1. (a) i. There are six terminal nodes. Three variables are used (Petal.length, Petal.width, Sepal.length). **[2 Marks]** (Seen Similar)

   ii. Following the tree from the root. At the first junction we examine the Petal.length variable which is 5.2cm, which is bigger than 2.45, so the first decision is Right. Then, for this individual, Petal.width is 1.6cm, which means we go left. The next variable is Petal.length again of 5.2cm, which is bigger than 4.95, so we go Right, which results in the output classification for this individual of `virginica`. **[2 Marks]** (Went over similar example quickly in class)

   (b) The problem is to fit a constant over a region with a least squares criterion. It is well known that the optimal estimator is the mean of those observations over the region. Or, with calculus, we can differentiate the error with the constant model and solve the equation:

$$2\sum_{i=1}^{n}(Y_i - \hat{c}) = 0, \tag{1}$$

   hence result. **[2 Marks]** (Simplest Regression Case)

   (c) The criterion for split variable j and split point s is

$$\min_{j,s}\left\{\min_{c_1}\sum_{X_i\in R_1(j,s)}(Y_i - c_1)^2 + \min_{c_2}\sum_{X_i\in R_2(j,s)}(Y_i - c_2)^2\right\}. \tag{2}$$

   The inner minimisations are solved using the method for $\hat{c}_m$ in part (b). Then, for a given j, the SSQ reduces to finding s that minimises

$$\sum_{X_i\in R_1(j,s)}(Y_i - \hat{c}_1)^2 + \min_{c_2}\sum_{X_i\in R_2(j,s)}(Y_i - \hat{c}_2)^2. \tag{3}$$

   This optimisation is easier than it looks. The expression only changes when s crosses a X_i boundary, so there are only $n+1$ quickly computable values of the expression in (3) and we choose the s that gives the minimum.

   We then repeat this calculation over all $j = 1, \ldots, p$ variables and choose the variable that reduces the overall SSQ the most. This is only $\mathcal{O}\{p(n+1)\}$ calculations. **[5 Marks]** (Bookwork)

   (d) A tree such as T_0 severely overfits the data. The tree is to reflective of the particular data set used to fit it and will probably not generalise well to new data.

The cost-complexity measure balances the measure of fit ($\sum_m n_m Q_m(T)$) with the size of the tree $|T|$. We want to minimise the combined quantity C_α. If α is small then we tend to grow a large tree which severely overfits. If α is large, then we are penalised for growing a large tree, so the tree tends to be small, but does not reflect the structure in the data. So, we need to choose a good α in between to balance the twin goals of goodness of fit and size of tree. **[3 Marks]** (Bookwork)

(e) For example. Classification trees are simple to interpret and explain and they are fast to compute. They can be inefficient in some cases (see next example) and also they are not continuous in their inputs: small changes in the input data can result in radically different trees. **[2 Marks]** (Bookwork)

(f) For example: two clusters aligned just above and just below $y = x$ and separated by that line. Trees are inefficient here as they need multiple splits on each axis to provide a "staircase" separator, whereas a linear discriminator might do a better job. A second possible example: one cluster is a ring around a second cluster centred in the middle of the ring (but separated from it). If this structure was hidden in many dimensions trees would find it hard to follow the ring. However, projection pursuit classification (based on regression used as a classifier) might find it easier to discover. **[4 Marks]** (Unseen, requires imagination)

2. (a) Define $K_h(x) = h^{-1}K(x/h)$.The kernel density estimator of $f(x)$ is

$$\hat{f}_h(x) = n^{-1}\sum_{i=1}^{n} K_h(X_i - x). \tag{4}$$

The KDE for $f(x, y)$ is

$$\hat{f}(x, y) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i)K_h(y - Y_i). \tag{5}$$

[2 Marks] (Bookwork)

(b) The conditional expectation $\mathbb{E}(Y|X = x)$ is

$$\begin{aligned} \mathbb{E}(Y|X = x) &= \int y f(y|x)\, dy & (6)\\ &= \int y \frac{f(x, y)}{f(x)}\, dy & (7)\\ &= \frac{\int y f(x, y)\, dy}{f(x)}. & (8) \end{aligned}$$

To estimate this, we can replace the densities f by their estimators $\hat{f}$ from part (a). Obtaining

$$\hat{\mathbb{E}}(Y|X=x) = \frac{\int y\hat{f}(x,y)}{\hat{f}(x)} \tag{9}$$

$$= \frac{\sum_{i=1}^{n} K_h(x-X_i)\int yK_h(y-Y_i)\,dy}{\sum_{i=1}^{n} K_h(x-X_i)}, \tag{10}$$

and the n^{-1} cancel. Now let's focus attention on the integral and substitute $v = y - Y_i$

$$\int yK_h(y-Y_i)\,dy = \int (v+Y_i)K_h(v)\,dv \tag{11}$$

$$= \int vK_h(v)\,dv + Y_i\int K_h(v)\,dv = Y_i \tag{12}$$

(the first integral $\to 0$, the second $\to 1$)

The first integral is zero as K is a symmetric kernel and

$$\int K_h(v)\,dv = h^{-1}\int K(v/h)\,dv = \int K(u)\,du = 1. \tag{13}$$

by substituting $u = v/h$ for the second integral. This then gives the required solution. **[3 Marks]** (Bookwork)

(c) The local quadratic fit follows the discontinuity the best, but seems to have a slightly too high variance to the left of the discontinuity, with some oscillations — although it's bias looks good throughout. The N-W and local linear estimator are good, but don't fit the discontinuity as well as the local quadratic. The N-W seems to be suffering from a bit of boundary bias at the right hand end. None of them really model the discontinuity that well. **[3 Marks]** (Seen Similar)

(d) i. Applying W to $y = f+\epsilon$ gives $Wy = Wf+W\epsilon \implies w = d+e$. **[1 Mark]**(Bookwork)

ii. Each e_i is a linear combination of ϵ_j for $j = 1,\ldots,n$ and since the sum of two normals is normal, this means that e_i must be normally distributed. Suppose the l.c. for e_i is given by coefficients $\{b_j^{(i)}\}_{j=1}^n$. Then

$$\mathbb{E}(e_i) = \mathbb{E}(\sum_{j=1}^{n} b_j^{(i)}\epsilon_j) = \sum_{j=1}^{n} b_j^{(i)}\mathbb{E}(\epsilon_j) = 0. \tag{14}$$

Finally,

$$\text{var}(e_i) = \text{var}\left\{\sum_{j=1}^{n} b_j^{(i)}\epsilon_j\right\} = \sum_{j=1}^{n} b_j^{(i)2}\,\text{var}(\epsilon_j) = \sigma^2\sum_{j=1}^{n} b_j^{(i)2} = \sigma^2, \tag{15}$$

since W is an orthogonal matrix, its columns have unit length. Hence, putting these three facts together implies $e_i \sim N(0, \sigma^2)$. **[4 Marks]** (Bookwork)

(e) i. For wide classes of functions, the wavelet coefficients of a function are sparse, in that most are zeros and few are non-zero and, typically, reasonably large. Hence, these can be modelled by a prior distribution where most are zero and a few are large. This translates to choosing γ large (to give a large coefficient), τ very small, close to zero, to give a density very concentrated around zero, its mean. Finally, p controls the proportion of non-zeroes in the prior representation of a function. For a sparsely represented function, you would expect p to be small, e.g. if $p = 0.02$, then two in a hundred of the coefficients would expect to be non-zero. **[3 Marks]** (Familiar, but not gone into detail before)

ii. Using standard Bayesian theory we have that

$$f(d|w) \propto f(w|d)f(d), \tag{16}$$

where $f(d) = g(d)$ is the prior distribution of the coefficients and $f(w|d)$ is the likelihood. Since, $w = d + e$ and e has the iid $N(0, \sigma^2)$ distribution above, we know that $w_i|d_i \sim N(d_i, \sigma^2)$. Since everything is independent, we can treat each coefficient separately. Let us write down all of our density functions. The two in the prior are

$$\begin{aligned}
\phi_{0,\tau^2}(d) &= (2\pi\tau^2)^{-1/2}\exp\{-d^2/(2\tau^2)\} && (17)\\
\phi_{0,\gamma^2}(d) &= (2\pi\gamma^2)^{-1/2}\exp\{-d^2/(2\gamma^2)\} && (18)\\
\phi_{d,\sigma^2}(w) &= (2\pi\sigma^2)^{-1/2}\exp\{-(w-d)^2/(2\sigma^2)\}. && (19)
\end{aligned}$$

So, using (16) we have to multiply the prior by likelihood. The prior is the sum of two very similar terms. For the first one:

$$\begin{aligned}
A &= \phi_{d,\sigma^2}(w)\phi_{0,\tau^2}(d) && (20)\\
&= K_1\exp\{-(w-d)^2/(2\sigma^2)\}\exp\{-d^2/(2\tau^2)\} && (21)\\
&= K_1\exp\{-(w-d)^2/(2\sigma^2) - d^2/(2\tau^2)\}, && (22)
\end{aligned}$$

where K_1 is some constant. Let's look at the argument of the exponential

$$\frac{(w-d)^2}{2\sigma^2}+\frac{d^2}{2\tau^2}=\frac{\tau^2(w-d)^2}{2\sigma^2\tau^2}+\frac{\sigma^2 d^2}{2\sigma^2\tau^2}=\frac{\tau^2(w-d)^2+\sigma^2 d^2}{2\sigma^2\tau^2}. \tag{23}$$

Rearranging the numerator

$$\begin{aligned}\tau^2w^2-2wd\tau^2+\tau^2d^2+\sigma^2d^2 &= \tau^2w^2-2wd\tau^2+(\sigma^2+\tau^2)d^2\\ &= (\sigma^2+\tau^2)\{d^2-2wdr^2+r^2w^2\}\end{aligned}$$

where $r^2=\tau^2/(\sigma^2+\tau^2)$. Now fiddling with the quadratic term gives

$$\begin{aligned}d^2-2wdr^2+r^2w^2 &= d^2-2wdr^2+w^2r^4-w^2r^4+r^2w^2\\ &= (d-r^2w)^2+w^2r^2(1-r^2)\end{aligned} \tag{24}$$

Since we are only interested in terms associated with d, because of the proportionality in (16), we can absorb the $w^2r^2(1-r^2)$ term (when divided by some other non-d factors and exponentiated) in the constant K_1 to form a new constant K_2. Then, the only thing in the exponential we are interested in is:

$$(\sigma^2+\tau^2)(d-r^2w)^2/(2\sigma^2\tau^2)=(d-r^2w)^2/(2\sigma^2r^2) \tag{25}$$

Hence, the first density is $\phi_{r^2w,r^2\sigma^2}$ and, similarly, the second one will be $\phi_{s^2w,s^2\sigma^2}$, where $s^2=\gamma^2/(\gamma^2+\sigma^2)$. Thus, the posterior density is

$$f(d|w)=(1-p)\phi_{r^2w,r^2\sigma^2}(d)+p\phi_{s^2w,s^2\sigma^2}(d). \tag{26}$$

The overall posterior mean can be obtained from the individual posterior means and weighted in the same way, giving:

$$d^*=(1-p)r^2w+ps^2w=w\{(1-p)r^2+ps^2\}. \tag{27}$$

[N.b.: an identical development, using a slightly different notation can be found in Section 3.10.1 of the Wavelet book by Nason (2008), as mentioned in lectures.] **[4 Marks]** (Unseen, but overview mentioned in lectures)