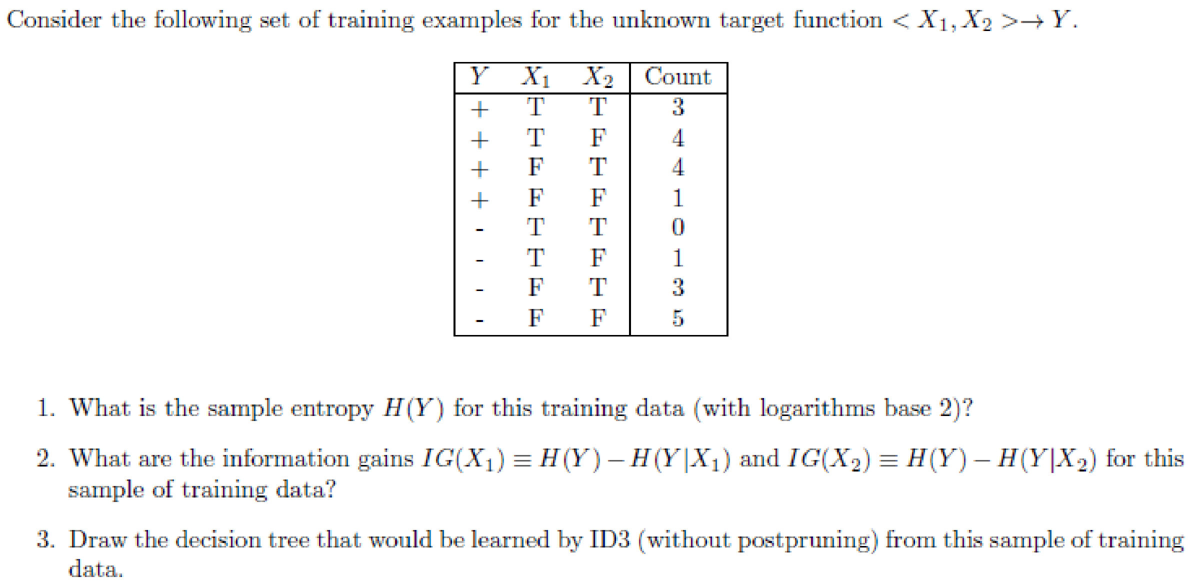
Machine Learning, Spring 2020: Homework 2

1. Linear Regression: In this problem, you will analyze potential relationships between baseball player salaries and the players' statistics. You have several features: batting average, on base, runs, hits, doubles, triples, home runs, RBI, walks, strike outs, and stolen bases. Your goal will be to see if any one of these listed features makes for a good linear relationship. Report your linear models for each feature, the overall error in the model, and the feature (if any) that you would advice is the best indicator for salary.

2. In this problem, you will work out the best feature to split on for the construction of a decision tree.



Consider the following set of training examples for the unknown target function $< X_1, X_2 > \rightarrow Y$.

Y	X_1	X_2	Count
+	T	T	3
+	T	F	4
+	F	T	4
+	F	F	1
-	T	T	0
-	T	F	1
-	F	T	3
-	F	F	5

1. What is the sample entropy $H(Y)$ for this training data (with logarithms base 2)?
2. What are the information gains $IG(X_1) \equiv H(Y) - H(Y|X_1)$ and $IG(X_2) \equiv H(Y) - H(Y|X_2)$ for this sample of training data?
3. Draw the decision tree that would be learned by ID3 (without postpruning) from this sample of training data.

3. In this problem, you will analyze the Pokemon dataset. The goal is to predict whether a Pokemon will be classified as legendary or not. For this exercise, construct a decision tree since you will be dealing with both continuous and categorical variables. You will submit your code, along with a visual display of the best decision tree for predicting whether a Pokemon should have the legendary status.