# Exercise 2 – Information Theory

1. Mutual Information
    a. Show that $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
    b. Conditional vs. unconditional Mutual Information
        i. Give an example for three random variables such that
        $$I(X;Y|Z) < I(X;Y)$$
        ii. Give an example for three random variables such that
        $$I(X;Y|Z) > I(X;Y)$$

2. Let $X$, $Y$, $Z$ three random variables who form a *Markov chain* $X \to Y \to Z$
Show that $X$, $Y$ are conditionally independent given $Z$, i.e.
$$p(x,z|y) = p(x|y)p(z|y)$$

3. Let $p(x,y)$ be given by

| $\boldsymbol{Pr(x,y)}$ | $\boldsymbol{x_1}$ | $\boldsymbol{x_2}$ |
|---|---|---|
| $\boldsymbol{y_1}$ | 1/3 | 1/3 |
| $\boldsymbol{y_2}$ | 0 | 1/3 |

Find
    a. $H(X), H(Y)$
    b. $H(X|Y), H(Y|X)$
    c. $H(X,Y)$
    d. $H(X) - H(Y)$
    e. $H(X), H(Y)$
    f. $I(X;Y)$
    g. Draw a Venn diagram that illustrates the quantities stated in the above bullets (“a” to “f”)

4. Grouping rule for Entropy
Let $p = (p_1, p_2, \dots, p_m)$ be a probability distribution on $m$ elements (i.e. $p_i \geq 0$) and $\sum_i p_i = 1$). Define a new distribution on $q$ on $m-1$ elements such that the distribution on the first $m-2$ elements is identical, and the probability of last element in $q$ is the sum of the last two probabilities in $p$, i.e.
$$q_1 = p_1\,, \qquad q_2 = p_2\,, \qquad \dots\,, \qquad q_{m-2} = p_{m-2}\,, \qquad q_{m-1} = p_{m-1} + p_m$$

Show that
$$H(p) = H(q) + (p_{m-1} + p_m)H(\nu)$$
where $\nu = \left(\frac{p_{m-1}}{p_{m-1}+p_m}, \frac{p_m}{p_{m-1}+p_m}\right)$ is a binary probability distribution

5. In general, Relative Entropy is not symmetric, namely $D(p||q) \neq D(q||p)$. Give an example for two ***not identical*** distributions, $p \not\equiv q$, such that $D(p||q) = D(q||p)$

6. Relative Entropy $D(p||q)$ and chi-square ($\chi^2$)
Show that the $\chi^2$ statistics
$$\chi^2 = \sum_x \frac{\left(p(x) - q(x)\right)^2}{q(x)}$$
is (twice) the first term in the Taylor series expansion of $D(p||q)$ around $q$. Thus,
$$D(p||q) = \frac{1}{2}\chi^2 + \cdots$$
namely chi-square is a first order approximation of the relative entropy.
<u>Hint:</u> Write $\frac{p}{q} = 1 + \frac{p-q}{q}$ and expand the $\log(\cdot)$

7. Min Relative Entropy under constraints
Let $p(x), q(x),\ x \in \mathcal{X}$ two probability mass function, and let $f_1, f_2, \dots, f_n$ where $f_j: \mathcal{X} \to \mathbb{R}$ be feature functions. Given expectation constraints on the features $\sum_x p(x) f_j(x) = \alpha_j \quad \forall j$ , what is the $p^*$ that minimize the relative entropy $D(p||q)$ ? Solve the following
$$p^* = \arg\max_{p \in \mathcal{P}} D(p||q)$$
Where $\mathcal{P} = \left\{p: E_p\left[f_j\right] = \alpha_j,\ \ \forall j\right\}$
Use Lagrange multipliers to derive an explicit form to $p^*$