AML solutions sheet 7.2

Jannik Gut

December 2020

Problem 2(Cluster evaluation)

1. purity

The maximal value of 1 can also be achieved by two full sets and i-1 empty sets.

$$U_1 = X = V_1 \quad U_{\neg 1} = V_{\neg 1} = \emptyset$$

This obviously adheres the restrictions and the maximal value is also achieved as:

$$\frac{1}{|X|}|U_1 \cap V_1| = \frac{|X|}{|X|} = 1$$

2. mutual information

U/V means either U or V

$$\forall p, p \geq 0, p_{UV} \leq p_{U/V} \quad \sum_j p_{UV}(i,j) = p_U(i)$$

$I(U,V) \geq 0$**:**
Use Jensen's from the solution, I can not do better.
$I(U,V) \leq H(U/V)$**:**

$$I(U,V) = \sum_i \sum_j p_{UV}(i,j) log_2 \frac{p_{UV}(i,j)}{p_U(i)p_V(j)} = \sum_i \sum_j p_{UV}(i,j) * (-1) * log_2 \frac{p_U(i)p_V(j)}{p_{UV}(i,j)}$$

$$I(U,V) = -\sum_i \sum_j p_{UV}(i,j) log_2 \frac{p_U(i)}{p_{UV}(i,j)} - \sum_i \sum_j p_{UV}(i,j) log_2 p_V(j)$$

$$I(U,V) = -\sum_i \sum_j p_{UV}(i,j) log_2 \frac{p_U(i)}{p_{UV}(i,j)} - \sum_j p_V(j) log_2 p_V(j)$$

$$I(U,V) = -\sum_i \sum_j p_{UV}(i,j) log_2 \frac{p_U(i)}{p_{UV}(i,j)} + H(V)$$

If the first part is non-negative then the inequality will hold:
The first part is split first:

$$p_{UV}(i,j) \geq 0$$

as stated at the beginning, p is a probability
The only thing left is:

$$log_2 \frac{p_U(i)}{p_{UV}(i,j)} = \alpha \geq 0$$

As written above we now that

$$p_{UV}(i,j) \leq p_U(i) \Rightarrow \frac{p_U(i)}{p_{UV}(i,j)} \geq 1 \Rightarrow log_2 \frac{p_U(i)}{p_{UV}(i,j)} = \beta \geq 0$$

Now (we can exchange V for U):

$$I(U,V) = H(V) - \sum \sum \alpha\beta$$

3.

Unlike mutual information, purity does not regulate on the size of the cluster.