HW2

211482559 207253899

Gini Impurity - A .1

על פי ההגדרה:

$$\varphi_{gini}(p) = 1 - \sum_{i=0}^{k} p^2 \qquad \sum_{i=0}^{k} p = 1$$

נרצה להוכיח:

$$\varphi_{gini}(p) \leq 1 - \frac{1}{k}$$

נשים לב ע"פ אי שיוון קושי שוורץ:

$$(\sum_{i=0}^{k} 1 * p)^2 \leq \sum_{i=0}^{k} p^2 \cdot \sum_{i=0}^{k} 1^2 = \sum_{i=0}^{k} p^2 \cdot k$$

נפשט את הביטוי השמאלי:

$$(\sum_{i=0}^{k} 1 * p)^2 = (\sum_{i=0}^{k} p)^2 = 1^2$$

סה"כ:

$$1 = (\sum_{i=0}^{k} 1 * p)^2 \leq \sum_{i=0}^{k} p^2 \cdot k$$

$$\frac{1}{k} \leq \sum_{i=0}^{k} p^2$$

$$\frac{1}{k} - 1 \leq \sum_{i=0}^{k} p^2 - 1$$

$$1 - \frac{1}{k} \geq 1 - \sum_{i=0}^{k} p^2$$

Gini Impurity - B .2

לפי הנתונים:

$$Pr(Y_i = j) = p_i$$

נרצה להוכיח:

$$\varphi_{gini}(p) = Pr(Y_1 \neq Y_2) = 1 - Pr(Y_1 = Y_2)$$

נשים לב:

$$Pr(Y_1 = Y_2) = \sum_{i=0}^{k} (Pr(Y_1 = j) \wedge Pr(Y_2 = j)) = \sum_{i=0}^{k} (p_i \cdot p_i) = \sum_{i=0}^{k} p_i^2$$

לכן:

$$1 - Pr(Y_1 = Y_2) = 1 - \sum_{i=0}^{k} p_i^2$$

סה"כ:

$$Pr(Y_1 \neq Y_2) = 1 - Pr(Y_1 = Y_2) = 1 - \sum_{i=0}^{k} p_i^2 = \varphi_{gini}(p)$$

Information Gain - A .3

נרצה להוכיח באינדוקציה:

$$h(\sum_{i=0}^{k} \lambda_i x_i) \geq \sum_{i=0}^{k} \lambda_i h(x_i)$$

בסיס האינדוקציה (k=2): נכון על פי הנתונים.

שלב האינדוקציה:

נניח כי:

$$h(\sum_{i=0}^{k} \lambda_i x_i) \geq \sum_{i=0}^{k} \lambda_i h(x_i)$$

ונראה שמתקיים:

$$h(\sum_{i=0}^{k+1} \lambda_i x_i) \geq \sum_{i=0}^{k+1} \lambda_i h(x_i)$$

$$h(\sum_{i=0}^{k+1} \lambda_i x_i) = h(\lambda_{i+1} x_{i+1} + \sum_{i=0}^{k} \lambda_i x_i) = h(\lambda_{i+1} x_{i+1} + (1 - \lambda_{i+1}) \sum_{i=0}^{k} \frac{\lambda_i}{1-\lambda_{i+1}} x_i) =$$

$$h(\lambda_{i+1} x_{i+1} + (1 - \lambda_{i+1}) \sum_{i=0}^{k} \frac{\lambda_i}{1-\lambda_{i+1}} x_i) \geq \lambda_{i+1} h(x_{i+1}) + (1 - \lambda_{i+1}) \sum_{i=0}^{k} \frac{\lambda_i}{1-\lambda_{i+1}} h(x_i) =$$

$$\lambda_{i+1} h(x_{i+1}) + (1 - \lambda_{i+1}) \sum_{i=0}^{k} \frac{\lambda_i}{1-\lambda_{i+1}} h(x_i) = \lambda_{i+1} h(x_{i+1}) + \sum_{i=0}^{k} \lambda_i h(x_i) = \sum_{i=0}^{k+1} \lambda_i h(x_i)$$

Information Gain - B .4

נרצה להוכיח:

$$IG(S,A) = H(S) - \sum_{v \in V(A)} \frac{|s_v|}{|s|} H(s_v) \geq 0$$

או בשקילות:

$$H(S) \geq \sum_{v \in V(A)} \frac{|s_v|}{|s|} H(s_v)$$

נפשט כל צד:

$$\sum_{v \in V(A)} \frac{|s_v|}{|s|} H(s_v) = - \sum_{v} p(v) \sum_{x} p(x|v) log(p(x|v))$$

$$H(S) = - \sum_{x} p(x) log(p(x)) = - \sum_{x} \sum_{v} p(v) p(x|v) log(\sum_{v} p(v) p(x|v))$$

לפי אי השוויון שהוכחנו קודם:

$$H(S) = - \sum_{x} (\sum_{v} p(v) p(x|v)) log(\sum_{v} p(v) p(x|v)) \geq$$

$$- \sum_{v} p(v) \sum_{x} p(x|v) log(p(x|v)) = \sum_{v \in V(A)} \frac{|s_v|}{|s|} H(s_v)$$