Problem 1 (60%). Consider the training data shown below. Here, $A$, $B$, and $C$ are attributes, and $Y$ is the class label.

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$C$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$y$</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>$y$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>$y$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>$y$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>$n$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$n$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$n$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>$n$</td>
</tr>
</tbody>
</table>

Suppose that we consider only decision trees each having 3 nodes (i.e., a root node and two leaves). Give the decision tree with the best empirical error. You need to explain your reasoning.

Answer. For the given input, there are only 6 possible decision trees having 3 nodes, which are:

(a) \[ \begin{array}{c}
 A \downarrow \rightarrow y \\
 y \\
 n 
\end{array} \]

(b) \[ \begin{array}{c}
 B \downarrow \rightarrow y \\
 y \\
 n 
\end{array} \]

(c) \[ \begin{array}{c}
 C \downarrow \rightarrow y \\
 y \\
 n 
\end{array} \]

(d) \[ \begin{array}{c}
 A \downarrow \rightarrow n \\
 n \\
 y 
\end{array} \]

(e) \[ \begin{array}{c}
 B \downarrow \rightarrow n \\
 n \\
 y 
\end{array} \]

(f) \[ \begin{array}{c}
 C \downarrow \rightarrow n \\
 n \\
 y 
\end{array} \]

Among them, the decision tree (b) has the lowest empirical error $1/4$ and, hence, is the answer.

Problem 2 (40%). Use the generalization theorem (in Lecture Notes 1) to estimate the generalization error of your decision tree in Problem 1. Again, we consider only the decision trees with 3 nodes. Your estimate should be correct with probability at least 99%.

Answer. Let $S$ be the training set given in Problem 1 and $H$ be the set of classifiers that can possibly be returned. Denote by $h$ the best decision tree we found in Problem 1. From the above solution, we know $|H| = 6$ and the empirical error $\text{err}_S(h) = 1/4$.

According to the generalization theorem, with probability at least $1 - \delta$, we have

\[
\text{err}_D(h) \leq \text{err}_S(h) + \sqrt{\frac{\ln(1/\delta) + \ln |H|}{2|S|}} \\
\leq 1/4 + \sqrt{\frac{\ln(1/\delta) + \ln 6}{16}}.
\]
By setting $\delta = 0.01$, we know with probability at least 0.99,

$$err_D(h) \leq 1/4 + \sqrt{\ln(1/0.01) + \ln 6 \over 16}.$$