Homework 1

Xiao-Qun Wang, Zili Ma

Q1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors p is small.

(b) The number of predictors p is extremely large, and the number of observations n is small.

(c) The relationship between the predictors and response is highly non-linear.

(d) The variance of the error terms, i.e. $\sigma^2 = \mathrm{Var}(\varepsilon)$, is extremely high.

Answer 1:

(a) Flexible method is better. For an extremely large sample size, the data set provides enough freedom to figure out a more complex flexible model, which give us a better estimation.

(b) Flexible method is worse than inflexible method. Flexible method can lead to the overfitting when the sample only has a limited observation number.

(c) Flexible method is better. For highly non-linear correlation, flexible method can give a better estimation, if the sample size is large enough.

(d) Flexible method performs worse in this case. The closely fitting result derived from the flexible method will overfit with a large variance the sample whose variance is very large.

Q2. We now revisit the bias-variance decomposition.

(a) Provide a sketch **of typical (squared) bias, variance, training error, test error,** and **Bayes (or irreducible) error curves**, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent.

(b) Explain why each of the five curves has the shape displayed in part (a).

Answer 2:

(a) See the hand-written page in the last part.

(b)

1. The value of **bias** continuously decreases when the flexibility increases. Because the assumed function used in the learning method is typically different from the function. Therefore, a method with less flexibility would introduce more error between the true function and our assumption than a more flexible method, in other words, during the estimation process, the inflexible method can result in more bias.
2. **Variance** value increases monotonically when the flexibility increases. Because for a more flexible method, the regression curve/plane in p dimension will approach to all the sample point more closely.
3. **Bayes error** or **irreducible** error keeps constant below training and test MSE curves for different flexibility. Due to the trade-off relationship between bias and variance, MSE is always larger than Bayer error.
4. The value of **training MSE** continuously decreases when the flexibility increase. Because increasing flexibility of the model will lead to the function of result approach to the training sample point closely or even pass it. Then, parameters will be very sensitive to those new data points. It results in an increasingly larger variance of the model.
5. **Test MSE** value has a U-shape when the flexibility increase. At the beginning, increased flexibility will decrease test value until the function reaches a reasonable fitting to the training sample. However, when the flexibility continues to increase, the function trained by the learning method will start to fit those points whose “patterns” in fact are caused by random noise. The overfitting to training samples increases the test MSE.

Q3. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

Answer 3:

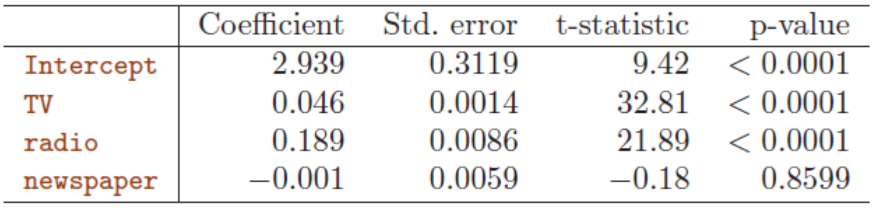
The **parametric statistical learning approach** uses some explicit assumptions to solve the problem. It reduces the problem of estimating f down to one of estimating a set of parameters. .The model using for estimation sometimes is given empirically, such as SLR or MLR. Typically. they involve a two-step model based approach

Non-parametric methods do not make explicit assumptions about the functional form of *f*. The fitting or learning process is done numerically to avoid the danger of using wrong model in the study.

Its advantages and disadvantages are as follows:

Methods	Advantages	Disadvantages
Parametric	1. Simplifies the problem by the given explicit model. 2. Typically, a smaller number of observations is needed in parametric methods than that of non-parametric approaches required	1. The pre-specific model could be not true, or not accurate enough to explain the data set in details. 2. Overfitting occurs when the model is too flexible.
Non-parametric	1. By avoiding the assumption of a particular functional form for *f*, they have the potential to accurately fit a wider range of possible shapes for *f*.	1. Typically, without the simplification of the problem, they require far more number of observation to fit datasets. 2. Overfitting occurs if a too low level of smoothness is used.

Q6. Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, *TV, radio,* and *newspaper,* rather than in terms of the coefficients of the linear model.



	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Answer 6:

In Table 3.4, with the corresponding p-value (smaller than 0.05) is significant, the coefficient for TV represents the average effect of increasing newspaper spending by $1,000 while holding radio and newspaper fixed. For a given amount of radio and newspaper advertising, spending an additional $1,000 on TV advertising leads to an increase in sales by approximately 46 units.

With the corresponding p-value (smaller than 0.05) is significant, the coefficient for radio represents the average effect of increasing newspaper spending by $1,000 while holding TV and newspaper fixed. For a given amount of TV and newspaper advertising, spending an additional $1,000 on radio advertising leads to an increase in sales by approximately 189 units.

The coefficient estimate for newspaper in this multiple regression model is close to zero, and the corresponding p-value (larger than 0.05) is no longer significant.

Q7. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the *i*th fitted value takes the form:

$$\hat{y}_i = x_i\hat{\beta},$$

Where

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i y_i\right) \Big/ \left(\sum_{i'=1}^{n} x_{i'}^2\right). \qquad (3.38)$$

Show that we can write

$$\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'}.$$

What is $a_{i'}$?

Answer 7:

See the attached paper written by hand.

Q8. Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$.

Answer 8:

The equation of (3.4) is as follows:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (3.4)$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x},$$

From the 2nd equation, if we substitute $\bar{\mathbf{x}}$ into this rearrange equation as follows,

$$\mathbf{y} = \hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_1\bar{\mathbf{x}} = \bar{\mathbf{y}}$$

Therefore, we can conclude that the linear regression passes through the point **$(\bar{\mathbf{x}}, \bar{\mathbf{y}})$.**

.