Adaptive Information Processing

Hints to the exercises

for Model complexity and the MDL principle

Bert de Vries and Tjalling Tjalkens
Signal Processing Group

March 5, 2009

Bayes and the Laplace method

1. The two envelope paradox. See `http://www.anc.ed.ac.uk/~amos/doubleswap.html` for an answer.

2. Consider the following integral (similar to the Beta integral)

$$F(\mu_1, \mu_2) = \int_{-\infty}^{\infty} \left(\frac{1}{1+e^{-a}}\right)^{\mu_1} \left(\frac{e^{-a}}{1+e^{-a}}\right)^{\mu_2} da.$$

   (a) Use Laplace's method to approximate this integral.

   Hint:

   Just follow the steps in the lecture notes.

   (b) Use the Beta integral

$$B(\mu_1, \mu_2) = \int_0^1 p^{\mu_1 - 1}(1-p)^{\mu_2 - 1}\, dp = \frac{\Gamma(\mu_1)\Gamma(\mu_2)}{\Gamma(\mu_1+\mu_2)}$$

   with

$$\Gamma(x+1) = x\Gamma(x)$$
$$\Gamma(1) = 1$$
$$\Gamma(0.5) = \sqrt{\pi}$$

   and compare your approximation with the actual values in the cases where $\mu_1 = \mu_2 = 0.5$ resp. $\mu_1 = \mu_2 = 1$.

   Hint:

   Consider the transformation from p to $\frac{1}{1+e^{-a}}$ in the Beta integral.

Universal data compression

1. The Shannon-Fano code and Huffman code.

   Consider a binary i.i.d. souce that generates $X_1, X_2, \ldots, X_n$ with the parameter $\theta = \Pr\{X = 1\} = 0.1$.

   Compute, for $n = 1, 2, 3$, the expected code wordlength for the Shannon-Fano code, with lengths

   $$l_C^*(x^n) = \lceil -\log_2 p(x^n) \rceil.$$

   Likewise for the Huffman procedure, see lecture notes Information Theory (5K020/5JJ40).

   Give your comments on this result, (and consider here the source entropy).

   Hint:

   No hint needed, just do it. For the Huffman procedure you can read the lecture notes of Information Theory I or any basic Information Theory textbook.

2. [Hard] Show that

   $$\bar{p}(x^n) < \sqrt{\frac{\pi}{2n}} e^{\frac{1}{3n}} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k},$$

   where

   $$\bar{p}(x^n) = \int_0^1 (1-\theta)^{N(0|x^n)} \theta^{N(1|x^n)} \, d\theta.$$

   Hint:

   It's hard and you're all on your own. You can make good use of Stirling's formula.

ML and MDL

1. Assume x^n are i.i.d. observations from $\mathcal{N}(\theta, 1)$, so the x_i's are independent Gaussians with unit variance but unknown mean $\theta \in \mathbb{R}$. We test two hypothesis, $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Otherwise said, we want to choose between the models

   $$\mathcal{M}_0 = \{\mathcal{N}(0, 1)\} \text{ and } \mathcal{M}_1 = \{\mathcal{N}(\theta, 1) | \theta \neq 0\}$$

   Derive that if we compute the ML probabilities for each model and then choose for the model with the largest ML probability we will never choose for $\mathcal{M}_0$ even if x^n was actually generated by $\mathcal{M}_0$.

   Hint:

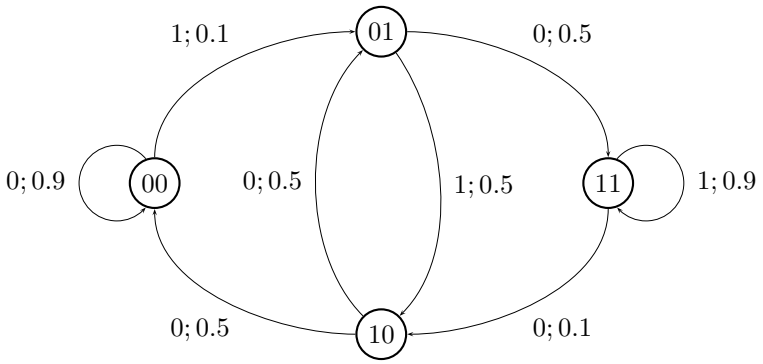
   The proof is based on the fact that for $\mathcal{M}_1$ we estimate the mean

and only if the estimated mean is equal to zero, $\mathcal{M}_0$ will be selected. This is an event of probability zero.

Compare this to the MDL result of pages 124–125. Doesn't that result seem more satisfying?

---

2. Consider this following 1^{th}-order binary Markov source. Next to the arrow from state a to state b is written $x; \Pr\{X_i = x, S_i = b | S_{i-1} = a\}$.



   (a) Determine the probability $\Pr\{X_i = 1\}$.
   Hint: Compute the stationary state distribution and then marginalize $\Pr\{X_i = 1, S_{i-1} = s$ to obtain $\Pr\{X_i = 1\}$.

   (b) Consider an "ideal" universal datacompression algorithm and we observe a sequence x^n that is typical for the source. How large must i be approximately to select the first order Markov model in stead of the memoryless model.

Hint:

Look at the discussion in the notes " Stochastic Complexity (MDL)". This is similar! Also the lecture notes Information Theory (5K020/5JJ40) will be helpful.

---

3. [Hard:] can you determine the number of suffix trees with maximal depth not more than D for $D = 0, 1, 2, \ldots, 10$?

Hint:

You should be on your own again but ok.

Let F_i be the number of binary trees of maximum depth at most i. Then

$$F_i = 1 + F_{i-1}^2$$

The 1 comes from the tree containing only a root and otherwise all trees with maximum depth i can be written as the tree of depth 1 having both a tree of depth at most $i-1$ at its zero node and its one node.

D	$\#trees$
0	1
1	2
2	5
3	26
	...
10	$1.437821978 \cdot 10^{181}$

(Unless I made an error in calculating this recursion.)

So the number of models is HUGE for reasonable D. e.g. we use this CTW method for text compression on a byte depth of 12 characters. This is $D = 96$. Still the amount of work is linear in the sequence length times D.
