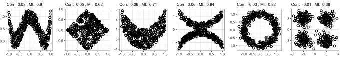
Introduction to Machine Learning

Information Theory
Joint Entropy and Mutual Information II



Learning goals

- Know mutual information as the amount of information of an RV obtained by another
- Know properties of MI

MUTUAL INFORMATION - COROLLARIES

Non-negativity of mutual information: For any two random variables, X, Y, $I(X;Y) \geq 0$, with equality if and only if X and Y are independent.

Proof: $I(X;Y) = D_{KL}(p(x,y) \| p(x)p(y)) \geq 0$, with equality if and only if $p(x,y) = p(x)p(y)$ (i.e., X and Y are independent).

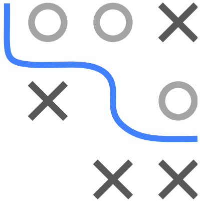
Conditioning reduces entropy (information can't hurt):

$$H(X|Y) \leq H(X),$$

with equality if and only if X and Y are independent.

Proof: $0 \leq I(X;Y) = H(X) - H(X|Y)$
Intuitively, the theorem says that knowing another random variable Y can only reduce the uncertainty in X. Note that this is true only on average.

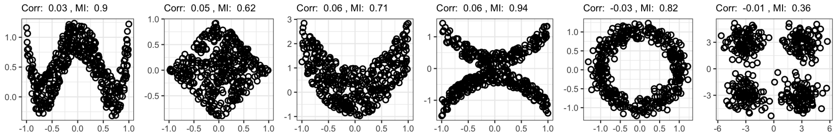
MUTUAL INFORMATION PROPERTIES



- MI is a measure of the amount of "dependence" between variables. It is zero if and only if the variables are independent.
- OTOH, if one RV is a deterministic function of the other, MI is maximal, i.e. entropy of the first RV.
- Unlike (Pearson) correlation, MI is not limited to real-valued RVs.
- Can use MI as a **feature filter**, sometimes called information gain.
- Can also be used in CART to select feature for split.
  Splitting on MI/IG = risk reduction with log-loss.
- MI invariant under injective and continuously differentiable reparametrizations.

MUTUAL INFORMATION VS. CORRELATION

- If two RVs are independent, their correlation is 0.
- But: two dependent RVs can have correlation 0 because correlation only measures linear dependence.



- Above: Many examples with strong dependence, nearly 0 correlation and much larger MI.
- MI can be seen as more general measure of dependence than correlation.

MUTUAL INFORMATION - EXAMPLE

Let X, Y be two correlated Gaussian random variables.
$(X, Y) \sim \mathcal{N}(0, K)$ with correlation ρ and covariance matrix K:

$$K = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

Then $h(X) = h(Y) = \frac{1}{2} \log\left((2\pi e)\sigma^2\right)$, and
$h(X, Y) = \frac{1}{2} \log\left((2\pi e)^2 |K|\right) = \frac{1}{2} \log\left((2\pi e)^2 \sigma^4 (1 - \rho^2)\right)$, and thus

$$I(X; Y) = h(X) + h(Y) - h(X, Y) = -\frac{1}{2} \log(1 - \rho^2).$$

For $\rho = 0$, X and Y are independent and $I(X; Y) = 0$.
For $\rho = \pm 1$, X and Y are perfectly correlated and $I(X; Y) \to \infty$.