Statistical Models and Data Analysis

Summer term 2018

Problem Set 11 **2.7.2018**

The solutions to this exercise should be ready by 2 pm on 9.7.2018. If you have any questions, please send me an email at stemmler@bio.lmu.de.

---

1. (Entropy) As described in the lectures (and the lecture script), the mutual information between two discrete random variables X and Y is

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

where $H(X) = -\sum_x p_X(X=x)\log_2\left[p_X(X=x)\right]$ is known as the entropy.

(a) Let X be a binary random variable, with p be the probability of success (in a trial) , and $q = 1-p$ the probability of failure. Compute the entropy $H(X)$ and **plot** it as a function $0 \leq p \leq 1$.

(b) Consider a spike train as a random binary string of 0's and 1's. We divide the train into N bins (with $N \gg 1$). Let $N_1 = pN$ be the number of 1's. $N_0 = (1-p)N$ will be the number of zeros. The entropy of such a train is the logarithm of the number of possible arrangements of 0's and 1's. In other words,

$$H(X) = \log_2\left(\frac{N!}{N_1!N_0!}\right)$$

Use Stirling's approximation for the factorial, $N! = N\ln N - N$ to show that you can write

$$H(X) = -\frac{N}{\ln 2}\left[p\ln p + (1-p)\ln(1-p)\right].$$

2. (Entropy and Information) Imagine that every letter in the English language occurs with equal likelihood. We ignore spaces and punctuation. There are 26 letters in the alphabet, and you receive a stream of these letters on your mobile device, which you are checking incessantly, instead of doing your homework (tsk, tsk!). Only now, because your wi-fi reception is poor, every 8th letter is garbled, i.e., replaced by a random letter in the alphabet.

(a) Compute the source entropy rate in bits/symbol.

(b) Compute the noise entropy rate. From the noise and source entropy rates, calculate the information rate, once again in bits/symbol.

(c) (No computation:) In reality, not all letters of the alphabet have the same probability. The letter 'e', for instance, is much more common than the letter 'q'. Qualitatively, how will this affect the entropy and information rates above? Assume that errors occur in the same way as before, i.e., a symbol is replaced by. a random symbol. All possible replacements are equally likely, such that $p = 1/26$ for any letter that results from replacement. Messaging often uses abbreviations like YOLO and LOL and 'u up'? How do abbrevations like this affect the source entropy?

(d) I told my cat YOLO once, and she gave me a quizzical look. Please explain.