CK0146: Exercise 02 (2017.1)

Exercise Q.1. In the English language, there are 27 possible letters `a-z` and the space `-` character. You are given a collection of N `English language documents`, for each document, you have to:

- Calculate and plot the probability distribution $P(x)$ over the 27 letters x.
- Calculate and plot the probability distribution $P(x, y)$ over the 27×27 possible (ordered) di-grams xy. Note: For this joint distribution, its two marginals $P(x)$ and $P(y)$ are identical.
- From this joint distribution, calculate and plot the conditional distributions *i*) $P(y|x)$, the conditional distribution of the second letter y, given the first letter x; and *ii*) $P(x|y)$, the conditional distribution of the first letter x, given the second letter y.
- Calculate and plot the Shannon information content $h(x) = \log_2(1/P(x))$ of an outcome x and the entropy of a randomly selected letter $H(X) = \sum_x P(x) \log_2(1/P(x))$, assuming its probability is given by $P(x)$. Entropy is additive for independent random variables ($H(X, Y) = H(X) + H(Y)$ iff $P(x, y) = p(x)p(y)$), is this true for the document?

For the i-th and the j-th document, you must calculate the relative entropy (or, Kullback-Leibler divergence) between each pair of probability distributions $P_i(x)$ and $P_j(x)$ over the same alphabet

$$D_{KL}(P_i||P_j) = \sum_x P_i(x) \log \frac{P_i(x)}{P_j(x)}, \quad \forall i, j.$$

In general, the relative entropy is not symmetric under interchange of the distributions P_i and P_j (in general, $D_{KL}(P_i||P_j) \neq D_{KL}(P_j||P_i)$). Plot the $N \times N$ matrix of KL divergences to show this.

Exercise Q.2. Let X and Y be two random variables with joint probability density function

$$f_{XY}(x, y) = \begin{cases} 6y, & 0 < y < x < 1 \\ 0, & \text{elsewhere} \end{cases}.$$

Plot this joint pdf, calculate and plot the marginal pdf $f_X(x)$ of X, calculate and plot the conditional pdf $f_{Y|X}(y|x)$ of Y given $X = x$, calculate and plot the conditional mean $E(Y|x)$ of Y given $X = x$.