1 Textbook Problem

1.1 Problem 5.1

1) Show that the truncated power basis functions in (5.3) represent a basis for a cubic spline with the two knots as indicated.

The truncated power basis functions are given by:

$$h_1(X) = 1 \quad h_3(X) = X^2 \quad h_5(X) = (X - \xi_1)_+^3$$

$$h_2(X) = X \quad h_4(X) = X^3 \quad h_6(X) = (X - \xi_2)_+^3$$

If these equations represent a basis for a cubic spline with two knots at ξ_1 and ξ_2, then we can write *any* cubic spline with knots at these locations as a linear combination of these basis functions.

By definition, any cubic spline with two knots at ξ_1 and ξ_2 is cubic in the regions $X < \xi_1$, $\xi_1 < X < \xi_2$, and $X > \xi_1$, and continuous. Therefore, write a function $f(z)$ as:

$$f(z) = \begin{cases} a_1X^3 + b_1X^2 + c_1X + d_1 & for \quad X \leq \xi_1 \\ a_2X^3 + b_2X^2 + c_2X + d_2 & for \quad \xi_1 \leq X \leq \xi_2 \\ a_3X^3 + b_3X^2 + c_3X + d_3 & for \quad X \geq \xi_2 \end{cases}$$

However, this function is subject to the condition of continuity at the boundary, implying that:

$$a_1\xi_1^3 + b_1\xi_1^2 + c_1\xi_1 + d_1 - a_2\xi_1^3 - b_2\xi_1^2 - c_2\xi_1 - d_2 = 0$$

$$a_3\xi_2^3 + b_3\xi_2^2 + c_3\xi_2 + d_3 - a_2\xi_2^3 - b_2\xi_2^2 - c_2\xi_2 - d_2 = 0$$

Now, clearly, the three lines in the function definition of $f(X)$ can be formed by simple linear combinations of $h_1(X)$, $h_2(X)$, $h_3(X)$, and $h_4(X)$. So it just remains to show that the basis functions $h_5(X)$ and $h_6(X)$ can be used to enforce the constraints.

Consider the first region; neither $h_5(X)$ nor $h_6(X)$ will contribute since both are 0 in this region. Hence, we see that for the first region, we can match the function exactly if we choose:

$$\tilde{f}(X) = a_1h_4(X) + b_1h_3(X) + c_1h_2(X) + d_1h_1(X)$$

Now, consider that we choose a coefficient γ for $h_5(z)$. Then, in the second region, we have:

$$(a_1 + \gamma)X^3 + (b_1 - 3\xi_1\gamma)X^2 + \left(c_1 + 3\xi_1^2\gamma\right)X + \left(d_1 - \gamma\xi_1^3\right)$$

If this is to match the equation for $f(z)$ in this region, we see that we need:

$$a_1 + \gamma = a_2 \Rightarrow a_1 - a_2 = -\gamma$$

$$b_1 - 3\xi_1\gamma = b_2 \Rightarrow b_1 - b_2 = 3\xi_1\gamma$$

$$c_1 + 3\xi_1^2\gamma = c_2 \Rightarrow c_1 - c_2 = 3\xi_1^2\gamma$$

$$d_1 - \gamma \xi_1^3 = d_2 \Rightarrow d_1 - d_2 = \xi_1^3 \gamma$$

Note, we can transform these four restrictions into the first constraint discussed earlier by noting that if these conditions are satisfied:

$$(a_1 - a_2)\,\xi_1^3 + (b_1 - b_2)\,\xi_1^2 + (c_1 - c_2)\,\xi_1 + (d_1 - d_2) = -\gamma \xi_1^3 + 3\gamma \xi_1^3 - 3\gamma \xi_1^3 + \gamma \xi_1^3 = 0$$

Therefore, satisfying the four conditions that cause the basis expansion to match $f(z)$ in region 2 is equivalent to satisfying the boundary constraint of the cubic spline. Now, we assume that γ **is** chosen so that the expansion matches $f(z)$ in region 2, and we repeat the same procedure for $h_6(z)$. Assuming we have a coefficient of ζ for $h_6(z)$. Then in the third region, the expansion is given by:

$$(a_2 + \zeta)\,X^3 + (b_2 - 3\zeta\xi_2)\,X^2 + \left(c_2 + 3\zeta\xi_2^2\right) X + \left(d_2 - \zeta\xi_2^3\right)$$

Again, if we are to select ζ so that this expansion matches $f(z)$ in the third region, we must have:

$$a_2 - a_3 = -\zeta$$

$$b_2 - b_3 = 3\zeta\xi_2$$

$$c_2 - c_3 = -3\zeta\xi_2^2$$

$$d_2 - d_3 = \zeta\xi_1^3$$

We can again transform these four restrictions into the second constraint now:

$$(a_2 - a_3)\,\xi_2^3 + (b_2 - b_3)\,\xi_2^2 + (c_2 - c_3)\,\xi_2 + (d_2 - d_3) = -\zeta\xi_2^3 + 3\zeta\xi_2^3 - 3\zeta\xi_2^3 + \zeta\xi_2^3 = 0$$

Since satisfying the four conditions that cause the basis expansion to match $f(z)$ in region 3 is equivalent to satisfying the boundary constraint of the cubic spline in that region, we can conclude that for any cubic spline with continuity constraints, it can be expressed as a weighted sum of the six suggested basis functions, where the weights chosen to make the function match the function definition will satisfy the boundary constraints.

2 Decision Tree Problem

2.1 Tree Root

Using information theory to measure the entropy, calculate which attribute is best to select for the root node of the tree.

Let CH be a random variable representing Credit History, D a random variable representing Debt, CO a random variable representing Collateral, I a random variable representing Income, and R a random variable representing Risk. Further, let $CH = -1$ be *bad*, $CH = 0$ be *unknown*, $CH = 1$ be *good*, $D = 0$ be low, $D = 1$ be high, $C = 0$ be none, $C = 1$ be adequate, $I = -1$ be \$0 to \$15k, $I = 0$ be \$15k to \$35k, $I = 1$ be over \$35k, $R = -1$ be low risk, $R = 0$ be moderate risk, and $R = 1$ be high risk.

2.2 Conservative Estimates

For this tree, we will estimate conservatively for all parameters. Therefore, the groupings we make will be:

1. Unknown Credit History is the same as Bad Credit History
2. \$0 to \$15K Income is the same as \$15K to \$30K Income
3. High Risk is the same as Moderate Risk

Then, we are presented with four possible decisions to make:

1. Bad vs. Good Credit History
2. Low vs. High Debt
3. None vs. Adequate Collateral
4. Low vs. High Income

Let $\mu_1^{(k)}$ denote a positive decision on the k^{th} choice, and $\mu_2^{(k)}$ denote a negative decision. Then, we have to evaluate four impurities at the root node:

$$I\left(\mu_1^{(1)}\right) = -\left(\frac{2}{9}\log_2\frac{2}{9} + \frac{7}{9}\log_2\frac{7}{9}\right) = 0.7642 \qquad I\left(\mu_2^{(1)}\right) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0.7642$$

$$I\left(\mu_1^{(2)}\right) = -\left(\frac{3}{7}\log_2\frac{3}{7} + \frac{4}{7}\log_2\frac{4}{7}\right) = 0.9852 \qquad I\left(\mu_2^{(2)}\right) = -\left(\frac{2}{7}\log_2\frac{2}{7} + \frac{5}{7}\log_2\frac{5}{7}\right) = 0.9710$$

$$I\left(\mu_1^{(3)}\right) = -\left(\frac{3}{11}\log_2\frac{3}{11} + \frac{8}{11}\log_2\frac{8}{11}\right) = 0.8454 \qquad I\left(\mu_2^{(3)}\right) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) = 0.9183$$

$$I\left(\mu_1^{(4)}\right) = -\left(\frac{0}{8}\log_2\frac{0}{8} + \frac{8}{8}\log_2\frac{8}{8}\right) = 0 \qquad I\left(\mu_2^{(3)}\right) = -\left(\frac{5}{6}\log_2\frac{5}{6} + \frac{1}{6}\log_2\frac{1}{6}\right) = 0.6500$$

$$I\left(\mu_1^{(1)}, \mu_2^{(1)}\right) = \frac{9}{14}I\left(\mu_1^{(1)}\right) + \frac{5}{14}I\left(\mu_2^{(1)}\right) = 0.8381$$

$$I\left(\mu_1^{(2)}, \mu_2^{(2)}\right) = \frac{1}{2}I\left(\mu_1^{(2)}\right) + \frac{1}{2}I\left(\mu_2^{(2)}\right) = 0.9242$$

$$I\left(\mu_1^{(3)}, \mu_2^{(3)}\right) = \frac{11}{14}I\left(\mu_1^{(3)}\right) + \frac{3}{14}I\left(\mu_2^{(3)}\right) = 0.8682$$

$$I\left(\mu_1^{(4)}, \mu_2^{(4)}\right) = \frac{8}{14}I\left(\mu_1^{(4)}\right) + \frac{6}{14}I\left(\mu_2^{(4)}\right) = 0.2786$$

Therefore, to maximize ΔI, we choose question (4) as the first question. Now we note that, once this decision is made, if the case is low-income, it will be high risk regardless of the other factors. Therefore, we conclude that this branch of the tree can be terminated. Therefore, we move on to the cases of high-income.

Now, we have three decisions to choose from. Let $\mu_{21}^{(k)}$ denote a positive decision on the k^{th} choice, and $\mu_{22}^{(k)}$ denote a negative decision. Then, we have to evaluate four impurities at the root node:

$$I\left(\mu_{21}^{(1)}\right) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) = 0.9183 \quad I\left(\mu_{22}^{(1)}\right) = -\left(\frac{3}{3}\log_2\frac{3}{3} + \frac{0}{3}\log_2\frac{0}{3}\right) = 0$$

$$I\left(\mu_{21}^{(2)}\right) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right) = 0.8113 \quad I\left(\mu_{22}^{(2)}\right) = -\left(\frac{2}{2}\log_2\frac{2}{2} + \frac{0}{2}\log_2\frac{0}{2}\right) = 0$$

$$I\left(\mu_{21}^{(3)}\right) = -\left(\frac{3}{3}\log_2\frac{3}{3} + \frac{0}{3}\log_2\frac{0}{3}\right) = 0 \quad I\left(\mu_{22}^{(3)}\right) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) = 0.9183$$

$$I\left(\mu_{21}^{(1)}, \mu_{22}^{(1)}\right) = \frac{1}{2}I\left(\mu_{21}^{(1)}\right) + \frac{1}{2}I\left(\mu_{22}^{(1)}\right) = 0.4592$$

$$I\left(\mu_{21}^{(2)}, \mu_{22}^{(2)}\right) = \frac{2}{3}I\left(\mu_{21}^{(2)}\right) + \frac{1}{3}I\left(\mu_{22}^{(2)}\right) = 0.5409$$

$$I\left(\mu_{21}^{(3)}, \mu_{22}^{(3)}\right) = \frac{1}{2}I\left(\mu_{21}^{(3)}\right) + \frac{1}{2}I\left(\mu_{22}^{(3)}\right) = 0.4592$$

We have a tie, so we choose question (1) as the next question. It becomes clear at this point that the data cannot be separated into a perfect tree, because two training instances have the same inputs and different outputs. Further, there is only one case whose risk is labeled as high at this point. Therefore, no matter how the tree is further split, this case will never be separated from the others. Further division will only increase the complexity of the tree without any appreciable difference in quality, so we conclude the tree at this point.

2.3 High Risk Estimates

For this tree, we will group parameters more liberally. Therefore, the groupings we make will be:

1. Unknown Credit History is the same as Good Credit History
2. \$15 to \$35K Income is the same as over \$35K Income
3. Low Risk is the same as Moderate Risk

Then, we are presented with four possible decisions to make:

1. Bad vs. Good Credit History
2. Low vs. High Debt
3. None vs. Adequate Collateral
4. Low vs. High Income

Let $\mu_1^{(k)}$ denote a positive decision on the k^{th} choice, and $\mu_2^{(k)}$ denote a negative decision. Then, we have to evaluate four impurities at the root node:

$$I\left(\mu_1^{(1)}\right) = -\left(\frac{3}{10}\log_2\frac{3}{10} + \frac{7}{10}\log_2\frac{7}{10}\right) = 0.8813 \quad I\left(\mu_2^{(1)}\right) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right) = 0.8113$$

$$I\left(\mu_1^{(2)}\right) = -\left(\frac{4}{7}\log_2\frac{4}{7} + \frac{3}{7}\log_2\frac{3}{7}\right) = 0.9852 \quad I\left(\mu_2^{(2)}\right) = -\left(\frac{2}{7}\log_2\frac{2}{7} + \frac{5}{7}\log_2\frac{5}{7}\right) = 0.8631$$

$$I\left(\mu_1^{(3)}\right) = -\left(\frac{3}{3}\log_2\frac{3}{3} + \frac{0}{3}\log_2\frac{0}{3}\right) = 0 \quad I\left(\mu_2^{(3)}\right) = -\left(\frac{6}{11}\log_2\frac{6}{11} + \frac{5}{11}\log_2\frac{5}{11}\right) = 0.9980$$

$$I\left(\mu_1^{(4)}\right) = -\left(\frac{4}{4}\log_2\frac{4}{4} + \frac{0}{4}\log_2\frac{0}{4}\right) = 0 \quad I\left(\mu_2^{(3)}\right) = -\left(\frac{3}{10}\log_2\frac{3}{10} + \frac{7}{10}\log_2\frac{7}{10}\right) = 0.8813$$

$$I\left(\mu_1^{(1)}, \mu_2^{(1)}\right) = \frac{9}{14}I\left(\mu_1^{(1)}\right) + \frac{5}{14}I\left(\mu_2^{(1)}\right) = 0.8613$$

$$I\left(\mu_1^{(2)}, \mu_2^{(2)}\right) = \frac{1}{2}I\left(\mu_1^{(2)}\right) + \frac{1}{2}I\left(\mu_2^{(2)}\right) = 0.9242$$

$$I\left(\mu_1^{(3)}, \mu_2^{(3)}\right) = \frac{11}{14}I\left(\mu_1^{(3)}\right) + \frac{3}{14}I\left(\mu_2^{(3)}\right) = 0.7818$$

$$I\left(\mu_1^{(4)}, \mu_2^{(4)}\right) = \frac{8}{14}I\left(\mu_1^{(4)}\right) + \frac{6}{14}I\left(\mu_2^{(4)}\right) = 0.6295$$

Again, we can maximize ΔI if we choose question (4) as the first question. Just as in the other case, we see that for low-income cases, the risk is always high. Therefore, this branch of the tree does not need to be further divided. If we consider only high-income cases, we evaluate further divisions:

$$I\left(\mu_{21}^{(1)}\right) = -\left(\frac{1}{8}\log_2\frac{1}{8} + \frac{7}{8}\log_2\frac{7}{8}\right) = 0.5436 \quad I\left(\mu_{22}^{(1)}\right) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 0.8113$$

$$I\left(\mu_{21}^{(2)}\right) = -\left(\frac{5}{7}\log_2\frac{5}{7} + \frac{2}{7}\log_2\frac{2}{7}\right) = 0.8631 \quad I\left(\mu_{22}^{(2)}\right) = -\left(0\log_2 0 + \log_2 1\right) = 0$$

$$I\left(\mu_{21}^{(3)}\right) = -\left(\frac{5}{5}\log_2\frac{5}{5} + \frac{0}{5}\log_2\frac{0}{5}\right) = 0 \quad I\left(\mu_{22}^{(3)}\right) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0.9710$$

$$I\left(\mu_{21}^{(1)}, \mu_{22}^{(1)}\right) = \frac{8}{10}I\left(\mu_{21}^{(1)}\right) + \frac{2}{10}I\left(\mu_{22}^{(1)}\right) = 0.6349$$

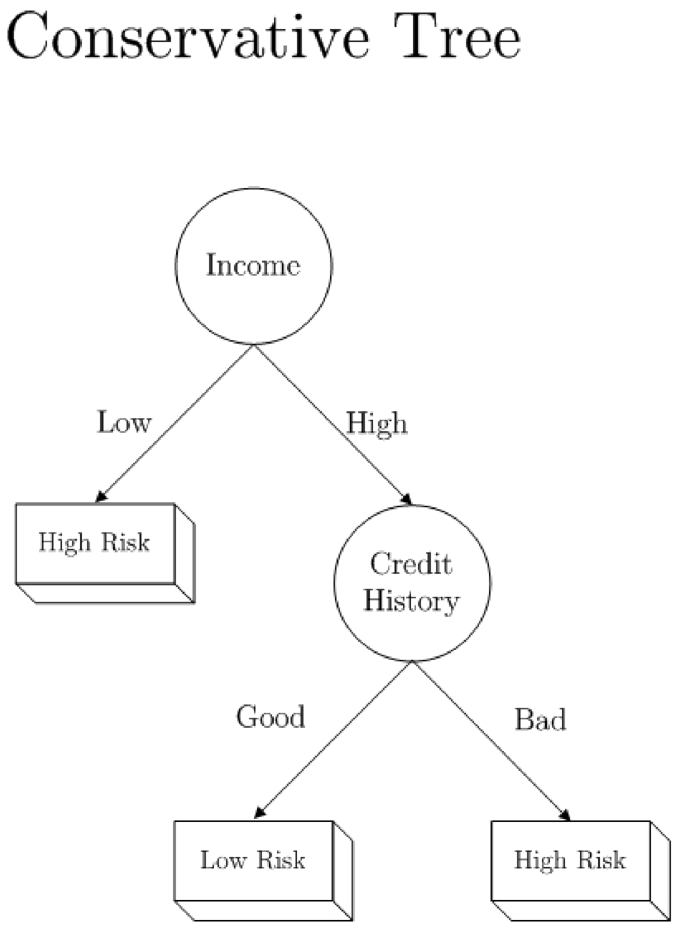
$$I\left(\mu_{21}^{(2)}, \mu_{22}^{(2)}\right) = \frac{5}{10} I\left(\mu_{1}^{(2)}\right) + \frac{5}{10} I\left(\mu_{2}^{(2)}\right) = 0.4855$$

$$I\left(\mu_{21}^{(3)}, \mu_{22}^{(3)}\right) = \frac{7}{10} I\left(\mu_{21}^{(3)}\right) + \frac{3}{10} I\left(\mu_{22}^{(3)}\right) = 0.6042$$

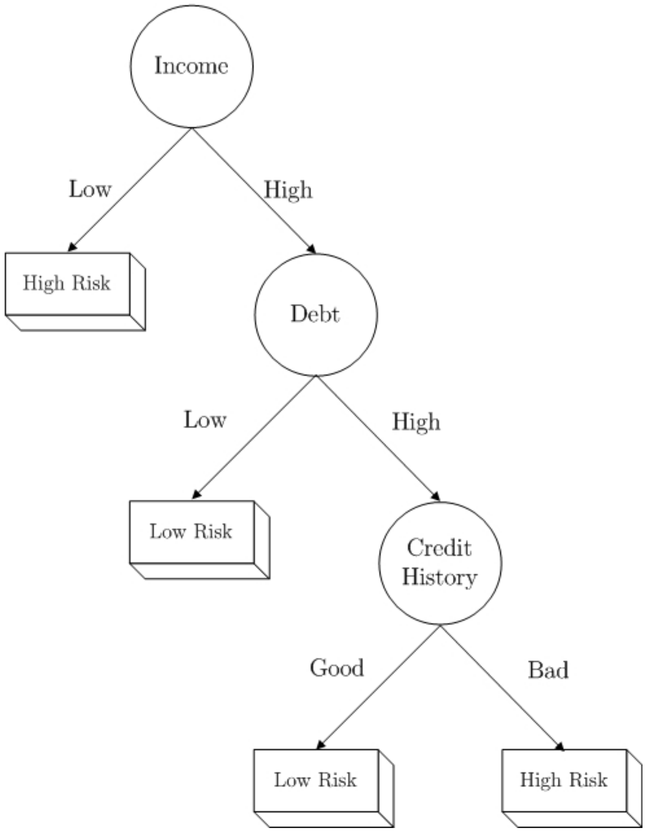
Therefore, we can conclude that to maximize ΔI, we choose question (3) as the next question. Now, we note that since the low-debt cases have a defined ouput, that branch of the tree does not need to be reduced further. It is visible that for the other branch, one of the two remaining high-risk cases is isolated by splitting on Credit History, but the remaining data is inseparable because two cases have the same inputs with different outputs. Therefore, we have reached the end of our tree design.

2.4 Tree Comparisons

The illustrations of the two decision trees are shown below:



Risky Business Tree



The same decision is made at the root of both trees, but the decision at the next level is different. In both cases, any inputs with low income are immediately classified as high risk. The conservative tree is smaller than the risky tree, and both trees will make exactly one classification error due to a lack of separability in the input data.

2.5 Construction of a Ternary Tree

A tree with ternary, rather than binary, decisions and classification can be determined by simply changing the impurity measurements to include three terms, instead of the two terms being considered. That is, we would take our usual functions $I\left(\mu_1^{(k)},\mu_2^{(k)}\right)$ into $I\left(\mu_1^{(k)},\mu_2^{(k)},\mu_2^{(k)}\right)$. The cardinality measures would be approximately the same, we would just have to be considering 3 terms instead of 2. Alternatively, we could combine two binary trees by considering, instead of a ternary output, two sets of binary outputs. one tree could then have the other tree as its ”parent node”.