Written assignment

1. Consider the following decision tree: (a) Draw the decision boundaries defined by this tree. Each leaf of the tree is labeled with a letter. Write this letter in the corresponding region of input space.

   (b) Give another decision tree that is syntactically different but defines the same decision boundaries. This demonstrates that the space of decision trees is syntactically redundant. Is this redundancy a statistical problem (i.e., does it affect the accuracy of the learned trees)? Is it a computational problem (i.e., does it increase or decrease the computational complexity of finding an accurate tree)?

2. In the basic decision tree algorithm, we choose the feature/value pair with the maximum information gain as the test to use at each internal node of the decision tree. Suppose we modified the algorithm to choose at random from among those feature/value combinations that had non-zero mutual information, but that we kept all other parts of the algorithm unchanged.

   (a) Prove that if a splitting feature/value combination has non-zero mutual information at an internal node, then at least one training example must be sent to each of the child nodes.

   (b) What is the maximum number of leaf nodes that such a decision tree could contain if it were trained on \( m \) training examples?

   (c) What is the maximum number of leaf nodes that a decision tree could contain if it were trained on \( m \) training examples using the original maximum mutual information version of the algorithm? Is it bigger, smaller, or the same as your answer to (b)?

   (d) How do you think this change would affect the accuracy of the decision trees produced on average? Why?

3. Consider the following training set:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Learn a decision tree from the training set shown above using the information gain criterion.

4. **MAP estimation.** Consider the problem of linear regression. We are given a set of observed data points \((X_i, t_i) : i = 1, \ldots, N\), where \(X\) is the input vector, and \(t\) is the target output. The goal is to estimate a set of linear coefficients \(W\) such that \(t\) can be predicted by \(W^TX\). In particular, we assume that \(t|X \sim N(W^TX, \sigma^2)\). Now we further assume that each coefficient \(w_i\) has a prior distribution \(N(0; \alpha^{-1})\). Please write down the posterior function of \(W\), and show that maximizing this posterior is equivalent to minimizing the least square objective with a \(L_2\) regularization term.

5. Use leave-one-out cross validation to select \(K\) for \(K\)-NN among three possible choices of \(K = 1, 2, 3\) for the following one dimensional data.

```
\[ \times \times \circ \circ \circ \circ \circ \circ \]
```

6. Please provide a neural network that expresses the following boolean function:

\[ (X_1 \land \neg X_2) \lor (X_2 \land X_3) \]

7. In class, we present the sum of squared error as the loss function for learning neural nets. We now consider a maximum likelihood objective for learning neural networks. In particular, consider a binary classification problem where the training examples are denoted by \((x_i, y_i), i = 1, \ldots, n\), and \(y_i \in \{0, 1\}\). We consider a network with a single output node whose activation function is a logistic sigmoid:

\[ \hat{y} = \sigma(a) = \frac{1}{1+\exp(-a)} \]

where \(a\) is the input to the activation function and we will interpret the output \(\hat{y}\) as the conditional probability of \(p(y = 1|x)\). Please derive the log likelihood function of \(w\), where \(w\) is the weights of the network. (Hint: this is essentially very similar to logistic regression objective).