Exam Formula Sheet

Given a discrete random variable *Y* taking *n* classes, Gini Impurity is defined as:

$Gini(Y) = \sum_{i=1}^{n} p_i(1 - p_i) = 1 - \sum_{j=1}^{n} p_i^2$ where $p_i = P(Y = y_i)$.

Using feature *A* to split into *m* partitions, the revised Gini impurity:

$$Gini_A(Y) = \sum_{j=1}^{m} \frac{|Y_j|}{|Y|} \times Gini(Y_j)$$

Change in Gini impurity (reduction in impurity) is given by:

$\Delta Gini(A) = Gini(Y) - Gini_A(Y)$

Given a discrete random variable (Y) taking *n* classes, Expected Entropy is:

$Info(Y) = -\sum_{i=1}^{n} p_i \, log_2(p_i)$ where $p_i = P(Y = y_i)$.

Using a feature *A* to split into *m* partitions, the expected entropy is:

$$Info_A(Y) = \sum_{j=1}^{m} \frac{|Y_j|}{|Y|} \times Info(Y_j)$$

Information gain by splitting according to A is given by:

$Gain(A) = Info(Y) - Info_A(Y)$