Introduction to Information Theory, Fall 2020

Practice problems for exercise class #7

You do **not** have to hand in these exercises, they are for your practice only.

0. **Exercises from MacKay:** 6.7

1. **Aritmetic coding with an IID source:** Consider a bitstring modelled by a sequence of IID random variables X_i with probabilities $P(\mathtt{0}) = 1/4$ and $P(\mathtt{1}) = 3/4$.

   (a) Order the bits as $\mathtt{0} < \mathtt{1}$. What ordering does this give rise to on bitstrings of length N?
   (b) Use the Shannon-Fano-Elias code to encode `010`.
   (c) Apply the arithmetic coding algorithm to encode `010`.

2. **Shannon-Fano-Elias code and arithmetic coding:**

   (a) Show that if you follow the arithmetic coding algorithm from the lecture without step 2, and then compute the binary expansion of $\frac{Q+R}{2}$ of length $\lceil \log(\frac{2}{P(x)}) \rceil$ you obtain the Shannon-Fano-Elias code.
   (b) Remark that when applying Shannon-Fano-Elias coding to x^N for large N, you will probably need high-precision arithmetic for computing $\lceil \log(\frac{2}{P(x)}) \rceil$. How does arithmetic coding you avoid this?
   (c) Explain why, in the arithmetic coding algorithm, step 3 comes *after* step 2.

3. **Language models: k-grams**

   (a) Let $k \geqslant 1$. We want to consider a language model, called the k-*gram* model, where the probability of a letter depends on the previous $k-1$ letters only. To make this formal, for $k = 1$ we would assume that $P(x_n | x_1 \ldots x_{n-1}) = P(x_n)$. For $k > 1$ we assume that $P(x_n | x_1 \ldots x_{n-1}) = P(x_n | x_{n-k+1} \ldots x_{n-1})$. We assume that these conditional probabilities are the same for each n. The $k = 1$ case thus corresponds to the IID case. Explain how to estimate these probabilities from the string x^N for large N.
   (b) Now imagine a language which consists of all English words, and which is such that the probability of a word only depends on the previous word, with the probabilities as in 'real' English. Think about how to sample from this language by hand only using a book (representative for the English language).

4. **Decompressing arithmetic codes:**

   (a) In Algorithm 1 a naive decompression algorithm is described. The inputs are the bitstring $b = b_1 b_2 \ldots$ and the length N of the source string x^N. We denote by $\mathcal{A}$ the (ordered) alphabet of the source, by $Q(x_n | x_1, \ldots, x_{n-1})$ and $R(x_n | x_1, \ldots, x_{n-1})$ the cumulative conditional probabilities, and by $0.b = 0.b_1 b_2 \ldots$ the number in $[0, 1)$ whose binary expansion is given by the bitstring b (followed by infinitely many zeros). Argue that this algorithm decompresses correctly.
   (b) Can you make a 'streaming' version of the decompression algorithm?

Algorithm 1 Decompress arithmetic coding

```
procedure DECOMPRESS(b, N)
    u ← 0
    p ← 1
    x ← []
    for n = 1, ..., N do
        for y ∈ A do
            U ← u + pQ(y|x_1, ..., x_{n−1})
            V ← u + pR(y|x_1, ..., x_{n−1})
            if 0.b ∈ [U, V) then
                x ← x + [y]
                u ← U
                p ← V − U
            end if
        end for
    end for
    return x
end procedure
```

5. **Learning on the fly (mathematics challenge):** For a streaming algorithm, you usually do not want to estimate the language model by looking at the whole string x^N, but rather learn them 'on the fly' as you are going through the message. In this exercise we will use Bayes rule to derive one such procedure for an IID source with unknown initial probabilities.

   (a) Suppose that we have an IID source X on an alphabet $\{a, b\}$, but we do not know the probabilities $P(a) = p_a$, $P(b) = p_b = 1 - p_a$. Show that if we assume a uniform prior on p_a (so with probability *density*[1] $P(p_a) = 1$ for $p_a \in [0, 1]$) and observe x^N with N_a times a and N_b times b, then Bayes rule tells use that

   $$P(p_a|x^N) = \frac{p_a^{N_a}(1-p_a)^{N_b}}{P(x^N)}$$

   where

   $$\begin{aligned} P(x^N) &= \int_0^1 P(x^N|p_a)P(p_a)dp_a \\ &= \int_0^1 p_a^{N_a}(1-p_a)^{N_b}dp_a. \end{aligned}$$

   (b) Use the fact that

   $$\int_0^1 p_a^{N_a}(1-p_a)^{N_b}dp_a = \frac{N_a!N_b!}{(N_a+N_b+2)}$$

   and the previous question to show that

   $$\begin{aligned} P(a|x^N) &= \int P(a|p_a)P(p_a|x^N)dp_a \\ &= \frac{N_a+1}{N_a+N_b+2}. \end{aligned}$$

[1] We will have to use a continuous version of Bayes rule here, where sums are replaced by integrals and probabilities by probability densities (which we didn't discuss and you don't have to know this for the exam).

This is known as Laplace's rule.

(c) Explain how to use Laplace's rule for arithmetic coding.

Notice that the rule we have derived depends on the prior! If we would have taken another prior we would have obtained different results.