Exercise Sheet #11

Fabian Schubert <fschubert@itp.uni-frankfurt.de>
Oren Neumann <neumann@itp.uni-frankfurt.de>

Problem 1 (*Variation of Information*) 7 Pts

The variation of information of two random variables can be regarded as a form of 'entropy distance' and is given by

$$D_H(X,Y) = H(X,Y) - I(X;Y);, \tag{1}$$

i.e., the difference between their joint entropy and the mutual information. Prove that this quantity satisfies the properties of a distance metric:

- $D_H(X,Y) \geq 0$
- $D_H(X,X) = 0$
- $D_H(X,Y) = D_H(Y,X)$
- $D_H(X,Z) \leq D_H(X,Y) + D_H(Y,Z)$

Problem 2 (*Mutual Information in a Markov Chain*) 7 Pts

Let $X_1 \to X_2 \to X_3 \to \cdots \to X_n$ form a Markov chain in this order; i.e., let

$$p(x_1, x_2, ..., x_n) = p(x_1)p(x_2|x_1) \cdots p(x_n|x_{n-1}) \ . \tag{2}$$

Show that the mutual information $I(X_1; X_2, ..., X_n)$ reduces to $I(X_1; X_2)$. *Hint: use the Markov property in the definition of the conditional entropy.*

Problem 3 (*Kullback-Leibler Divergence and Maximum Likelihood*) 6 Pts

Suppose you have a random variable x that follows some underlying probability distribution $p_x(x)$. Furthermore, suppose you want to model this random variable by proposing some distribution $p_{\text{model}}(x|\theta)$ that depends on a parameter θ. You have access to a set of N samples of x, i.e. $\{x_1, ..., x_N\}$ and you would like to estimate θ from these. One way to do so is by maximizing the likelihood function

$$\theta_{\text{MLE}} = \arg\max_\theta \prod_{i=1}^{N} p_{\text{model}}(x_i|\theta) \ . \tag{3}$$

Show that for $N \to \infty$, this yields the same result as

$$\theta_{\text{KL}} = \arg\min_{\theta} D_{KL}\left(p_x(x) || p_{\text{model}}(x|\theta)\right) , \tag{4}$$

which is the minimum with respect to θ of the Kullback-Leibler divergence between the true underlying distribution and your model distribution.
Hint: Start with the expression for θ_{KL} *and show that this becomes* θ_{MLE} *for large* N.