Measures of Information

- Hartley defined the first information measure:
  - H = n log s
  - n is the length of the message and s is the number of possible values for each symbol in the message
  - Assumes all symbols equally likely to occur
- Shannon proposed variant (Shannon's Entropy)

$$H = \sum_i p_i \cdot \log \frac{1}{p_i}$$

  - weighs the information based on the probability that an outcome will occur
  - second term shows the amount of information an event provides is inversely proportional to its prob of occurring

Three Interpretations of Entropy

- The amount of information an event provides
  - An infrequently occurring event provides more information than a frequently occurring event
- The uncertainty in the outcome of an event
  - Systems with one very common event have less entropy than systems with many equally probable events
- The dispersion in the probability distribution
  - An image of a single amplitude has a less disperse histogram than an image of many greyscales
    - the lower dispersion implies lower entropy

Definitions of Mutual Information

- Three commonly used definitions:
  - 1) $I(A,B) = H(B) - H(B|A) = H(A) - H(A|B)$
    - Mutual information is the amount that the uncertainty in B (or A) is reduced when A (or B) is known.
  - 2) $I(A,B) = H(A) + H(B) - H(A,B)$
    - Maximizing the mutual info is equivalent to minimizing the joint entropy (last term)
    - Advantage in using mutual info over joint entropy is it includes the individual input's entropy
    - Works better than simply joint entropy in regions of image background (low contrast) where there will be low joint entropy but this is offset by low individual entropies as well so the overall mutual information will be low

Definitions of Mutual Information II

$$I(A,B) = \sum_{a,b} p(a,b) \cdot \log\left(\frac{p(a,b)}{p(a)\,p(b)} \right)$$

- This definition is related to the Kullback-Leibler distance between two distributions
- Measures the dependence of the two distributions
- In image registration I(A,B) will be maximized when the images are aligned
- In feature selection choose the features that minimize I(A,B) to ensure they are not related.

Additional Definitions of Mutual Information

- Two definitions exist for normalizing Mutual information:
  - Normalized Mutual Information:

$$NMI(A,B) = \frac{H(A) + H(B)}{H(A,B)}$$

  - Entropy Correlation Coefficient:

$$ECC(A,B) = 2 - \frac{2}{NMI(A,B)}$$

Derivation of M. I. Definitions

$$H(A,B) = \sum_{a,b} p(a,b) \cdot \log(\, p(a,b)), \text{where } \; p(a,b) = p(a \mid b) \cdot p(b)$$

$$H(A,B) = \sum_{a,b} [p(a \mid b) \cdot p(b)] \cdot \log[\, p(a \mid b) \cdot p(b)]$$

$$H(A,B) = \sum_{a,b} [p(a \mid b) \cdot p(b)] \cdot \{\log[\, p(a \mid b)] + \log[\, p(b)]\}$$

$$H(A,B) = \sum_{a,b} p(a \mid b) \cdot \log[\, p(a \mid b)] \cdot p(b) + \sum_{a,b} p(b) \cdot \log(\, p(b)) \cdot p(a \mid b)$$

$$H(A,B) = \sum_{a} p(a \mid b) \cdot \log[\, p(a \mid b)] \cdot \sum_{b} p(b) + \sum_{b} \sum_{a} p(a \mid b) \cdot p(b) \cdot \log(\, p(b$$

$$H(A,B) = \sum_{a} p(a \mid b) \cdot \log[\, p(a \mid b)] + \sum_{b} p(b) \cdot \log(\, p(b))$$

$$H(A,B) = H(A \mid B) + H(B)$$

$$\text{therefore } \; I(A,B) = H(A) - H(B \mid A) = H(A) + H(B) - H(A,B)$$

Properties of Mutual Information

- MI is symmetric: I(A,B) = I(B,A)
- I(A,A) = H(A)
- I(A,B) <= H(A), I(A,B) <= H(B)
  - info each image contains about the other cannot be greater than the info they themselves contain
- I(A,B) >= 0
  - Cannot increase uncertainty in A by knowing B
- If A, B are independent then I(A,B) = 0
- If A, B are Gaussian then:

$$I(A,B) = -\tfrac{1}{2}\log(1-\rho^2)$$

Mutual Information based Feature Selection

- Tested using 2-class Occupant sensing problem
  - Classes are RFIS and everything else (children, adults, etc).
  - Use edge map of imagery and compute features
    - Legendre Moments to order 36
    - Generates 703 features, we select best 51 features.
- Tested 3 filter-based methods:
  - Mann-Whitney statistic
  - Kullback-Leibler statistic
  - Mutual Information criterion
    - Tested both single M.I., and Joint M.I. (JMI)

Mutual Information based Feature Selection Method

- M.I. tests a feature's ability to separate two classes.
  - Based on definition 3) for M.I.

$$I(A,B) = \sum_{a}\sum_{b} p(a,b)\cdot \log\left(\frac{p(a,b)}{p(a)p(b)}\right)$$

  - Here A is the feature vector and B is the classification
    - Note that A is continuous but B is discrete
  - By maximizing the M.I. We maximize the separability of the feature
    - Note this method only tests each feature individually