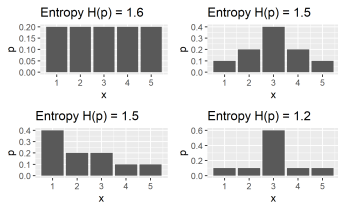


# Introduction to Machine Learning

## Information Theory

## Joint Entropy and Mutual Information I



### Learning goals

- Know the joint entropy
- Know conditional entropy as remaining uncertainty
- Know mutual information as the amount of information of an RV obtained by another

# JOINT ENTROPY

- Recap: The **joint entropy** of two discrete RVs  $X$  and  $Y$  with joint pmf  $p(x, y)$  is:

$$H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log(p(x, y)),$$

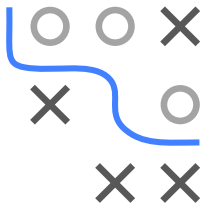
which can also be expressed as

$$H(X, Y) = -\mathbb{E} [\log(p(X, Y))].$$

- For continuous RVs  $X$  and  $Y$  with joint density  $p(x, y)$ , the differential joint entropy is:

$$h(X, Y) = - \int_{\mathcal{X} \times \mathcal{Y}} p(x, y) \log p(x, y) dx dy$$

For the rest of the section we will stick to the discrete case. Pretty much everything we show and discuss works in a completely analogous manner for the continuous case - if you change sums to integrals.



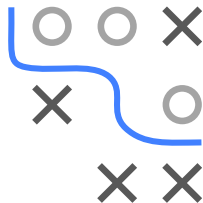
# CONDITIONAL ENTROPY

- The **conditional entropy**  $H(Y|X)$  quantifies the uncertainty of  $Y$  that remains if the outcome of  $X$  is given.
- $H(Y|X)$  is defined as the expected value of the entropies of the conditional distributions, averaged over the conditioning RV.
- If  $(X, Y) \sim p(x, y)$ , the conditional entropy  $H(Y|X)$  is defined as

$$\begin{aligned} H(Y|X) &= \mathbb{E}_X[H(Y|X = x)] = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= -\mathbb{E}[\log p(Y|X)]. \end{aligned}$$

- For the continuous case with density  $f$  we have

$$h(Y|X) = - \int f(x, y) \log f(x|y) dx dy.$$



# CHAIN RULE FOR ENTROPY

The **chain rule for entropy** is analogous to the chain rule for probability and derives directly from it.

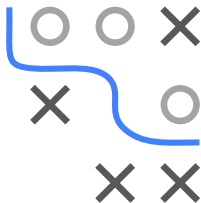
$$H(X, Y) = H(X) + H(Y|X)$$

**Proof:**

$$\begin{aligned} H(X, Y) &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) p(y|x) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= - \sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= H(X) + H(Y|X) \end{aligned}$$

n-variable version:

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1).$$



# JOINT AND CONDITIONAL ENTROPY

The following relations hold:

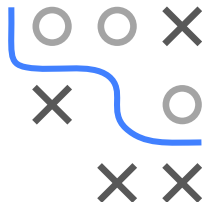
$$H(X, X) = H(X)$$

$$H(X|X) = 0$$

$$H((X, Y)|Z) = H(X|Z) + H(Y|(X, Z))$$

Which can all be trivially derived from the previous considerations.

Furthermore, if  $H(X|Y) = 0$  and  $X, Y$  are discrete RV, then  $X$  is a function of  $Y$ , so for all  $y$  with  $p(y) > 0$ , there is only one  $x$  with  $p(x, y) > 0$ . Proof is not hard, but also not completely trivial.



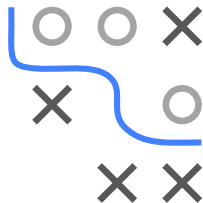
# MUTUAL INFORMATION

- The MI describes the amount of info about one RV obtained through another RV or how different their joint distribution is from pure independence.
- Consider two RVs  $X$  and  $Y$  with a joint pmf  $p(x, y)$  and marginal pmfs  $p(x)$  and  $p(y)$ . The MI  $I(X; Y)$  is the Kullback-Leibler Divergence between the joint distribution and the product distribution  $p(x)p(y)$ :

$$\begin{aligned} I(X; Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= D_{KL}(p(x, y) \| p(x)p(y)) \\ &= \mathbb{E}_{p(x, y)} \left[ \log \frac{p(X, Y)}{p(X)p(Y)} \right]. \end{aligned}$$

- For two continuous random variables with joint density  $f(x, y)$ :

$$I(X; Y) = \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dx dy.$$

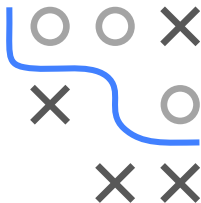


# MUTUAL INFORMATION

We can rewrite the definition of mutual information  $I(X; Y)$  as

$$\begin{aligned} I(X; Y) &= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x, y) \log \frac{p(x|y)}{p(x)} \\ &= - \sum_{x,y} p(x, y) \log p(x) + \sum_{x,y} p(x, y) \log p(x|y) \\ &= - \sum_x p(x) \log p(x) - \left( - \sum_{x,y} p(x, y) \log p(x|y) \right) \\ &= H(X) - H(X|Y). \end{aligned}$$

So,  $I(X; Y)$  is reduction in uncertainty of  $X$  due to knowledge of  $Y$ .



# MUTUAL INFORMATION

The following relations hold:

$$I(X; Y) = H(X) - H(X|Y)$$

$$I(X; Y) = H(Y) - H(Y|X)$$

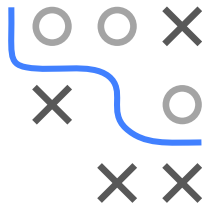
$$I(X; Y) \leq \min\{H(X), H(Y)\} \text{ if } X, Y \text{ are discrete RVs}$$

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$

$$I(X; Y) = I(Y; X)$$

$$I(X; X) = H(X)$$

All of the above are trivial to prove.

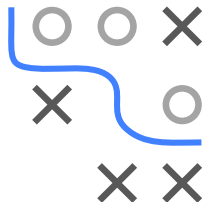




# MUTUAL INFORMATION - EXAMPLE

Let  $X, Y$  have the following joint distribution:

	$X_1$	$X_2$	$X_3$	$X_4$
$Y_1$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
$Y_2$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
$Y_3$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
$Y_4$	$\frac{1}{4}$	0	0	0



Marginal distribution of  $X$  is  $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$  and marginal distribution of  $Y$  is  $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ , and hence  $H(X) = \frac{7}{4}$  bits and  $H(Y) = 2$  bits.