معیارهای R2 و RSS در بخش قبل مدل 4 بهترین مدل انتخاب شد.

در هر مرحله یک ویژگی به آن اضافه کرده و معیارprediction risk AIC را که مطابق رابطه زیر است:

$$\ell_S - |S| \qquad (14.26)$$

را محاسبه میکنیم تا جایی که مدل بهبود پیدا کند.

خروجی زیر بیانگر مراحل انجام این روش است:

Current AIC only using Feature4 : -1370.65051699
AIC after adding Feature1 :-1116.83583082
AIC after adding Feature2 :-1147.55091986
AIC after adding Feature3 :-1084.17621322
AIC after adding Feature5 :-1120.05396117
AIC after adding Feature6 :-1147.01321782
AIC after adding Feature7 :-1116.83579677
AIC after adding Feature8 :-1004.97688032
Model Uses Feature (4) (8)
RSS: 22166.4775617 and R2: 0.727625844192

AIC after adding Feature1 :-872.740330977
AIC after adding Feature2 :-999.718528724
AIC after adding Feature3 :-889.561997673
AIC after adding Feature5 :-961.908005422
AIC after adding Feature6 :-1003.72193249
AIC after adding Feature7 :-872.740420133
Model Uses Feature (4) (8) (1)
RSS: 11386.1577047 and R2: 0.860090757132

AIC after adding Feature2 :-851.359379569
AIC after adding Feature3 :-521.045021233
AIC after adding Feature5 :-781.628381942
AIC after adding Feature6 :-866.334549718
AIC after adding Feature7 :-873.712752701
Model Uses Feature (4) (8) (1) (3)
RSS: 1952.13146085 and R2: 0.976012870913

AIC after adding Feature2 :-521.310488203
AIC after adding Feature5 :441.903468335
AIC after adding Feature6 :-522.034365154
AIC after adding Feature7 :-521.966801426
Model Uses Feature (4) (8) (1) (3) (5)
RSS: 15.7514852688 and R2: 0.999806451093

AIC after adding Feature2 :441.77316912
AIC after adding Feature6 :441.317793578
AIC after adding Feature7 :442.467031705
Model Uses Feature (4) (8) (1) (3) (5) (7)
RSS: 15.6288231417 and R2: 0.999807958324

AIC after adding Feature2 :442.535265977
AIC after adding Feature6 :441.777067551
Model Uses Feature (4) (8) (1) (3) (5) (7) (2)
RSS: 15.5455694531 and R2: 0.999808981317

AIC after adding Feature6 :441.860387784
Model Uses Feature (4) (8) (1) (3) (5) (7) (2) (6)
RSS: 15.5203189642 and R2: 0.999809291586

د:

با توجه به تنیجه بخش قبل مشاهده شد که بعد از اضافه کردن ویژگی های (4) (8) (1) (3) (5) به مدل، معیارAIC افزایش قابل توجهی نداشته است.

Model Uses Feature (4) (8) (1) (3) (5)
RSS: 15.7514852688 and R2: 0.999806451093

ه:

مدلی که از همه ی ویژگی ها استفاده می‌کند در آخر قسمت قبل مطابق زیر بدست آمده بود.

Model Uses Feature (4) (8) (1) (3) (5) (7) (2) (6)
RSS: 15.5203189642 and R2: 0.999809291586

معیار خطای LOOCV برای این مدل به شکل زیر به دست می‌آید:

$$\widehat{R}_{CV}(S) = \sum_{i=1}^{n} \left(\frac{Y_i - \widehat{Y}_i(S)}{1 - U_{ii}(S)} \right)^2 \qquad (14.28)$$

با استفاده از روش کنار گذاشتن یک نمونه از داده ها و یادگیری مدل و اندازه گیری RSS ، مقدار RSS میانگین این مدل ها که همان Leave One Out Cross Validation میباشد، 15.52 محاسبه شده است.

و:

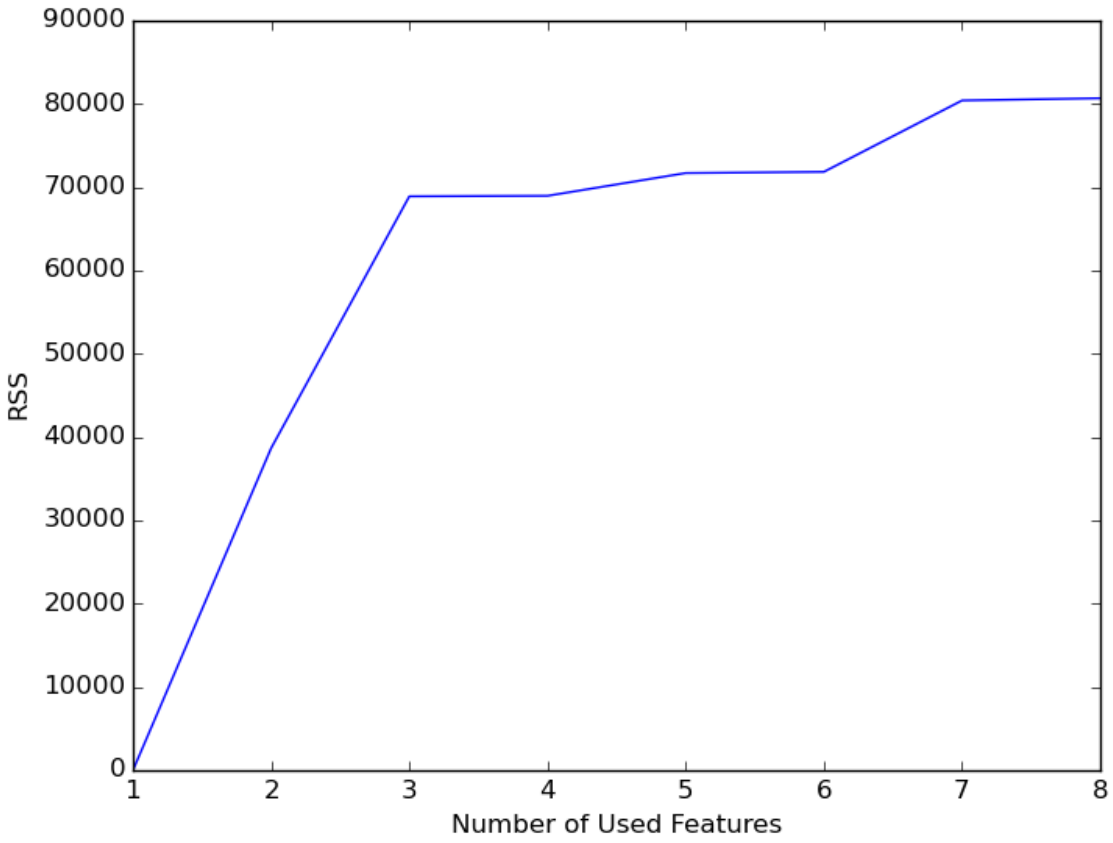
در فرایند backward و مدل بدست آمده در بخش قبل ابتدا همه ویژگی ها را استفاده میکنیم سپس ازاول یکی یکی آن ها را کم میکنیم، یعنی به ترتیب ویژگی های (4) (8) (1) (3) (5) (7) (2) (6) حذف می‌شوند. مشاهده میشود که در ابتدا RSS به شدت افزایش می‌یابد.

در واقع به جای جستجوی کل فضای حالت فقط حالت هایی که منتهی به جواب از حالت قبل بود را بررسی کردیم. همانطور که انتظار داشتیم، بهترین مدل با این جستجوی حریصانه، همان مدلیست که از همه ویژگی ها استفاده کند.

نتایج به دست آمده:

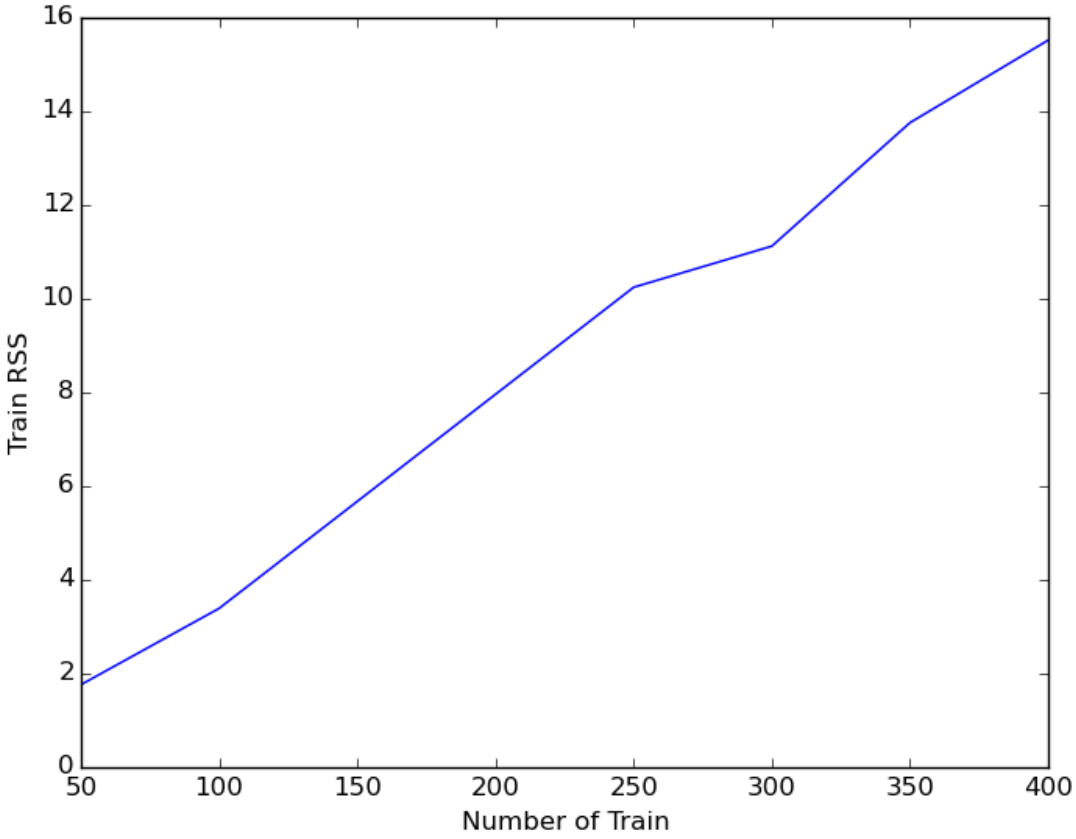
Features	LOOCV
8	15.48
7	38683.45
6	68912.41
5	68975.43
4	71707.02
3	71850.29
2	80421.39
1	80690.97

نمودار بر حسب تعداد ویژگی به شکل زیر است:



ز)

بهترین مدل که با تمام ویژگی ها آموزش داده می‌شود را انتخاب کرده و با تعداد داده های 50, 100,250, 300, 350, 400 آموزش می‌دهیم. RSS Train های به دست آمده بر حسب تعداد داده های آموزشی در نمودار زیر رسم شده است:



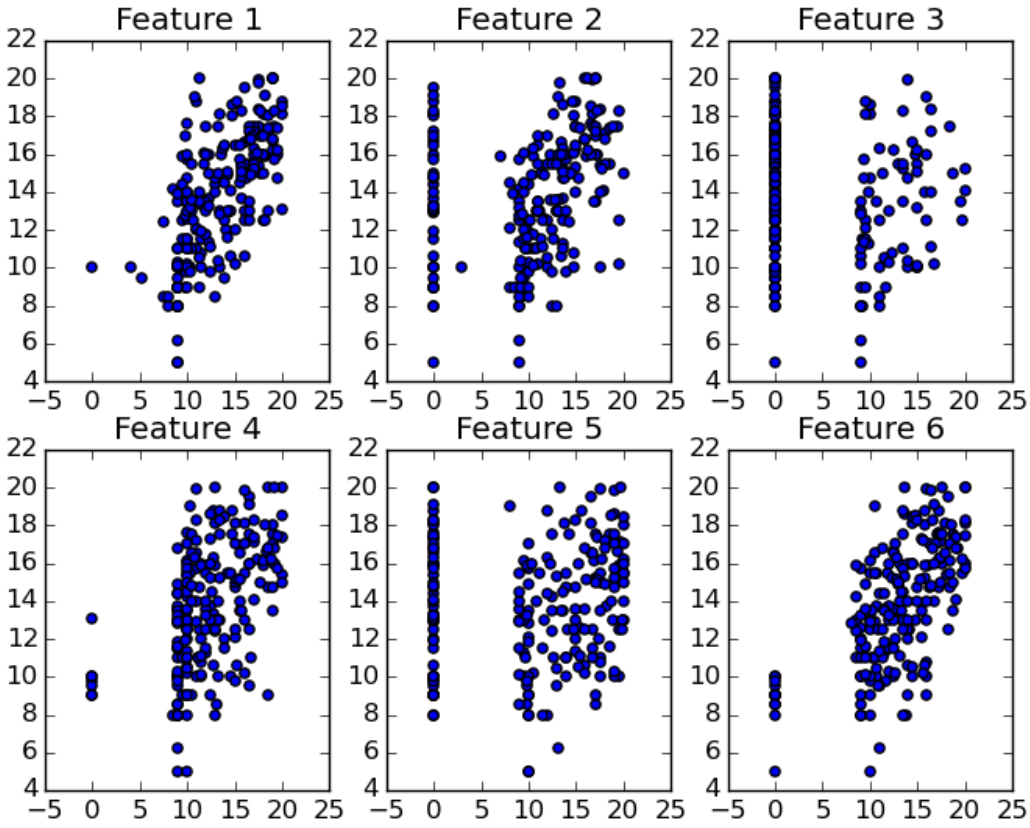
نتیجه این که با افزایش تعداد داده تست، معیار RSS هم تقریبا خطی افزایش میابد، یعنی میان تعداد داده آزمایش و انتخاب مدل بهتر رابطه مستقیم وجود دارد.

بخش دوم)

مجموعه داده های نمرات:

الف)

نمودار Scatter مرتبط با هر ویژگی کشیده شده است. حاصل شکل زیر است:

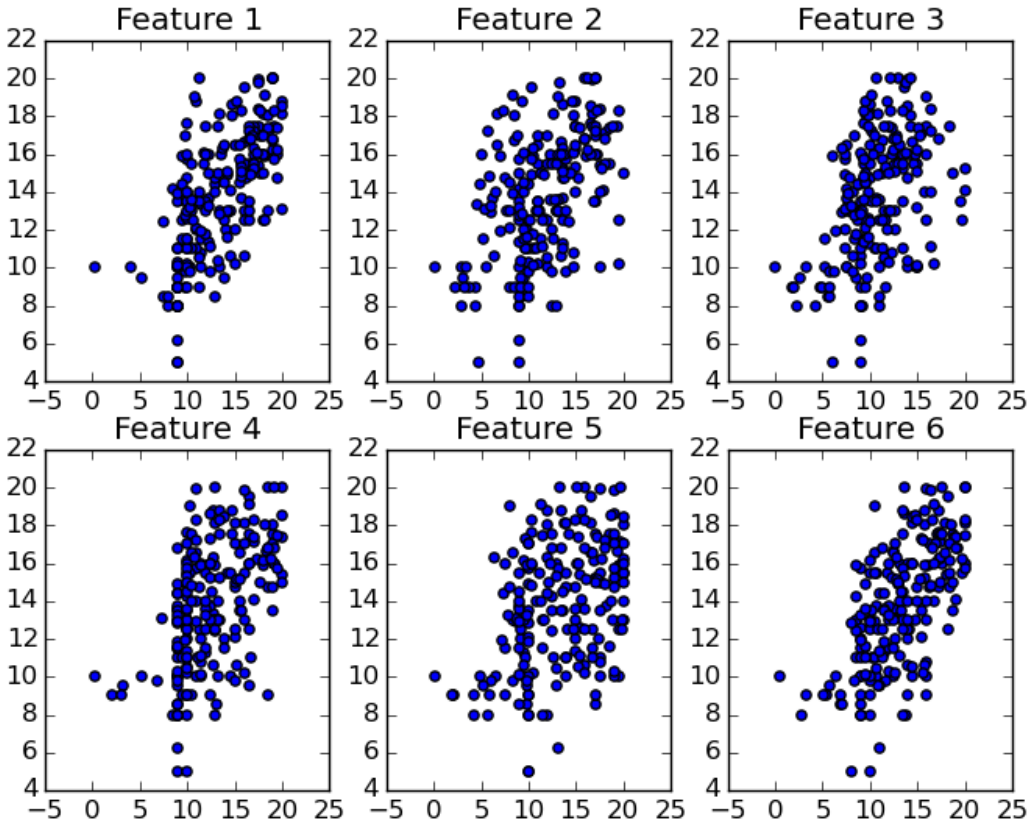


از آنجایی میدانیم جنس داده ها نمره دانشجویان است، انتظار میرود که تمام همبستگی ها مثبت باشند که همین موضوع مشاهده می‌شود. مقادیر این همبستگی و مقدار داده پرت در جدول زیر جمع بندی شده است:

	Strength	Slope	Outliers
F1	Strong	+	Moderate
F2	Moderate	+	High
F3	Very Weak	+	Very High
F4	Moderate	+	Moderate

F5	Very Weak	+	Very High
F6	Very Strong	+	Very Low

ب) همانطور که گفته شد برچسب ها مقادیر نامشخص ندارد. پس فقط ۶ نمره دیگر را بررسی میکنیم. ساده ترین کار برای حذف مقادیر نامشخص، میانگین تمام داده هاست که بدیهیست خطای بالایی دارد. در این مساله از روش نویز گاوسی استفاده شده است. در این روش برای هر دانشجو، یک توزیع نرمال با میانگین نمرات آن دانشجو در نظر میگیریم و نمره حدس زده شده برای نمره نامشخص را از این توزیع با یکبار نمونه برداری به دست می آوریم. در این توزیع، واریانس ۱ در نظر گرفته شده است. دلیل این کار این است که تعداد درس ها کم است و واریانس یک یا دو عدد قابل اطمینان نیست.
نتیجه این روش به شکل زیر است:

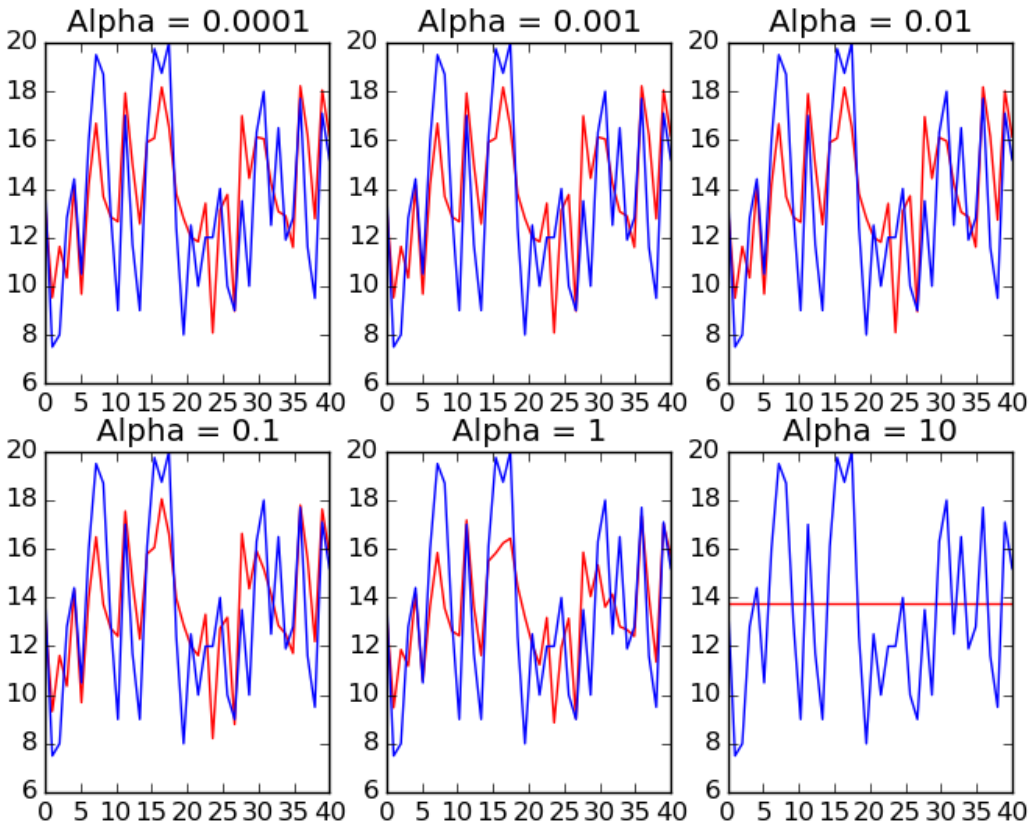


ج) برای پیاده سازی روش Lasso که رابطه آن مطابق زیر است:

$$\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \qquad (14.31)$$

از تابع موجود در کتابخانه Sci-kit استفاده شده است.

د) مدل حاصل از مقادیر λ های 0.0001, 0.001, 0.01, 0.1, 1, 10 به شکل زیر در آمده است:



همانطور که انتظار می رفت، با افزایش مقدار λ ، مدل ساده تر می‌شود. این موضوع به این دلیل اتفاق می افتد که وقتی جریمه ضرایب بالا افزایش پیدا کند، مدل به سمت ضرایب ساده تر و در نتیجه مدل ساده تر می‌رود.

از نمودار های به دست آمده مشخص است که مدل با آلفا کم یعنی 0.0001 بهترین مدل در بین این ۶ مدل است. برای این مدل معیار RSS و R^2 به صورت زیر است:

```
RSS: 284.84, R2: 0.43
```

ه) پیشبینی داده های بدون برچسب حاصل از این مدل در فایل predictions.csv در کنار گزارش قرار داده شده است.

منابع:

[1].Larry Wasserman, All of Statistics: A Concise Course in Statistical Inference, Springer, 2010.
[2].http://www.uow.edu.au/student/qualities/statlit/module3/5.4interpret/index.html