علیرضا درویشی ۹۶۱۰۹۶۷۴
تمرین سری اول
تحلیل رگرسیون بهار ۹۸-۹۹
۹۸/۱۲/۹

۱.
الف)
چون تعداد متغیرهای توضیح دهنده بسیار کم است ولی تعداد داده‌ها به اندازه‌ی کافی زیاد است، بهتر است برای پیش‌بینی پارامترهای بیشتری را تعیین کنیم و این یعنی درجه آزادی بیشتری برای تابع مورد نظر انتخاب کنیم.
ب)
تعداد متغیرهای توضیح دهنده زیاد است و تعداد داده‌ها بسیار کم است پس نمی‌توان از پارامترهای زیادی برای پیش‌بینی استفاده کرد و بهتر است درجه آزادی کمتری برای تابه انتخاب کنیم.
ج)
وقتی رابطه‌ی تابع با متغیرها بسیار غیرخطی است، باید برای مدل سازی این تابع غیرخطی از پارامترهای زیادی استفاده کنیم و این به معنی درجه آزادی بیشتر تابع است
د)
اگر واریانس نویز زیاد باشد، با انتخاب درجه آزادی زیاد، تغییرات نویز را دنبال خواهیم کرد که گزینه‌ی خوبی نیست. پس بهتر است برای جلوگیری از $overfit$ درجه‌ی آزادی کمی انتخاب کنیم.
۲.
الف)
سوال $reggresion$ است و چون هدف توضیح دادن دارد پس $inference$ است. در این سوال $n = 500, p = 3$ است.
ب)
سوال $classification$ است و هدف فهمیدن این است که محصول موفق خواهد بود یا خیر پس $prediction$ است. در این سوال $n = 20, p = 13$ است.
ج)
سوال $reggresion$ است و چون هدف پیش بینی است پس مسئله $prediction$ است.
$n = 51, p = 3$
۵.
مزیت یک مدل $flexible$ پارامتر های بیشتری برای بدست آوردن و توضیح بهتر متغیر پیشبینی شونده است اما از مشکلات یک مدل $flexible$ امکان $overfit$ شدن مدل است. در واقع امکان دارد مدل درحال دنبال کردن نویزهای سیستم باشد.
در صورتی که داده‌ها از نویز کمی آمده باشند یا رابطه خیلی غیر خطی باشد یا تعداد داده‌ها بسیار زیاد باشد اما متغیرهای پیش‌بینی کننده کم باشند، بهتر است که از یک مدل $flexible$ استفاده شود.
۷.
الف)

obs	X1	X2	X3	Y	distance from (0,0,0)
1	0	3	0	Red	3
2	2	0	0	Red	2
3	0	1	3	Red	3.16
4	0	1	2	Green	2.24
5	-1	0	1	Green	1.41
6	1	1	1	Red	1.73

ب)

$k = 1 \Rightarrow \ Y = Green$

ج)

$k = 3 \Rightarrow Y = Red$

د)

باید k کوچک باشد چون با k های بزرگ، مدل نقاط نزدیک به هم را بسیار بیشتر شبیه به هم پیشبینی می‌کند چون همسایه ها مشابهی هم خواهند داشت. اما به دنبال مدلی هستیم که بیشتر غیرخطی باشد و تغییرات سریعتر را هم پیشبینی کند. پس باید فلکسیبل‌تر باشد.

۸.

در فایل $Q8.R$ پاسخ داده شده.