15-681 Machine Learning

Homework # 2 **due: Sept 15**

1. **Mistake-bound model.** A "2-CNF formula" is a conjunction of items each of which is a disjunction of up to 2 literals. For instance, the rule:

   `(cloudy OR warm) AND (windy) AND (weekend OR holiday)`

   is a 2-CNF formula.

   (a) How would you represent the concept "windy and at least 2 out of 3 of {sunny, warm, weekend}" as a 2-CNF?

   (b) Give an algorithm to learn 2-CNF formulas over n boolean features in the mistake-bound model. Your algorithm should run in polynomial-time per example (so the "halving algorithm" is not allowed). How many mistakes does it make at most?

2. **PAC model.** Consider the problem of learning the concept class $C = \{[0, a] : a > 0\}$. In class we saw that the strategy: "Pick $\frac{1}{\varepsilon}\ln(\frac{1}{\delta})$ examples and hypothesize the interval $[0, b]$ where b is the largest positive example seen" succeeds in the PAC model (with probability $\geq 1-\delta$ the error is $\leq \varepsilon$). Prove that after $\frac{1}{\varepsilon}\ln(\frac{2}{\delta})$ examples, with probability $\geq 1 - \delta$, *every* interval $[0, b]$ consistent with the data has error $\leq \varepsilon$.

3. **Representations.** Prove that any 1-decision list can be written as an equivalent linear threshold function. That is, for any set of examples consistent with a decision list, the positive examples are linearly separable from the negative examples. (A linear threshold function is a function like "$2x_1 + 5x_2 - 4x_3 \geq 3$.")

   For instance, the decision list: "$[x_1 \rightarrow +]$, else $[x_2 \rightarrow -]$, else $+$" can be written as: $x_1 - x_2 \geq 0$.

   As a hint, try out some examples.

4. **Decision Trees.**

   (a) Write a decision tree for the concept in 2(a) above.

   (b) Show the decision tree that would be learned by ID3 assuming it is given the four training examples for the *Enjoy-Sport?* target concept shown in Table 2-1 of Chapter 2.