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 $\sum_i \alpha_i = 1$.

**Solution 1:** : by *Ian Fette*

We offer the following proof that information gain is always nonnegative. (The exact number of positive and negative examples, $n_0, n_1, p_0, p_1$ are actually not important for this proof.)

Assume that the classes we are trying to distinguish between are represented by $X$ and that the attribute we are splitting on is $K$. Then let us denote $P(X, K)$ to be the joint PDF of $X$ and $K$. We can obtain the marginal density $P(X)$ by summing over values of K, and vice versa. (i.e. $P(X) = \sum_K P(X, K)$ and $P(K) = \sum_X P(X, K)$.)

Our proof is as follows:

$$IG(X, K) = H(X) - H(X|K) \tag{1}$$

$$IG(X, K) = \sum_X -P(X) \log_2 P(X) - \sum_K P(K) \sum_X (-P(X|K) \log_2 P(X|K)) \tag{2}$$

$$-IG(X, K) = \sum_X P(X) \log_2 P(X) - \sum_K P(K) \sum_X (P(X|K) \log_2 P(X|K)) \tag{3}$$

$$-IG(X, K) = \sum_X \sum_K P(X, K) \log_2 P(X) - \sum_K P(K) \sum_X (P(X|K) \log_2 P(X|K)) \tag{4}$$

$$-IG(X, K) = \sum_X \sum_K P(X, K) \log_2 P(X) - \sum_K \sum_X (P(K)P(X|K) \log_2 P(X|K)) \tag{5}$$

$$-IG(X, K) = \sum_X \sum_K P(X, K) \log_2 P(X) - \sum_K \sum_X (P(X, K) \log_2 P(X|K)) \tag{6}$$

$$-IG(X, K) = \sum_X \sum_K P(X, K)(\log_2 P(X) - \log_2 P(X|K)) \tag{7}$$

$$-IG(X,K) = \sum_X \sum_K P(X,K) \left( \log_2 \left( \frac{P(X)}{P(X|K)} \right) \right) \tag{8}$$

$$-IG(X,K) = \sum_X \sum_K P(X|K)P(K) \left( \log_2 \left( \frac{P(X)}{P(X|K)} \right) \right) \tag{9}$$

$$-IG(X,K) = \sum_K P(K) \sum_X P(X|K) \left( \log_2 \left( \frac{P(X)}{P(X|K)} \right) \right) \tag{10}$$

$$-IG(X,K) \leq \sum_K P(K) \left( \log_2 \left( \sum_X \frac{P(X|K)P(X)}{P(X|K)} \right) \right) \tag{11}$$

$$-IG(X,K) \leq \log_2 \left( \sum_K \sum_X \frac{P(K)P(X|K)P(X)}{P(X|K)} \right) \tag{12}$$

$$-IG(X,K) \leq \log_2 \left( \sum_K \sum_X P(K)P(X) \right) \tag{13}$$

$$-IG(X,K) \leq \log_2 \left( \sum_K P(K) \sum_X P(X) \right) \tag{14}$$

$$-IG(X,K) \leq \log_2 \left( \sum_K P(K) \right) \tag{15}$$

$$-IG(X,K) \leq \log_2(1) \tag{16}$$

$$-IG(X,K) \leq 0 \tag{17}$$

$$IG(X,K) \geq 0 \tag{18}$$

In this proof, lines 11 and 12 are both applications of Jensen's inequality. On line 11, $\sum_X P(X|K) = 1$, and by definition each probability is nonnegative. The same argument applies for the application of Jensen's inequality on line 12.

**Solution 2:**

**Lemma**: $f(x) = -x\log_2 x - (1-x)\log_2(1-x)$ is a concave function where $x \in (0,1)$.

**Proof**:

$$f'(x) = -\log_2 x + \log_2(1-x)$$
$$f''(x) = -\frac{1}{ln2} \cdot \frac{1}{x(1-x)}$$

Since $x \in (0,1)$, we have $f''(x) < 0$.

Known from concave function's property that if $f$ is twice continuously differentiable function on R. Then $f$ is concave if and only if $f'' \leq 0$. So we have that $f(x)$ is concave. Q.E.D.

Information gain from using attribute $k$ at this node is:

$IG = I(\frac{p}{p+n}, \frac{n}{p+n}) - \sum_{i=0}^{1} \frac{p_i+n_i}{p+n} I(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i})$

We want to show that $IG \geq 0$.

$$\sum_{i=0}^{1} \frac{p_i+n_i}{p+n} I(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}) \tag{19}$$

$$\begin{aligned}
&= \frac{p_0+n_0}{p+n} I(\frac{p_0}{p_0+n_0}, \frac{n_0}{p_0+n_0}) + \frac{p_1+n_1}{p+n} I(\frac{p_1}{p_1+n_1}, \frac{n_1}{p_1+n_1}) && (20)\\
&= \frac{p_0+n_0}{p+n} f(\frac{p_0}{p_0+n_0}) + \frac{p_1+n_1}{p+n} f(\frac{p_1}{p_1+n_1}) && (21)\\
&\leq f(\frac{p_0+n_0}{p+n} \cdot \frac{p_0}{p_0+n_0} + \frac{p_1+n_1}{p+n} \cdot \frac{p_1}{p_1+n_1}) && (22)\\
&= f(\frac{p_0+p_1}{p+n}) && (23)\\
&= f(\frac{p}{p+n}) && (24)\\
&= I(\frac{p}{p+n}, \frac{n}{p+n}) && (25)
\end{aligned}$$

Line (22) uses the following form of Jensen's Inequality:

$$\sum_x p(x)f(x) \leq f(\sum_x p(x)x)$$

where $\sum_x p(x) = 1, p(x) \geq 0, f(x)$ is concave.

So that $IG = I(\frac{p}{p+n}, \frac{n}{p+n}) - \sum_{i=0}^{1} \frac{p_i+n_i}{p+n} I(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}) \geq 0$. Finally we have shown that the information gain is non-negative.