

UFO: Supplementary material

1. Consistency measures and their bounds

In this section we provide the proofs of the claims, stated in the paper, concerning the lower bounds for consistencies measured between pairs of features. The proofs are given following the formulations of corresponding claims for three out of the four consistency measures, considered in the paper, namely: Jaccard Index, Mutual Information and Suspicious coincidence. The bound for the case of L2 obviously follows from the triangular inequality.

1.1. Jaccard Index (JI)

Claim 2.1

$$JI[F_i; F_j] \geq \frac{JI[F_i; C] + JI[F_j; C] - 1}{\frac{1}{JI[F_i; C]} + \frac{1}{JI[F_j; C]} - 1} \quad (1)$$

Proof.

Here for convenience we treat F_i , F_j and C (binary row vectors of length N in the paper) as sets of indices of ones in them. Bar above a set means its complement.

Consider:

$$\begin{aligned} |F_i \cap C| &= JI[F_i, C] \cdot |F_i \cup C| \geq JI[F_i, C] \cdot |C| \\ |F_j \cap C| &= JI[F_j, C] \cdot |F_j \cup C| \geq JI[F_j, C] \cdot |C| \end{aligned} \quad (2)$$

Hence:

$$|\bar{F}_j \cap C| = |C \setminus (F_j \cap C)| = |C| - |F_j \cap C| \leq |C| - JI[F_j, C] \cdot |C| \quad (3)$$

And thus:

$$\begin{aligned} |F_i \cap F_j| &\geq |(F_i \cap C) \cap (F_j \cap C)| = |(F_i \cap C) \setminus (\bar{F}_j \cap C)| \geq \\ &\geq |F_i \cap C| - |\bar{F}_j \cap C| \geq JI[F_i, C] \cdot |C| - (|C| - JI[F_j, C] \cdot |C|) = \\ &= (JI[F_i, C] + JI[F_j, C] - 1) \cdot |C| \end{aligned} \quad (4)$$

Moreover:

$$|C| \geq |F_i \cap C| = JI[F_i, C] \cdot |F_i \cup C| \geq JI[F_i, C] \cdot |F_i| \quad (5)$$

which means:

$$|F_i| \leq \frac{|C|}{JI[F_i, C]} \quad (6)$$

Thus:

$$|F_i \setminus C| \leq |F_i| - |C| \leq \frac{|C|}{JI[F_i, C]} - |C| \quad (7)$$

Similarly:

$$|F_j \setminus C| \leq |F_j| - |C| \leq \frac{|C|}{JI[F_j, C]} - |C| \quad (8)$$

Hence:

$$\begin{aligned} |F_i \cup F_j| &\leq |C \cup (F_i \setminus C) \cup (F_j \setminus C)| \leq |C| + |F_i \setminus C| + |F_j \setminus C| \leq \\ &\leq |C| + \frac{|C|}{JI[F_i, C]} - |C| + \frac{|C|}{JI[F_j, C]} - |C| = \\ &= \left(\frac{1}{JI[F_i, C]} + \frac{1}{JI[F_j, C]} - 1 \right) \cdot |C| \end{aligned} \quad (9)$$

Combining all the above expressions we finally get:

$$\begin{aligned} JI[F_i; F_j] &= \frac{|F_i \cap F_j|}{|F_i \cup F_j|} \geq \\ &\geq \frac{(JI[F_i, C] + JI[F_j, C] - 1) \cdot |C|}{\left(\frac{1}{JI[F_i, C]} + \frac{1}{JI[F_j, C]} - 1 \right) \cdot |C|} = \\ &= \frac{JI[F_i, C] + JI[F_j, C] - 1}{\frac{1}{JI[F_i, C]} + \frac{1}{JI[F_j, C]} - 1} \end{aligned} \quad (10)$$

■

1.2. Mutual Information (MI)

Claim 2.2

Assuming that F_i and F_j are conditionally independent given class C , then:

$$MI[F_i; F_j] \geq MI[F_i; C] + MI[F_j; C] - H(C) \quad (11)$$

Proof.

Here for convenience we treat F_i , F_j and C (binary row vectors of length N in the paper) as binary random variables with joint distribution given by empirical distribution computed on the vectors (this is the ML approximation).

Consider:

$$\begin{aligned} H(F_i, F_j) &\leq H(F_i, F_j, C) = - \sum_{F_i, F_j, C} P(F_i, F_j, C) \log(P(F_i, F_j, C)) = \\ &= - \sum_{F_i, F_j, C} P(F_i, F_j, C) \log(P(F_i, F_j|C) \cdot P(C)) = \\ &= - \sum_{F_i, F_j, C} P(F_i, F_j, C) \log(P(F_i|C)P(F_j|C) \cdot P(C)) = \\ &= - \sum_{F_i, F_j, C} P(F_i, F_j, C) \log(P(F_i|C)) - \sum_{F_i, F_j, C} P(F_i, F_j, C) \log(P(F_j|C)) - \sum_{F_i, F_j, C} P(F_i, F_j, C) \log(P(C)) = \\ &= - \sum_{F_i, C} P(F_i, C) \log(P(F_i|C)) - \sum_{F_j, C} P(F_j, C) \log(P(F_j|C)) - \sum_C P(C) \log(P(C)) = \\ &= H(F_i|C) + H(F_j|C) - H(C) \end{aligned} \quad (12)$$

Thus:

$$\begin{aligned}
MI(F_i, F_j) &= H(F_i) + H(F_j) - H(F_i, F_j) \geq H(F_i) + H(F_j) - [H(F_i|C) + H(F_j|C) - H(C)] = \\
&= [H(F_i) - H(F_i|C)] + [H(F_j) - H(F_j|C)] - H(C) = \\
&= MI(F_i, C) + MI(F_j, C) - H(C)
\end{aligned} \tag{13}$$

■

1.3. Suspicious Coincidence (SC)

Claim 2.3

Assume the events $F_i = 1$ and $F_j = 1$ are conditionally independent given the event $C = 1$, then:

$$SC[F_i; F_j] \geq SC[F_i; C] \cdot SC[F_j; C] \cdot P(C = 1) \tag{14}$$

Proof.

Here for convenience we treat F_i , F_j and C (binary row vectors of length N in the paper) as binary random variables with joint distribution given by empirical distribution computed on the vectors (this is the ML approximation).

$$\begin{aligned}
SC[F_i; F_j] &= \frac{P(F_i = 1, F_j = 1)}{P(F_i = 1)P(F_j = 1)} = \\
&= \frac{P(F_i = 1, F_j = 1|C = 1)P(C = 1) + P(F_i = 1, F_j = 1|C = 0)P(C = 0)}{P(F_i = 1)P(F_j = 1)} \geq \\
&\geq \frac{P(F_i = 1, F_j = 1|C = 1)P(C = 1)}{P(F_i = 1)P(F_j = 1)} = \\
&= \frac{P(F_i = 1, C = 1)P(F_j = 1, C = 1)P(C = 1)}{P(F_i = 1)P(C = 1)P(F_j = 1)P(C = 1)} = \\
&= SC[F_i; C] \cdot SC[F_j; C] \cdot P(C = 1)
\end{aligned} \tag{15}$$

■