Exercise Sheet 2 (mini)

Probabilistic models of modern AI

vassilis.papadopoulos@epfl.ch – www.vassi.life/teaching/aiproba

References :

- Cover & Thomas, 'Elements of Information Theory', Chapter 2 [link]

Throughout the exercise we use the mixed discrete/continuous notation to denote the probability distributions of a random variable X. We write $p(x) = \mathrm{P}(X = x)$[1], or sometimes $p(x_i) \equiv P(X = x_i) \equiv p_i$ when the possible values of X are discrete. To denote the full probability distribution, we sometimes write just p. In the case of a continuous random variable ($X \in \mathbb{R}$), $p : \mathbb{R} \to [0, 1]$ is a function, whereas for a discrete random variable ($X \in \{x_1, \cdots x_n\}$) $p \in [0, 1]^n$ is a vector. For all exercises, you can assume the case of X being discrete, and most often all proofs carry over to the continuous case seamlessly.

When writing sums/integrals, we will often use the unified notation as in $\sum_x p(x)f(x)$. This should be interpreted as being an integral $\int_x p(x)f(x)dx$ if X is a continuous random variable, or a sum $\sum_i p(x_i)f(x_i)$ if X is discrete.

Exercise 1 Proper scoring rules encore

Consider $s_i(r)$ to be the score assigned to a prediction $r \in \mathbb{R}^n$, if event $i \in [[1, n]]$ is realized.

1. Find all the *proper* scoring rules such that $s_i(r) = f(r_i)$, that is, the score $s_i(r)$ depends only on the probability assigned to the event i, which is the one that actually happened. Let's call such a scoring rule *proper and independent*
2. Why is this a reasonable property for the scoring rule to have?

Exercise 2 Entropy

For a probability distribution $p(x)$, the entropy of the distribution is defined as $H(p) = -\sum_x p(x) \log_a p(x)$. In physics, we usually use ln, namely $a = e$, whereas in computer science, we use $a = 2$. Here we will use $\log \equiv \log_2$. Often, we say that the entropy of a distribution quantifies the degree of uncertainty in the outcome of the distribution. If the probability distribution p_X is associated to a random variable X, we sometimes write $H(X) = H(p_X)$. Note that values that the random variable can take do NOT affect the value of its entropy.

1. Compute the entropy of the discrete distribution $p(x_i) = \delta_{i,j}$ where $\delta_{i,j} = 1$ iff $i = j$
2. Compute the entropy of the uniform discrete distribution over n possibilities
3. Compute the entropy of the Normal distribution $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
   a) Why is it independent of μ ?
4. For a discrete random variable with n outcomes, find the distribution p_i that maximizes the entropy.
5. Let two discrete independent random variables X_1 and X_2 (with probability distribution p_1 and p_2). Denote O_1, O_2 two finite sets of the possible values of X_1, X_2. We construct a third random variable $X_3 = (X_1, X_2)$ whose outcomes are in $O_1 \times O_2$. Compute $H(X_3)$ as a function of $H(X_1)$ and $H(X_2)$.

Next week, we will see that the entropy of a random variable is the central quantity related to compression; It tells us that if we want to transmit the outcomes of a random variable X to somebody, the average number of bits that we will need to send is $H(X)$.

[1] Technically, when X is continuous we should introduce a density function $f(x)$, such that $\mathrm{P}(X = x) = f(x)dx$, but let's not care too much about this at least for now.