ITDM - LEZIONE 05

DEL 15/10/2019

In this lecture we will define other measures of information.

KULLBACK-LEIBLER DIVERGENCE

Given two p.m.f. $p(x), q(y)$ associated to two r.v. having the same range, we can define the Kullback-Leibler divergence as follows

$$D_{KL}(p \| q) := \sum_{i=1}^{N_x} p(x_i) \cdot \log \frac{p(x_i)}{q(x_i)}$$

Notice that $D_{KL}(p \| q)$ is a measure of DIFFERENCE between two pmf. It is NOT a measure of DISTANCE between two pmf because its not symmetric.

$$D_{KL}(p \| q) \neq D_{KL}(q \| p)$$

VARIATION OF INFORMATION

Its defined as
$$\begin{aligned} V(X,Y) &:= H(X|Y) + H(Y|X) \\ &= H(X,Y) - I(X,Y) \\ &= H(X) + H(Y) - 2 \cdot I(X,Y) \\ &= 2H(X,Y) - H(X) - H(Y) \end{aligned}$$

This measure is SYMMETRIC $V(X,Y) = V(Y,X)$. Also, we note that $V(X,X) = 0$ because $H(X|X) = 0$. Finally, $V(X,Y)$ satisfies the TRIANGULAR INEQUALITY. We can thus say that $V(X,Y)$ is a DISTANCE between r.m.f.s

$V(X,Y)$ is a measure of DISSIMILARITY between rmfs. If we want a measure of SIMILARITY we can take the inverse $\frac{1}{V(X,Y)}$.

In DATA MINING similarity is also known as ASSOCIATION.

NORMALIZED MEASURES OF INFORMATION ~ (18:50)

The measures introduced so far depend on the particular values of $H(X)$ and $H(Y)$. To have measures which are more global, in the sense that they do not depend too much on the particular values of $H(X)$ and $H(Y)$ we introduce here some NORMALIZED MEASURES.

• NORMALIZED JOINT ENTROPY

Defined as,

$$\mu_{JE}(X,Y) := 1 - \frac{I(X,Y)}{H(X)+H(Y)} \in [\tfrac{1}{2}, 1]$$

• NORMALIZED CONDITIONAL ENTROPY

Defined as,

$$\mu_{CE}(X, Y) := \frac{I(X, Y)}{H(X, Y)} \in [0, 1]$$

• NORMALIZED MEASURE OF INFORMATION (TYPE I)

Defined as $\eta_{MI} := \dfrac{I(X,Y)}{H(X,Y)} \in [0, 1]$

If $X = Y$ then $\eta_{MI}(X,Y) = 1$.

If $X \perp Y$ then $\eta_{MI}(X, Y) = 0$.

• NORMALIZED MEASURE OF INFORMATION (TYPE II)

Defined as, $\eta_{MI2}(X, Y) := \dfrac{H(X) + H(Y)}{H(X, Y)}$

Notice that $\eta_{MI2}(X,Y) = 1 + \eta_{MI}(X, Y)$, which means that $\eta_{MI2}(X,Y) \in [1,2]$.

• NORMALIZED MEASURE OF INFORMATION (TYPE III)

Defined as,

$$\eta_{MI_3} := \frac{I(X, Y)}{\sqrt{H(X) \cdot H(Y)}} \in [0, 1]$$

It measures the association between X and Y. A value of 0 indicates that $X \perp Y$, while a value of 1 indicates that $Y = f(X)$.

OSS: In general normalized measures are useful to compare different associations, since we scale by the quantity of information of the specific r.v.

• Entropy Correlation Coefficient

Defined as,

$$\eta_{CC} := \sqrt{\frac{2 \cdot I(X,Y)}{H(X) + H(Y)}} \in [0,1]$$

• Symmetric Uncertainty

Defined as,

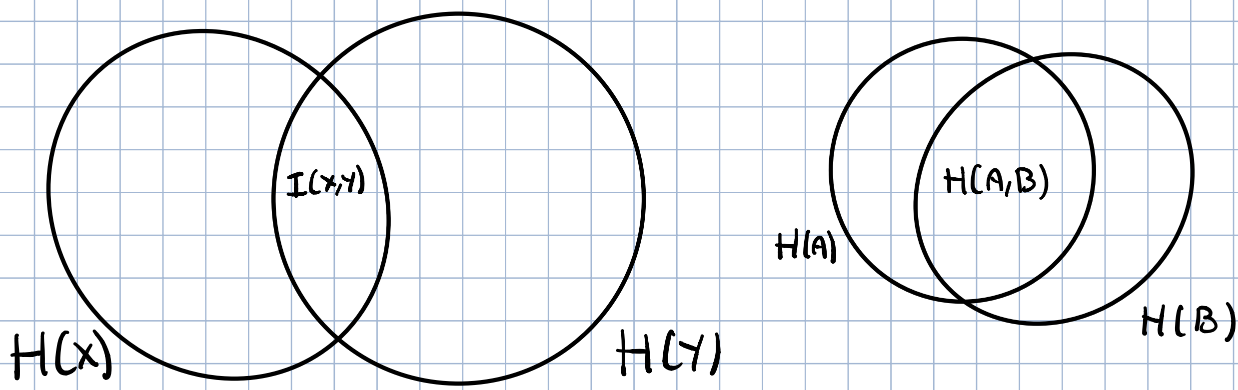
$$\eta_{SU}(X,Y) := \frac{2 \cdot I(X,Y)}{H(X) + H(Y)} \in [0,1]$$

• Variation of Information

Defined as,

$$\eta_{VI}(X,Y) := \frac{V(X,Y)}{H(X,Y)}$$

OSS: Consider the following entropy models:



If we increase the entropy of X and Y we also increase their mutual information $I(X,Y)$. Normalized measures allow us to not be worried about this type of increase of $I(X,Y)$ in the final result.

If we don't use normalized measure we could infer that X and Y are more associated than A and B simply because $H(X)$ and $H(Y)$ are higher than $H(A)$ and $H(B)$ and therefore $I(X,Y) > I(A,B)$.