1 Homework

1. 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:

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

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

2. Given

$$Gain = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{(H_L + H_R) + \lambda} - \gamma, \tag{2}$$

show that,

- if $\lambda = \gamma = 0$, then $Gain$ must be greater or equal to 0.
- if $\lambda \neq 0$, then $Gain$ may be less than 0.
- if $\gamma \neq 0$, then $Gain$ may be less than 0.