ZHANG Weiwen 20583632 wzhangbu@connect.ust.hk

1. 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. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.

Ans: (a) Better. Because n is large but p is small, model would less likely to be overfitted.

(b) worse. Model is very likely to be overfitted. Thus performance would getting worse

(c) better. More flexible can fit data better if model's non-linear

(d) Worse. Flexible model would be affected by errors and would diverge away from true situation.

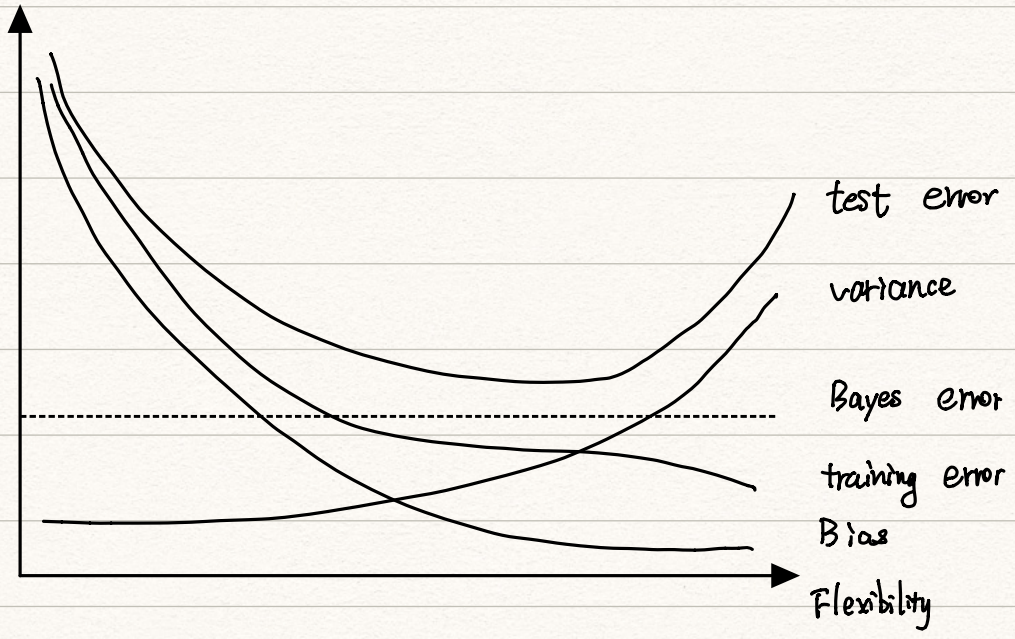
3. We now revisit the bias-variance decomposition.

   (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

the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

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

Ans: (a)



(b) As the model becoming more flexible, it would fit the training set better and better. Thus the traing error would decrease and bias would be smaller as well. But when flexibility increases excessively, it would be overfitted and test error would stop decreasing and turn increasing instead. As the flexibility increases, variance would increase in the whole process. And irreducible error would be constant because training can do nothing to it.

1. 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.

Ans: Null hypotheses should be, above terms have no relationship with sales in multiple regression i.e. $\beta_1 = \beta_2 = \beta_3 = 0$ While based on p-values, we can draw conclusion that TV and Radios are associated with sales, and newspapers and sales have no relationship in multiple regression.

5. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form

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

where

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

Show that we can write

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

What is $a_{i'}$?

Note: We interpret this result by saying that the fitted values from linear regression are linear combinations *of the response values.*

Ans: $\hat{y}_i = x_i \hat{\beta} = x_i \dfrac{\sum_{j=1}^{n} x_j y_j}{\sum_{j'=1}^{n} x_{j'}^2}$

$= \dfrac{\sum_{j=1}^{n} x_j y_j x_i}{\sum_{j'=1}^{n} x_{j'}^2}$

$$= \sum_{j=1}^{n} \frac{x_j x_j}{\sum_{j'=1}^{n} x_{j'}^2} \cdot y_j$$

adjust index $$= \sum_{i'=1}^{n} \frac{x_{i'} x_i}{\sum_{j=1}^{n} x_j^2} y_{i'}$$

So $$a_i' = \frac{x_i x_i'}{\sum_{j=1}^{n} x_j^2}$$

1. Using a little bit of algebra, prove that (4.2) is equivalent to (4.3). In other words, the logistic function representation and logit representation for the logistic regression model are equivalent.

Ans. Know that (4.2) $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$

(4.3) $\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$

(4.3) $\Rightarrow$ $p(x) + p(x) e^{\beta_0+\beta_1 x} = e^{\beta_0+\beta_1 x}$

$p(x) = e^{\beta_0+\beta_1 x}(1-p(x))$

$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$ $\Leftrightarrow$ (4.5)

2. It was stated in the text that classifying an observation to the class for which (4.12) is largest is equivalent to classifying an observation to the class for which (4.13) is largest. Prove that this is the case. In other words, under the assumption that the observations in the kth class are drawn from a $N(\mu_k, \sigma^2)$ distribution, the Bayes' classifier assigns an observation to the class for which the discriminant function is maximized.

Ans: (4.12): $p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(x-\mu_k)^2)}{\sum_{l=1}^{k} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(x-\mu_l)^2)}$

Noted that $(x-\mu_k)^2 = x^2 - 2x\mu_k + \mu_k^2$

Now we take log and rearrange:

$$\log P_k(x) = \log(\pi_k) + \log\left[\frac{\frac{1}{\sqrt{2\pi}\sigma}\left(-\frac{1}{2\sigma^2}(x^2)\right)}{\sum_{l=1}^{K}\pi_l\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}\right]$$

$$+\left(-\frac{1}{2\sigma^2}(-2x\mu_k+\mu_k^2)\right)$$

$$\Rightarrow \log P_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) + \log\left[\frac{\frac{1}{\sqrt{2\pi}\sigma}\left(-\frac{1}{2\sigma^2}(x^2)\right)}{\sum_{l=1}^{K}\pi_l\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}\right]$$

where last term is independent of k.

Then the class that maximizes above equation is equivalent to the one that maximizes

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

3. This problem relates to the QDA model, in which the observations within each class are drawn from a normal distribution with a class-specific mean vector and a class specific covariance matrix. We consider the simple case where $p = 1$; i.e. there is only one feature.

   Suppose that we have K classes, and that if an observation belongs to the kth class then X comes from a one-dimensional normal distribution, $X \sim N(\mu_k, \sigma_k^2)$. Recall that the density function for the one-dimensional normal distribution is given in (4.11). Prove that in this case, the Bayes' classifier is *not* linear. Argue that it is in fact quadratic.

   Hint: For this problem, you should follow the arguments laid out in Section 4.4.2, but without making the assumption that $\sigma_1^2 = \ldots = \sigma_K^2$.

Ans: Known that $f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$

And by Bayes theorem where $P(Y=k|X=x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K}\pi_l f_l(x)}$

we find $P(Y=k|X=x) = \frac{\pi_k\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K}\frac{1}{\sqrt{2\pi}\sigma_l}\exp\left(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2\right)}$

take the log and rearrange:

$$\log P_k(x) = \log(\pi_k) - \log(\sigma_k) + \log\left(\frac{\frac{1}{\sqrt{2\pi}}}{\sum_{l=1}^{K} \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2\right)}\right) - \frac{x^2}{2\sigma_k^2} + \frac{\mu_k x}{\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2}$$

In this way, Baye's classifier is qualtratic in terms of x^2.