Continuation of Assignment 2

Question 2: Decision Tree The table below contains a small training set. Each line includes an individual's education, occupation choice, years of experience and an indication of salary. Your task is to create a complete decision tree including the number of low's and highs, entropy at each step and the information gain for each feature gained at each node in the tree.

Instance	Education Level	Career	Years of Experience	Salary
1	High School	Management	Less than 3	Low
2	High School	Management	3 to 10	Low
3	College	Management	Less than 3	High
4	College	Service	More than 10	Low
5	High School	Service	3 to 10	Low
6	College	Service	3 to 10	High
7	College	Management	More than 10	High
8	College	Service	Less than 3	Low
9	High School	Management	More than 10	High
10	High School	Service	More than 10	Low

The base entropy, from 4 lows and 6 highs is

$$\text{info}[4,6] = -\frac{4}{10}\log_2\frac{4}{10} + -\frac{6}{10}\log_2\frac{6}{10} = 0.971$$

To determine which feature to have the first node on, calculate the info gain from education, career and years of experience respectively.

$$\begin{aligned}\text{info}[[4,1],[2,3]] = \text{info}[0.722, 0.971] = 0.8465 \rightarrow \text{Gain} = 0.1245\\ \text{info}[[2,3].[4,1]] = \text{info}[0.971, 0.733] = 0.8465 \rightarrow \text{Gain} = 0.1245\\ \text{info}[[2,1],[2,1],[2,2]] = \text{info}[0.918, 0.918, 1] = 0.9508 \rightarrow \text{Gain} = 0.020\end{aligned}$$

Education gives the greatest entropy gain. If high school is chosen, then its base info is

$$\text{info}[4,1] = -\frac{4}{5}\log_2\frac{4}{5} + -\frac{1}{5}\log_2\frac{1}{5} = 0.722$$

Determine whether to have the next node on career or experience, respectively.

$$\begin{aligned}\text{info}[[2,1],[2,0]] = \text{info}[0.918, 0] = 0.551 \rightarrow \text{Gain} = 0.171\\ \text{info}[[1,0],[2,0],[1,1]] = \text{info}[0,0,1] = 0.4 \rightarrow \text{Gain} = 0.322\end{aligned}$$

Years of experience gives the greatest entropy gain. Finally, career has a base info of

$$\text{info}[1,1] = -\frac{1}{2}\log_2\frac{1}{2} + -\frac{1}{2}\log_2\frac{1}{2} = 1$$

If college was chosen, then its base info is

$$\text{info}[2,3] = -\frac{2}{5}\log_2\frac{2}{5} + -\frac{3}{5}\log_2\frac{3}{5} = 0.971$$

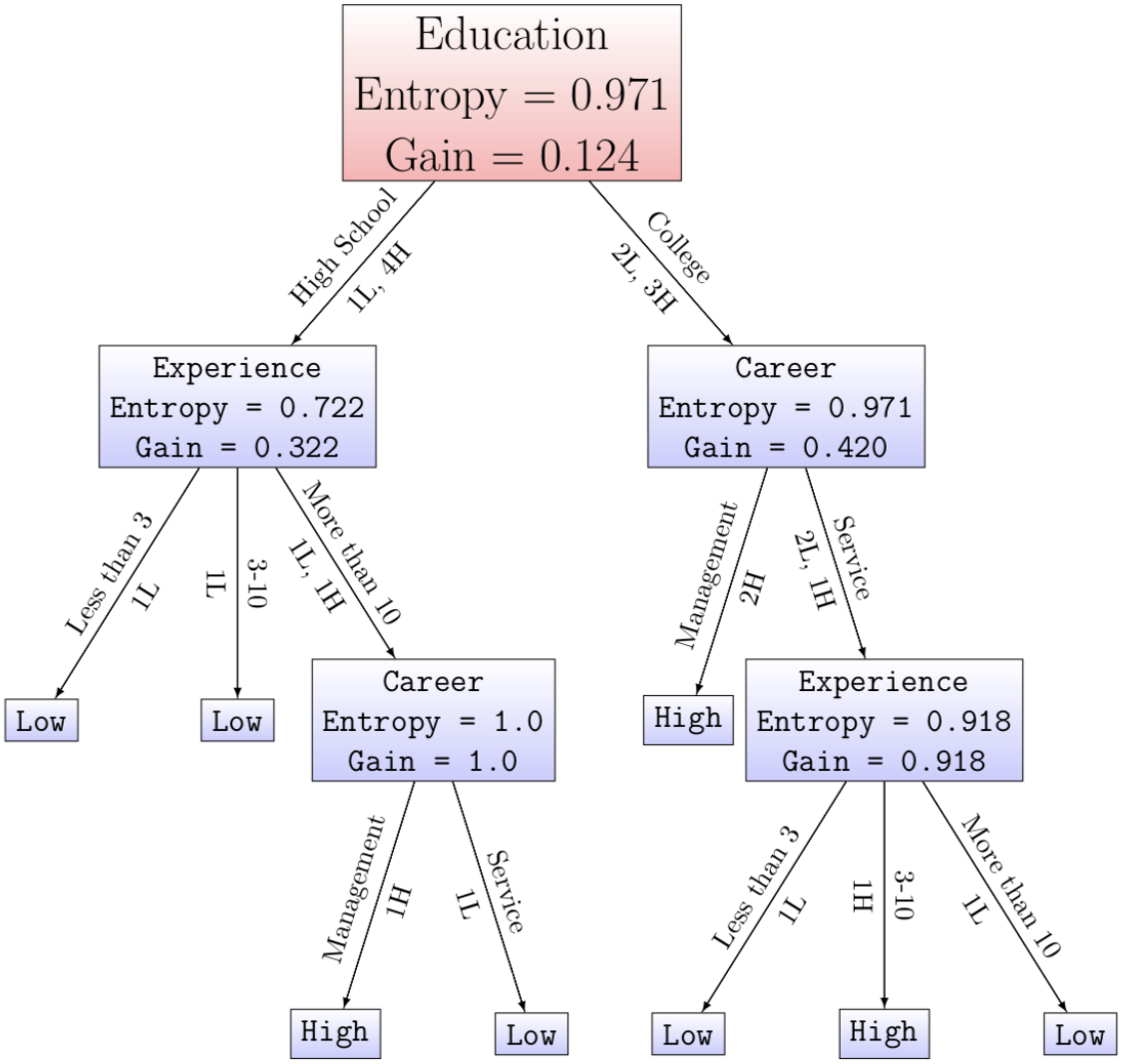
Determine whether the information gain from career or experience is greater.

$$\text{info}[[0,2],[2,1]] = \text{info}[0,0.918] = 0.551 \rightarrow \text{Gain} = 0.420$$
$$\text{info}[[1,1],[0,1],[1,1]] = \text{info}[1,0,1] = 0.8 \rightarrow \text{Gain} = 0.171$$

Career gives the greatest entropy gain. Finally, years of experience has a base info of

$$\text{info}[2,1] = -\frac{2}{3}\log\frac{2}{3} + -\frac{1}{3}\log\frac{1}{3} = 0.918$$

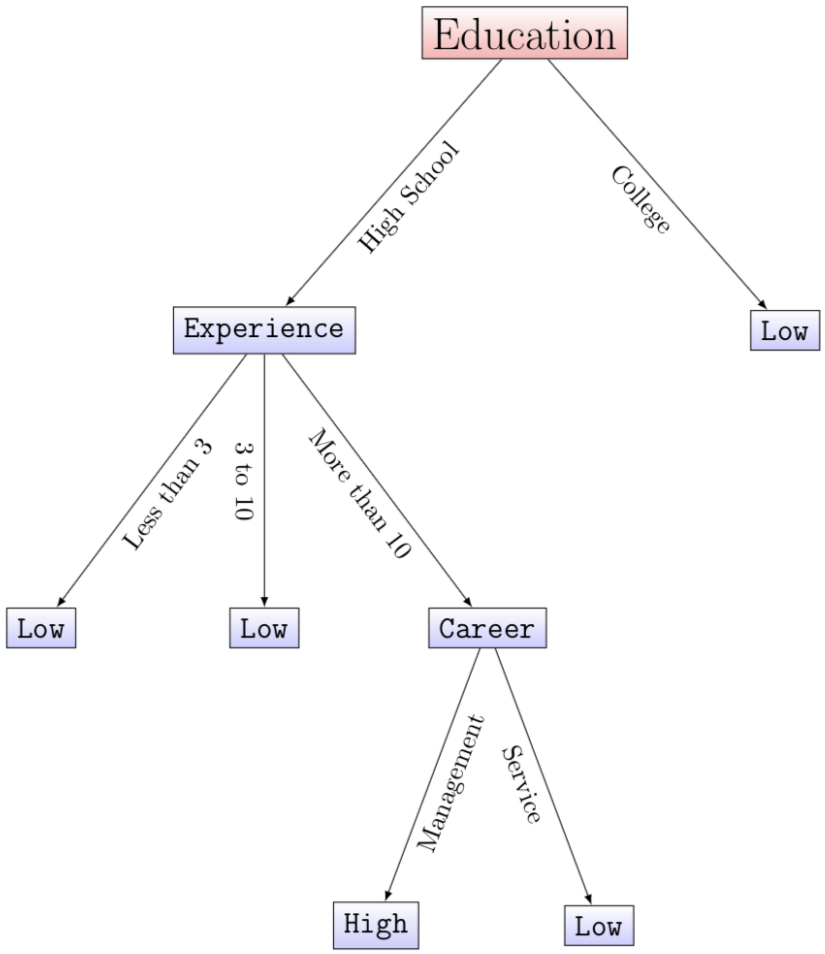
The decision tree is displayed below.



Prune the tree above using the validation data below. Show all work.

Instance	Education Level	Career	Years of Experience	Salary
1	High School	Management	More than 10	High
2	College	Management	Less than 3	Low
3	College	Service	3 to 10	Low

Instance 1 appears with the proper salary in the decision tree above. Instance 2 and 3 show a different salary than predicted. Going down the college path and picking management gives a predicted high salary whereas instance 2 gives a low salary. Replace that leaf node with L. Now, going down the college path and picking service and 3 to 10 years of experience gives a predicted high salary whereas instance 3 gives a low salary. Replace that leaf node with L. Note that now all leaf nodes extended from the experience stem is low salary; therefore replace the entire stem with L. Also note that now both careers predict low salary; therefore replace the entire career stem with L. The new decision tree after pruning looks like the following.



Question 3: SVM using Weka
Apply SVM with several different kernels and hyper-parameter choices to the **veh-prime.arff** file. Import this file into Weka and then select the SMO classifier found under classifiers/function. Use 10 fold cross-validation. Make kernel and hyper-parameter choices by clicking on "SMO ... " appearing next to Choose.
Make 5 runs of the algorithm. Select PolyKernel with exponent option 1, 2 and 4. Then select RBFKernel with gamma set to 0.01 and 1.0. For each run, record the number of correctly and incorrectly classified instances. Explain why some of the choices do not work well.

Kernel and Hyper-Parameter	# of Correct Instances	# of Incorrect Instances	Accuracy
PolyKernel, exponent = 1	717	129	84.75%
PolyKernel, exponent = 2	810	36	95.74%
PolyKernel, exponent = 4	791	55	93.49%
RBFKernel, gamma = 0.01	614	232	72.57%
RBFKernel, gamma = 1.0	764	82	90.30%

The polynomial kernel is defined as

$$K(x, y) = (1 + x^T y)^Q$$

The exponent will control the nature of the polynomial function to be determined. When the exponent is 1, the model under fits the data whereas when the exponent is 4, the model overfits the data. The exponent chosen in between, 2, fits the data better with a higher number of correct instances.

The RBF kernel is defined as

$$K(x, y) = e^{-\gamma||x-y||^2}$$

The γ value will control the nature of the exponential function to be determined. For low gamma values, the shape of the bell-shaped curve representing the distances between the vectors is wide. Therefore it will create a lot of inaccuracies and under fit the data. For high gamma values, the shape of the bell-shaped curve is narrow and so a lot more data points will be captured by the model. Thus a model with a high gamma value will do better than one with a low gamma value.

Question 4: Kernels
Assume that $x = (x_1, x_2)$ is a two dimensional vector and function K is defined as

$$K(x, z) = x_1 \cdot z_1 + x_1 \cdot e^{z_2} + z_1 \cdot e^{x_2} + e^{x_2+z_2}$$

Prove that K is a kernel.

$K(x, z)$ is a kernel if the Mercer's condition is fulfilled. Namely, $K(x, z)$ is a kernel if and only if it is symmetric and if the following matrix is positive semi-definite for any x, z.

$$X = \begin{bmatrix} K(x, x) & K(x, z) \\ K(z, x) & K(z, z) \end{bmatrix}$$

To show that $K(x, z)$ is symmetric, show that $K(x, z) = K(z, x)$.
Suppose $x = (x_1, x_2)$ and $z = (z_1, z_2)$. Then

$$\begin{aligned} K(x, z) &= x_1 \cdot z_1 + x_1 e^{z_2} + z_1 e^{x_2} + e^{x_1 + x_2} \\ &= x_1(z_1 + e^{z_2}) + e^{x_2}(z_1 + e^{z_2}) \\ &= (x_1 + e^{x_2})(z_1 + e^{z_2}) \end{aligned}$$

This is equivalent to saying

$$K(z, x) = (z_1 + e^{z_2})(x_1 + e^{x_2})$$

Therefore

$$K(x, z) = K(z, x)$$

and so $K(x, z)$ is symmetric. This proves the first part of Mercer's condition.

To show that X is positive semi-definite, show that $\det(X) \geq 0$ and $\text{tr}(X) \geq 0$.
First rewrite X using the full equations.

$$\begin{aligned} K(x, x) &= K((x_1, x_2), (x_1, x_2)) = (x_1 + e^{x_2})^2 \\ K(x, z) &= K((x_1, x_2), (z_1, z_2)) = (x_1 + e^{x_2})(z_1 + e^{z_2}) \\ K(z, x) &= K((z_1, z_2), (x_1, x_2)) = (z_1 + +e^{z_2})(x_1 + e^{x_2}) \\ K(z, z) &= K((z_1, z_2), (z_1, z_2)) = (z_1 + e^{z_2})^2 \end{aligned}$$

So

$$X = \begin{bmatrix} (x_1 + e^{x_2})^2 & (x_1 + e^{x_2})(z_1 + e^{z_2}) \\ (z_1 + +e^{z_2})(x_1 + e^{x_2}) & (z_1 + e^{z_2})^2 \end{bmatrix}$$

Prove that $\det(X) \geq 0$.

$$\begin{aligned} \det(X) &= (x_1 + e^{x_2})^2 (z_1 + e^{z_2})^2 - (x_1 + e^{x_2})(z_1 + e^{z_2})(z_1 + +e^{z_2})(x_1 + e^{x_2}) \\ &= (x_1 + e^{x_2})^2 (z_1 + e^{z_2})^2 - (x_1 + e^{x_2})(x_1 + e^{x_2})(z_1 + +e^{z_2})(z_1 + e^{z_2}) \\ &= (x_1 + e^{x_2})^2 (z_1 + e^{z_2})^2 - (x_1 + e^{x_2})^2 (z_1 + e^{z_2})^2 \\ &= 0 \end{aligned}$$

Prove that $\text{tr}(X) \geq 0$.

$$\text{tr}(X) = \underbrace{(x_1 + e^{x_2})^2}_{\geq 0} + \underbrace{(z_1 + e^{z_2})^2}_{\geq 0} \geq 0$$

Since both properties of matrix X being positive semi-definite is fulfilled, matrix X is positive semi-definite and so the second part of Mercer's condition is fulfilled.

Hence K is a kernel. $\square$