



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده مهندسی برق

Statistical learning Assignment 1

Mohammad hasan shammakhi

محمد حسن شماخی

۹۳۱۲۳۰۵۳

we may need to talk
how we can share our work with others
please come to the office to talk.

Chapter 2 Question 5:

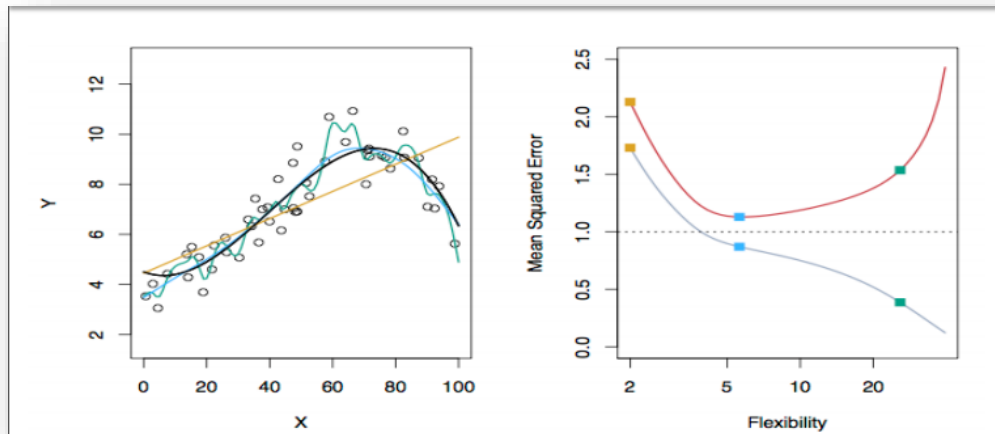
از مزایای انعطاف پذیری بالای تابع تخمین می توان گفت که موجب به کمتر شدن خطای مربع تفاضلات Training set که این کم شدن خطا همراه با کاهش خطای Test set نسبت به تابع تخمین تا قبل از آغاز مرحله Overfitting خواهد بود.

از معایب آن می توان به پیچیده تر شدن مدل و امکان Overfitting اشاره کرد. *variance زیاد*

اما کی کدوم روش ارجعیت دارد؟

I am not sure about it!

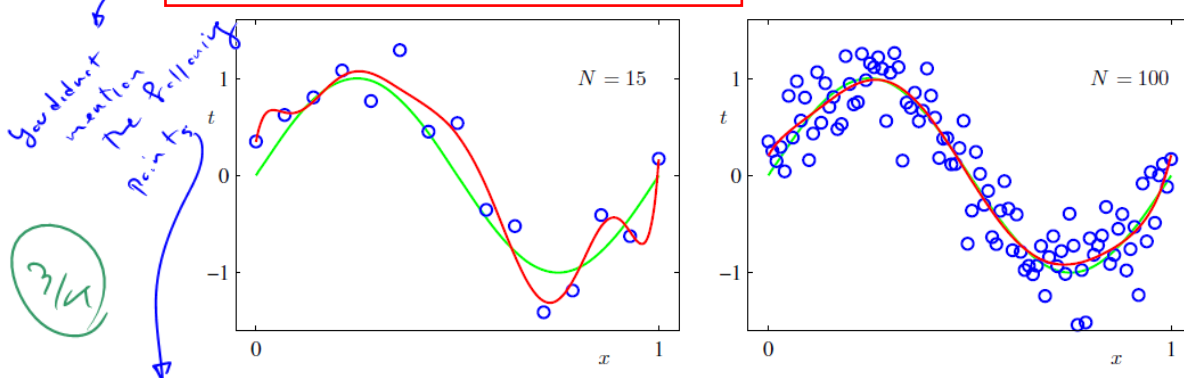
زمانی که ما درجه تابع را زیاد کردیم و RSS بهبود یافت و از طرفی این طور می توان گفت که رنج ضرایب زیادی بزرگ نشد (چرا که در این حالت این طور حس می شود که تخمین در حال تلاش زیاد برای ماندن در دور رقابت است و متناسب با داده ها نمی باشد) اما راه اصلی که قبلا در کلاس نیز بیان شده رسم نمودار RSS برای داده های آموزشی و تست هست.



what does mean here?

اما نکته حائز اهمیت این است که اگر تعداد داده ها کم باشد ممکن است Test set نسبت به تابع تخمینی که برای Training set در نظر گرفته شده، طوری باشد که Overfitting رخ ندهد (مانند شکل زیر) بنابراین باید تعداد مناسبی داده جمع آوری کرد.

hazuri khematetun manzuramo goftam
bargerefte az tahlile bishop hast un jomle



*if we want good prediction we might want lower flexible model.
if we want to model the exact training set more flexible is good.*

4/4

Chapter2 Question8:

کد این سوال به صورت **ch2e8.R** می باشد.

در داده هایی که Outstate صفر است با احتمال بالایی Private ، no است.

در حالی که ارتباط Elite با Outstate دقیقاً برعکس است.

Chapter3 Question3:

با توجه به فرضیات مسئله داریم.

$$Y = B_0 + B_1 * GPA + B_2 * IQ + B_3 * Gender + B_4 * GPA * IQ + B_5 * GPA * Gender$$

که برای مرد و زن بودن به صورت زیر می شود.

$$Y = \begin{cases} B_0 + B_1 * GPA + B_2 * IQ + B_3 + B_4 * GPA * IQ + B_5 * GPA & \text{for Woman} \\ B_0 + B_1 * GPA + B_2 * IQ + B_4 * GPA * IQ & \text{for Man} \end{cases}$$

که مقدار IQ و GPA ثابت هستند پس فرض میکنیم:

$$A = B_0 + B_1 * GPA + B_2 * IQ + B_4 * GPA * IQ$$

در نتیجه داریم:

$$Y = \begin{cases} A + B_3 + B_5 * GPA & \text{for Woman} \\ A & \text{for Man} \end{cases} = \begin{cases} A + 35 - 10 * GPA \\ A \end{cases}$$

بنابراین برای مقادیر زیاد GPA مرد بودن موجب Y بیشتری نسبت به زن بودن می شود. ✓

با توجه به مقادیر $GPA=4$ و $IQ=110$ خواهیم داشت:

$$A = B_0 + B_1 * GPA + B_2 * IQ + B_4 * GPA * IQ = 50 + 20 * 4 + 0.07 * 110 + 35 * 1 + 0.01 * 4 * 110 - 10 * 4 * 1 = 137.1 \quad \checkmark$$

(C) غلط. ✓

با توجه به اینکه ضریب b_4 در $IQ * GPA$ ضرب می شود لذا تاثیر آن زیاد است و کوچکی آن دلیل بر بی ارزشی آن نمی باشد

مثلاً اگر GPA برابر ۸ باشد تاثیر ضریب b_4 از ضریب b_2 بیشتر می شود.

3/3

Chapter3 Question4:

in ghemat ro hozuri khematetun goftam manzuramo, lotfan dar surate emkan eslah farmayid. albate ghabul daram ke nesbat be khata bar ruye training set eshtebah mikardam

در ابتدا باید گفت از اینکه $n=100$ هست نمیتوان گفت n به اندازه کافی بزرگ هست یا کوچک چرا که بستگی به پراکندگی و

میزان تغییرات پارامتر x دارد که مثلا x بین ۰ تا ۵۰ است یا بین ۰ تا ۱۰۰۰

اگر n را زیاد فرض کنیم:

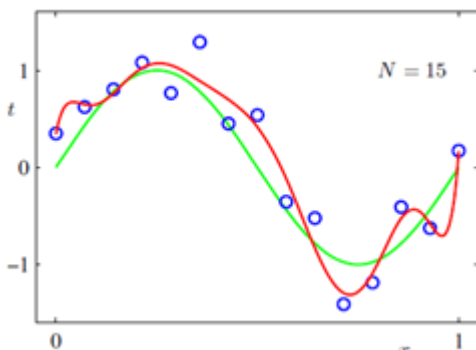
برای قسمت a با توجه به اینکه مدل خطی هست پس با تقریب تابع مدل با تابعی درجه ۳ نسبت به تابع خطی RSS زیاد می-

شود.

اما برای قسمت b و مدل غیرخطی نمی توان دقیقا اظهار نظر کرد چرا که مثلا برای تابعی که شکل درجه ۲ دارد شاید RSS

مدل خطی بهتر از مدل درجه ۳ باشد و یا برعکس که به تعداد داده ها و محل قرار گرفتن آنها بستگی دارد.

اما اگر n را کم در نظر بگیریم علاوه بر اینکه برای دو قسمت a و b نمی توان اظهار نظر کرد شاید جواب اشتباه هم بگیریم.



مثلا برای شکل:

با اینکه شکل مدل درجه ۳ هست ولی مدل درجه ۱۵ جواب بهتری

نسبت به مدل درجه ۳ می دهد چرا که مقدار RSS برای تعداد دیتا

کم به نوعی شانسی است.

Chapter3 Question9:

(c) بله در مدل سازی خطی با توجه به p value ها متوجه می شویم وزن و منشا تولید و سال تولید بسیار مهم است.

سال بیشترین تاثیر گذاری را داشته است.

کد مطالب گفته شده : [ch3e9.R](#)

explanation on part d-e-f is

missing.

dar code neveshte shode ghesmate e , f ro anjam dadam va chon unmoghe compile ro va export information az code ro balad nabudam natunestam mesle Assignment 2 untor gozarah benevisam, vagarna hamuni ke dakhele code neshun midad ro mizashtam ghesmat d ro ham dars nadade budid leverage va outlier ro, va ba'd az ersale taklif khundamesh.

Ostad kheyli sakhtgirane negah kardid. vagarna khodaish masalan malume ke baraye e vaghti * gozashtam ham baraye weight ham year va ham jofesh zarib midim vali vaghti / mizarim faghat baraye year*weight zarib midahim

you should explain your findings! as

in part e and f you need to study interaction effect of all vars. not just weight and origin!

Chapter3 Question7:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad TSS = \sum (y_i - \bar{y})^2$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\left\{ \begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^K (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^K (x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \bar{x} &= \bar{y} = 0 \end{aligned} \right\} \rightarrow \hat{\beta}_0 = 0$$

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

$$= \frac{-\sum_{i=1}^n (\beta_1 x_i)^2 + 2 \sum_{i=1}^n y_i \beta_1 x_i}{\sum_{i=1}^n y_i^2} = \frac{-\beta_1^2 \sum_{i=1}^n (x_i)^2 + 2 \beta_1 \sum_{i=1}^n y_i x_i}{\sum_{i=1}^n y_i^2} =$$

$$= \frac{-\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right)^2 \sum_{i=1}^n x_i^2 + 2 \left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right) \sum_{i=1}^n y_i x_i}{\sum_{i=1}^n y_i^2} = \frac{(\sum_{i=1}^n x_i y_i)^2}{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}$$

$$= \text{cor}(x, y)^2$$

chon javabi ke dar c bedast amade,
javabe b ham hast, bekhatere hamin
naneveshte budam

Chapter3 Question10:

c) model: Sales=13.04-0.05*Price+0.02*Urban+1.2*US

(d) پیش بینی کننده Urban چرا که مقدار t-statistic نزدیک صفر دارد.

(e) بنابراین با توجه به مقدار p-value و عدد t-statistic Urban را حذف کرده و دوباره مدل سازی می‌کنم.

(f) به دنبال کمترین خطا رفته که برای تحقق آن نسبت به متغیرها مشتق گرفته و برابر صفر قرار می‌دهم اما با توجه به ربط نداشتن ویژگی Urban و محاسبه آن در مدل اول، لذا مورد دوم بهتر است.

کد مطالب گفته شده: [ch3e10.R](#)

Chapter3 Question14:

$$a) Y=2.13+1.44*X1+1.01*X2$$

(c) با توجه به t-statistic مقادیر X1 و X2 نمیتوان هیچ یک را حذف کرد.

(e و d) در این حالت t-statistic بیشتر شده و متغیرها وابسته تر بنظر میرسن که با توجه به نمودار قبلی به cor بالای آنها پی می‌بریم.

کد مطالب گفته شده: [ch3e14.R](#)

f, g code bude va dar code ch3e14
matlab khaste shode ra anjam dadam