MDL exam, Thursday May 22nd 2014, 10.30-13.30, Room 401

You can gain 100 points in total with exercises 1–5; you get 10 bonus points for free, so you can obtain 110 points in total.
Your final score is given by min{total number of points achieved/10, 10}. As you know, if you find an exercise very difficult, it may be wiser not to spend too much time on it but rather try another exercise first.

1. *Bernoulli Rules* [20].

   a) Consider binary data $x^n \in \{0,1\}^n$ and a model consisting of just three distributions: Bernoulli with parameters $0, 0.5$ and 1. We are going to code data using the Normalized Maximum Likelihood (NML) code relative to this model (see e.g. Chapter 6, Section 6.2 in the MDL book). Give a formula for the worst-case regret as a function of n. For what sequence(s) of length n is the actual regret of the NML code equal to the worst-case regret?

   b) Suppose we have a model consisting of a finite number of M distributions $P_1, \ldots, P_M$ defined over data of length n. We now change the model by adding one more distribution P_{M+1}.

   (i) Prove that the worst-case regret achieved by the NML code for the new, larger model is at least as large as for the original model.

   (ii) Under what condition on $P_1, \ldots, P_{M+1}$ do we get that the NML code for both the original and the enlarged model achieve equal worst-case regret?

   (iii) Give an example that shows that the worst-case regret achieved by the predictive plug-in universal code based on maximum-likelihood estimators can actually be strictly *smaller* for the larger model than for the smaller model. If the maximum likelihood estimator is undefined or not unique for data x^i, then you may assume that your predictive code predicts/codes based on P_1. (you don't have to give a full mathematical proof for your answer. It is sufficient to give an example and argue informally that for that example, the worst-case regret for the larger model will be strictly smaller than for the smaller model).

2. *Is it Real?* [20] Consider the Rational Bernoulli model. $\mathcal{B}_{\mathbb{Q}} = \{P_\theta | \theta \in [0,1] \cap \mathbb{Q}\}$ where $\mathbb{Q}$ stands for the set of rational numbers (the set of numbers which can be written as p/q for integer p and q). As always, $P_\theta(x^n) := \theta^{n_1}(1-\theta)^{n_0}$.

   We compare the rational Bernoulli model to the ordinary Bernoulli model.

   a) Which model is larger?

   b) Compute the difference between the complexity terms (the log of the normalizing sum in the NML distribution) for the Bernoulli and the rational Bernoulli model.

   c) Design a two-part code L such that for every $P \in \mathcal{B}_{\mathbb{Q}}$, there exists a fixed constant $C_P > 0$ (dependent on P but not n) such that for all n and x^n, we have:

   $$L(x^n) < -\log P(x^n) + C_P. \tag{1}$$

   d) Does the NML code satisfy (1)?

3. *Fat Tails* [20]. Let the sample space $\mathcal{X} = \mathbb{N} = \{1, 2, 3, 4, \ldots\}$ be equal to the natural numbers.

a) Let Q be any probability distribution on $\mathcal{X} = \mathbb{N}$ with decreasing probabilities, i.e. $Q(1) \geq Q(2) \geq Q(3) \geq Q(4) \geq \ldots$. Show that if $E_Q[X] < \infty$, then we must also have $H(Q) < \infty$, where H is the entropy (*hint:* You already score nearly all available points if you can show this for distributions Q of the form $Q(x) = Cx^{-\alpha}$ for some $\alpha > 0$ and a constant C)

b) Give an example of a distribution P' on $\mathcal{X} = \mathbb{N}$ with $E_{P'}[X] = \infty$ but $H(P') < \infty$.

4. *Model Selection* [20]. Suppose data $x_1, x_2, \ldots, x_n \in \{0,1\}^n$ are modeled by a list of models $\mathcal{M}_0, \mathcal{M}_1, \ldots$, where $\mathcal{M}_K$ is the K-th order Markov model. In all exercises below, you do not have to formally proof your answer (that would go beyond what you learned in class). We just require you to provide a well-motivated guess.

   (a) Suppose the data x^n are sampled from the 10th-order Markov model (i.e. the "true distribution" is a 10-th order Markov chain that cannot be rewritten as a 9-th order chain). For each initial segment of the data x^i, we do MDL model selection based on two-part codes. What Markov order will be selected (with high probability according to the true distribution) by MDL for small i, for intermediate i and for large i?

   (b) Suppose the data x^n are equal to the binary expansion of $\pi = 3.1415\ldots$. What Markov order do you think will be selected by MDL for small i, for intermediate i and for large i? You may use the following intriguing fact: intensive computer experiments have shown that, for all n up until $n = 10^9$, the following holds: let $j \ll n$. Then for all sequences $y^j \in \{0,1\}^j$, the number of occurences of y^j followed by a 1 in x^n is approximately equal to the number of occurrences in x^n of y^j followed by a 0. Although there is no proof, most mathematicians think that this actually holds not just for $n \leq 10^9$, but for all n.

   (c) Suppose the data are equal to the binary expansion of $\pi = 3.1415\ldots$, but whenever there appears a 0, the next bit is replaced by a 1. So if y^n are the first n bits of π written in binary, then for $1 \leq i \leq n$, $x_i = y_i$ if $x_{i-1} \neq 0$, and $x_i = 1$ otherwise. What Markov order do you think will be selected by MDL for small i, for intermediate i and for large i?

5. *Hardy-Weinberg* [20]. In population biology, one is often interested in the probability of the alleles of some gene. For example, the "human eye color gene" can have alleles blue, brown and green. Here we study the simple case of genes with only two alleles, say a and b. Every human or animal has two alleles for each gene, one stemming from the mother, the other from the father. Thus, in the population, one encounters four different combinations: aa, ab, ba, bb. However, ab and ba express themselves in the same way. This means that the combination ab (a from mother, b from father) leads to the same observable trait as the combination ba (b from mother, a from father). For example, the Scarlet Tiger Moth has a gene which determines how many white spots it has. Allele a codes for "many spots", b codes for "few spots", and moths with ab or ba will have an intermediate number of spots; but there will be no visible difference between moths with ab or ba. For ease of analysis we will encode all three outcomes as a number: -1 stands for aa, 1 stands for bb, and 0 stands for $\{ab, ba\}$.

   According to the celebrated *Hardy-Weinberg law*, the probability θ that an individual has obtained allele a from the mother is the same as the probability that it has attained a from the father, and the two probabilities are independent. We may think of the Hardy-Weinberg law as imposing a one-dimensional probability model $\mathcal{M}_1 = \{P_\theta \mid \theta \in [0,1]\}$ on the sample space with three (not four!) outcomes $\mathcal{X} = \{-1, 0, 1\}$, with

$$P_\theta(X = 1) = \theta^2, P_\theta(X = -1) = (1-\theta)^2, P_\theta(X = 0) = 2\theta(1-\theta). \tag{2}$$

(a) Derive equation (2) from the assumption that (a) the probability of getting a from the father is the same as the probability from getting it from the mother, and (2) the two probability distributions are independent

(b) Show that the Hardy-Weinberg model (2) is an exponential family with carrier $r(-1) = r(1) = 1/2$, $r(0) = 1$ (Hint: write down equations, one for each value of x, relating θ and the canonical parameter β; then solve these equations).

(c) Is the parameterization given in (2) equal to either the canonical or the mean-value parameterization? Explain your answer.

(d) Compute the Fisher information and Jeffreys' prior for the Hardy-Weinberg model.