Adaptive Information Processing
Exercises
for Model complexity and the MDL principle

Bert de Vries and Tjalling Tjalkens
Signal Processing Group

March 5, 2009

Abstract

The exercises are intended to illustrate the results and deepen your understanding. Their level is sometimes higher than expected for the exam. The mark [Hard:] indicates an exercise above the exam level.

Tjalling Tjalkens, 3 March, 2006, Updated for reduced contents, March 2009.

Bayes and the Laplace method

1. The two envelope paradox.

   See `http://www.anc.ed.ac.uk/~amos/doubleswap.html` for a nice paradox that can be solved using the Bayes approach. The answer is also given on the web site.

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.

   (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

$$\begin{aligned} \Gamma(x+1) &= x\Gamma(x) \\ \Gamma(1) &= 1 \\ \Gamma(0.5) &= \sqrt{\pi} \end{aligned}$$

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

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).

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

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

   So, we use a *uniform* prior over θ.

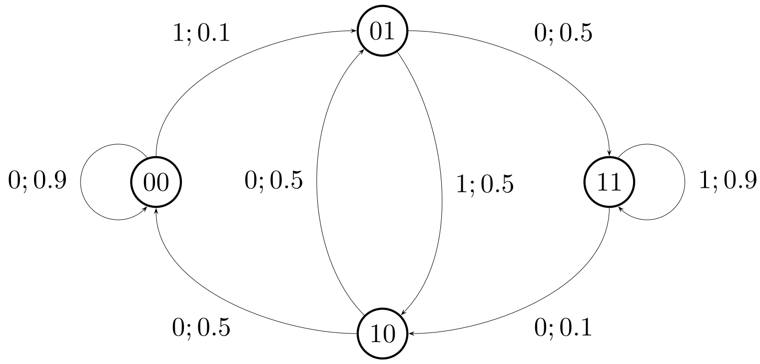
ML and MDL

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 $\mathcal{M}_0$.

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.

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