Homework 4

- *Homework deadline: 10:30am on Nov 8*

1. **Hidden Markov Model [40 pts]**. In class we defined a forward looking variable $\alpha_{t+1}(i) = P(O_1, \cdots, O_{t+1} \wedge q_{t+1} = s_i)$. We also defined a backward looking variable $\beta_t(i) = P(O_{t+1}, \cdots, O_n | s_t = i)$.

   (a) [**15 pts**] Use the above two definitions, show how to get the following equation.

   $$P(q_t = s_i | O_1, \cdots, O_n) = \frac{\alpha_t(i)\beta_t(i)}{\sum_j \alpha_t(j)\beta_t(j)}$$

   .

   (b) [**25 pts**] Use the above two definitions and (a), show how to get the following equation.

   $$P(q_t = s_i, q_{t+1} = s_j | O_1, \cdots, O_n) = \frac{\alpha_t(i)P(q_{t+1} = s_j | q_t = s_i)P(o_{t+1} | s_j)\beta_{t+1}(j)}{\sum_j \alpha_t(j)\beta_t(j)}$$

2. **Decision Trees and Neural Nets [35 pts]**. In this problem we compare decision trees and neural nets on two functions, namely the *majority function* and the *parity function*. Both functions are logic functions from n inputs to 1 output. Majority function returns TRUE if more than half of its inputs are TRUE. Parity function returns TRUE if and only if an odd number of inputs are TRUE.

   (a) [**12 pts**] Can a perceptron be used to solve

   - the majority function?
   - the parity function?

   For each of these two functions, if you say yes present the perceptron that solves it, and if you say no explain why.

   (b) [**12 pts**] Can a decision tree be used to solve

   - the majority function?
   - the parity function?

   For each of these two functions, if you say yes explain how to construct a decision tree that solves it, and if you say no explain why.

   (c) [**11 pts**] For each of the functions in (a) for which you said no, can it be solved using a two layered neural network (one input layer and one hidden layer)? If so, present the neural nets that solves the problem(s); if no, how many layers would you need to solve it (assume there are n inputs)?

3. **Decision Trees [25 pts]**. Denote by n and p the set of negative and positive samples at a specific internal node in a decision tree. Show that if an attribute k divides the set of samples into p_0 and n_0 (for $k = 0$), and p_1 and n_1 (for $k = 1$), then the information gain from using attribute k at this node is greater or equal to 0. Hint: you may want to use the following version of Jensen's inequality:

$$\sum_{i=1}^{v} \alpha_i \log x_i \leq \log(\sum_{i=1}^{v} \alpha_i x_i)$$

where $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$.