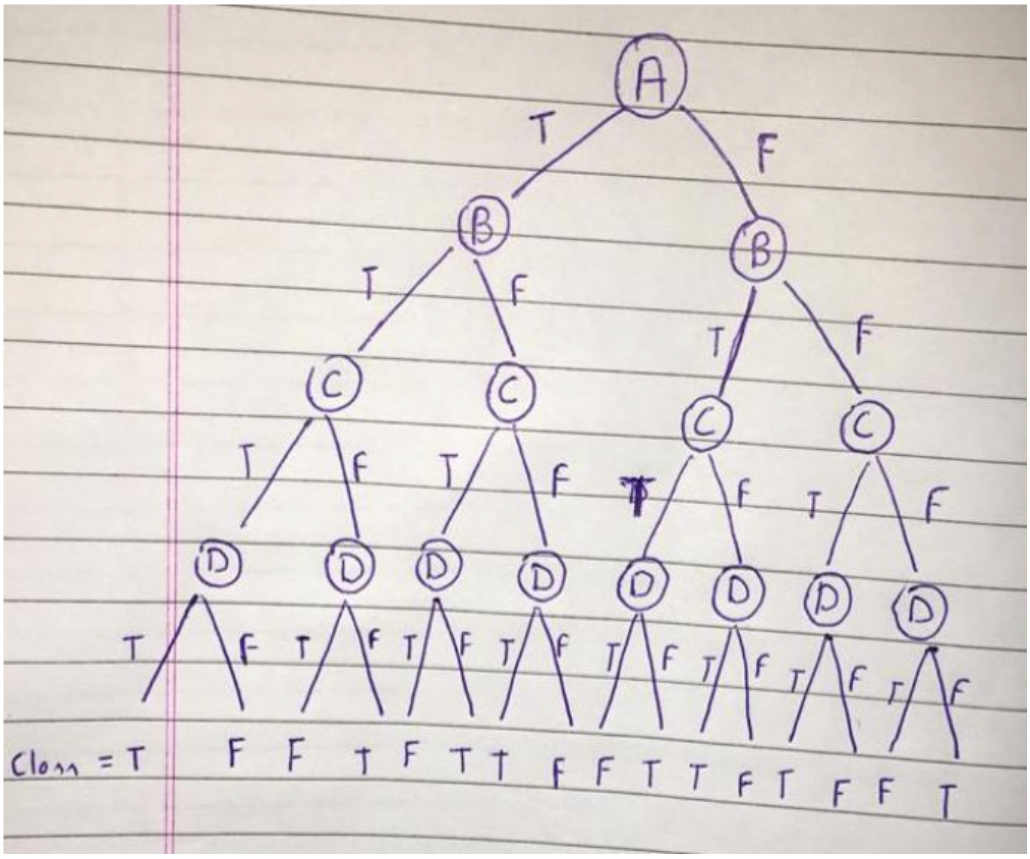
1.1 Tan, Chapter 3

Exercise 2:

Question: Draw a full decision tree for the odd parity function, where only when the count of True is odd is the class label True, of four Boolean attributes A,B,C and D. Is it possible to simplify the tree?



This tree cannot be simplified, because if there exactly even no of attributes are True which means at each stage we want the value of all 4 attributes. Therefore, we cannot simplify the tree.

Exercise 3:

Question: Consider the training examples showing in Table 3.5 for a binary classification problem.

1) Compute the Gini index for the overall collection of training examples

   Gini = 1- 2*[(10/20)^2] = 0.5. The 10 is from the amount of each class there is since there are a total of 20 classes but 10 per class then the total gini is 1-two times that.

2) Compute the Gini index for the Customer ID attribute

   The Gini index of Customer ID is 0. Since the gini index of one customer is 0 so for all of the customer ID its 0.

3) Compute the Gini index for the Gender attribute

   The Gini index of Gender is 1-(6/10)^2-(4/10)^2=0.48. The weighted gini index of each gender is 0.48.

4) Compute the Gini index for the Car Type attribute using multiway split

   The Gini of a Family car is 1-[(1/4)^2]+[(3/4)^2]=0.375. The Gini Index for Sports Car is 0. The Gini of a Luxury Car is 1-[(1/8)^2+(7/8)^2] = 0.218. The Gini Index of the Car Type is (4/20)*0.375+(8/20)*0+(8/20)*0.218 = 0.1625

5) Compute the Gini index for the Shirt Size attribute using multiway split

   The Gini Index for Small Shirt Size is 1-[(3/5)^2+(2/5)^2]=0.48. The Gini Index of Medium Shirt Size is 1-[(3/7)^2+(4/7)^2]=0.49. The Gini Index of a Large Shirt Size is 1-[(2/4)^2+(2/4)^2]=0.5. The Gini Index of an Extra Large Shirt Size is 1-[(2/4)^2+(2/4)^2]=0.5. The total Gini Index is (5/20)*0.48+(7/20)*0.49+(4/20)*0.5+(4/20)*0.5=0.49.

6) Which attribute is better, Gender, Car Type, or Shirt Size?

   The better attribute is Car Type since it has the lowest Gini Index out of all the attributes.

7) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.

It should not be used as an attribute since it is unique per customer and an unique attribute should not be used for testing.

Exercise 5:

Question: What are the two roles played by a classification model in data mining?

The two roles that a classification model plays in data mining are predictive and descriptive modeling. Predictive modeling is a tool that uses statistical techniques, machine learning, and data mining to discover facts in order to make predictions about unknown future events. Descriptive modeling is used to "Describe", or summarize raw data and try to know something about the data that is interpretable by humans.

1.2 Tan, Chapter 4

Exercise 18:
Question: Derive the dual Lagrangian for the linear SVM with nonseparable data where the objective function is

$$f(\mathbf{w}) = \frac{\|\mathbf{w}\|^2}{2} + C\left(\sum_{i=1}^{N} \xi_i\right)^2.$$

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j} - C\left(\sum_i \xi_i\right)^2.$$

1.3 Multiclass classification
Using the confusion matrix from multiclass.Rmd notebook (from Lecture 7), create a binary-class confusion matrix using the “one-vs-many” strategy for each class. Then, for each class, compute the sensitivity, specificity and precision to two decimal places. Show all work, including the binary class confusion matrices.