STAT 8330 FALL 2015 ASSIGNMENT 1

Peng Shao

September 6, 2015

▶ **Exercises 2.5. Solution.**

(1).

- advantage: can fit many different functional forms; low bias; usually predict more accurately
- disadvantage: overfitting problem; sually hard to interpret; high variance

(2). If our goal is to predict more accurately, it will usually be best to choose a more flexible approach.

(3). If our goal is to make some inferences, we prefer choosing a less flexible approach because the relation between response and predictor is more explicit.

▶ **Exercises 2.6. Solution.**

(1). The essential difference between parametric and non-parametric approach is that, the parametric make an assumption of the form of f, which can reduce problem of estimating f down to one of estimating a set of parameter, but non-parametric do not make explicit assumptions about the functional form of f.

(2).

- advantage: it is easier to estimate parameter; the relation between response and predictor is more explicit;
- disadvantage: the model we choose will usually not match the true unknown form of f; sometimes need more assumption.

▶ **Exercises 2.10. Solution.**

▶ **Exercises 3.5. Solution.**

▶ **Exercises 3.15. Solution.**

▶ **Exercises 4.3. Solution.**

▶ **Exercises 4.10. Solution.**

We know that we classify X into kth class based on Bayes' classifier if

$$p_k(x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

is largest among all $p_l(x)$, $l = 1, 2, ..., K$. For 1 dimension, the density of x from kth class is

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$$

In comparing two classes k and l, it is sufficient to look at the log-ratio, and we see that

$$\begin{aligned}
\log\left(\frac{p_k(x)}{p_l(x)}\right) &= \log\left(\frac{\pi_k}{\pi_l}\right) + \log\left(\frac{f_k(x)}{f_l(x)}\right) \\
&= \log\left(\frac{\pi_k}{\pi_l}\right) + \log\left(\frac{\sigma_l}{\sigma_k}\right) - \frac{(x-\mu_k)^2}{2\sigma_k^2} + \frac{(x-\mu_l)^2}{2\sigma_l^2} \\
&= \left(-\frac{(x-\mu_k)^2}{2\sigma_k^2} - \log\sigma_k + \log\pi_k\right) - \left(-\frac{(x-\mu_l)^2}{2\sigma_l^2} - \log\sigma_l + \log\pi_l\right) \\
&= \delta_k(x) - \delta_l(x)
\end{aligned}$$

Then the Beyes' classifier can be be defined as

$$C(x) = \arg\max_k \delta_k(x)$$

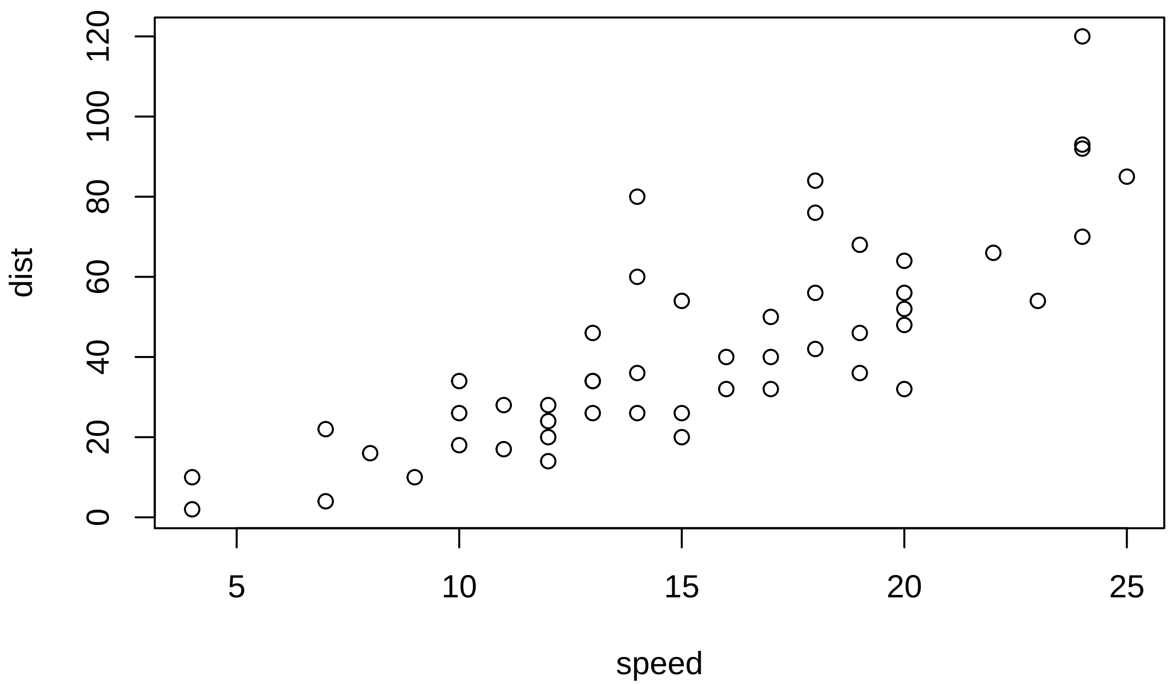
▶ **Exercises 4.13.** **Solution.**

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot. s