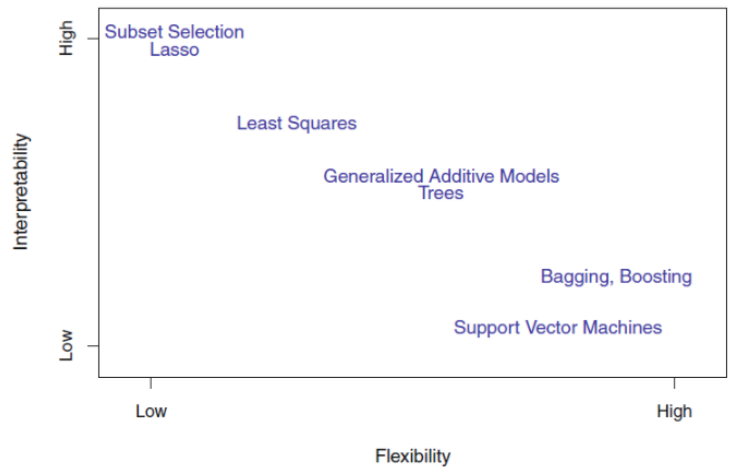
Chapter 9.1/9.2
Introduction

The **supervised learning** means the outcomes from a **sample space** are accessible, which could be integrated with the corresponding predictors, the input, to generate a statistical model. The **cluster analysis** is an example of **unsupervised learning**.

The supervised statistical models consist of a variety of methodologies, from some linear ones to those are purely nonlinear. In this chapter, the **generalized additive models**, **trees**⋯, are used to deal with nonlinear types of data, to some degree. Sometimes, the balance should be achieved between **flexibility** and **interpretability**, **prediction** and **inference**, **variance** and **bias**, **mean squared error** and **overfitting**. Thus, the pros and cons often coexist among different statistical models, as the following picture illuminates, depending on our problems to solve.

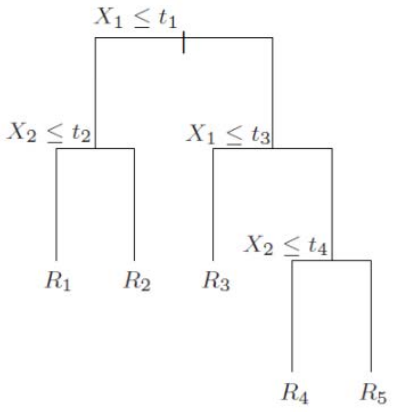


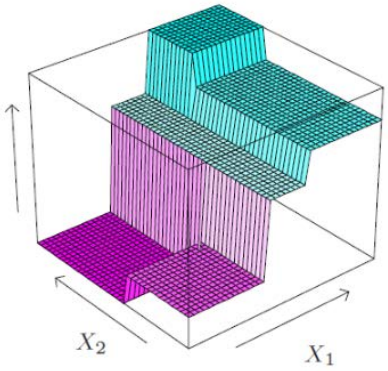
9.1.1
In this chapter, the text talks about the concepts of GAM corresponding to **regression settings** and **classification settings**, which reflect the continuous or discrete types of **response** values. Some deeper background knowledge of **regression** is necessary here.

The reading also includes some technical issues about how to **fit** additive models-the **algorithms**, referring to its building blocks as the **scatterplot smoother**, specifically speaking, the **cubic smoothing spline**.

9.1.2-9.1.3
Examples are given to illustrate the procedure of GAM. The summary mentions some limitations to it.

9.2
9.2.1
The **tree-based regression and classification (CART)** are good at interpretation, at the cost of the lower predictive ability.





9.2.2

A **generalized tree function** is given as:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m). \qquad (9.10)$$

to minimize the bias or sum of squared error, **Cm** is parameterized as

$$\hat{c}_m = \text{ave}(y_i | x_i \in R_m). \qquad (9.11)$$

A **greedy algorithm** is proposed to feasibly minimize SE, even though potentially being short-sighted.

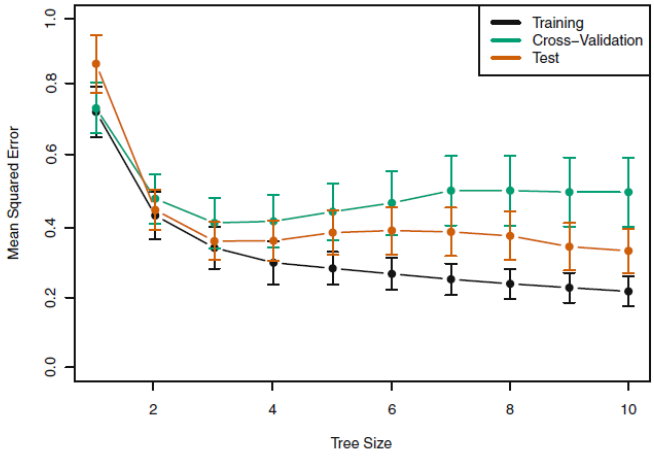
The increment of the **tree size** leads to more flexibility, as well as the possibility of overfitting, which should be controlled by a process called, **cost-complexity pruning**.

$$\begin{aligned} N_m &= \#\{x_i \in R_m\}, \\ \hat{c}_m &= \frac{1}{N_m} \sum_{x_i \in R_m} y_i, \\ Q_m(T) &= \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2, \end{aligned} \qquad (9.15)$$

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|. \qquad (9.16)$$

For certain α, a subtree could be found to minimize Cα(T); on the other hand, the α could be adaptively selected by five or tenfold **cross-validation.** The final subtree, after being pruned, is thus determined.

The relationship between the mean squared error and the tree size of CV, training, test samples in an example.



9.2.3

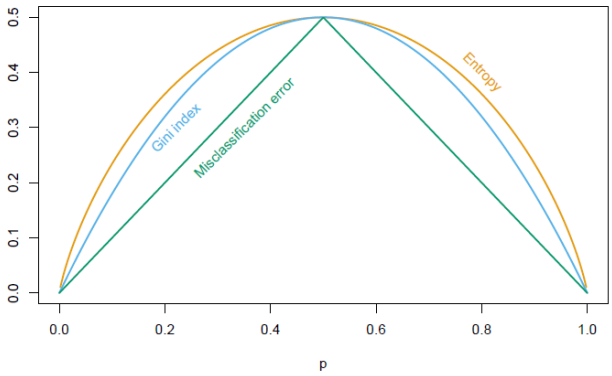
The **classification tree** is a little different from the regression tree, as the response values are of a qualitative nature. The tree is growing in the same way by **recursive binary splitting.** But the 'SE' is replaced by the **classification error rate** as:

$$E = 1 - \max_k(\hat{p}_{mk}).$$

defining the **class** of this node as:

$$\text{class } k(m) = \arg\max_k \hat{p}_{mk}$$

The **Gini index** and **cross-entropy** are also used to reveal the impurity of certain nodes, especially when **growing** the tree (here's an example); in pruning situations, each of these three could be selected.



Corresponding functions:

$$\begin{aligned}
&\text{Misclassification error:} && \tfrac{1}{N_m}\textstyle\sum_{i\in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}.\\
&\text{Gini index:} && \textstyle\sum_{k\neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk}).\\
&\text{Cross-entropy or deviance:} && -\textstyle\sum_{k=1}^{K} \hat{p}_{mk}\log\hat{p}_{mk}.
\end{aligned} \tag{9.17}$$

9.2.4
Other specific and advanced issues about tree-based method are discussed.

9.2.5
An example of the classification tree.
The notions of **sensitivity**, **specificity**, and **receiver operating characteristic curves of** trees and GAM models.