MDL exam, 22 May 2012

You start off with one point, and can earn up to 10 points — if you decide to answer the Bonus Question 3b, you can get up to 11 points. Do not hesitate to ask for help if you don't understand a question.

1. When in Doubt, Normalize

We will investigate two very different models for binary data of some fixed length n. The first model $\mathcal{M}_1 = \{P_\theta \mid \theta \in \{0, \frac{1}{2}, 1\}\}$ consists of just three Bernoulli distributions, extended to n outcomes and parameterised by the mean as usual.

(a:1) Calculate the minimax regret, i.e. the smallest worst-case regret $\max_{x^n} \mathcal{R}(P, \mathcal{M}_1, x^n)$ that can be achieved by some P. What distribution P achieves this? Would you call model $\mathcal{M}_1$ "simple" or "complex"?

(b:1) We now impose the constraint that we will use an (idealised) two-part code for $\mathcal{M}_1$. Describe the two-part code that minimises the worst-case regret. How much larger is the worst-case regret compared to what you found in the previous question? What code would you prefer in practice?

The second model $\mathcal{M}_2 = \{P_\alpha \mid 0 < \alpha < \infty\}$ is somewhat unusual: its distributions are defined as $P_\alpha(x^n) = 1$ if the first n digits of the binary expansion (behind the 'binary' rather than 'decimal' point) of $\pi^{-\alpha}$ coincide with x^n, and 0 otherwise. Here π is the well-known constant, 3.14.... For example, for sufficiently small α, we have $P_\alpha(1^n) = 1$. For this second model, we will ask roughly the same questions:

(c:1) First, calculate the maximum likelihood for data x^n, i.e. $\max_{0<\alpha<\infty} P_\alpha(x^n)$, as a function of x^n. Next, calculate the minimax regret, i.e. the smallest worst-case regret $\max_{x^n} \mathcal{R}(P, \mathcal{M}_2, x^n)$ that can be achieved by some P. What distribution P achieves this? Would you call model $\mathcal{M}_2$ "simple" or "complex"?

(d:1) Now consider data x^n where each $x_i \in \mathcal{X}$ and $\mathcal{X}$ is the set of positive natural numbers. Let $\mathcal{M}_3 = \{P_\theta \mid \theta \in \Theta_3\}$ be any model with infinite minimax regret, so that the NML distribution is undefined. For example, $\mathcal{M}_3$ could be the Poisson model. One way of modifying NML so that it becomes well-defined is to include a prior distribution W on the (countable) set of parameters

$$\hat{\Theta}_n := \{\theta \in \Theta_3 : \theta = \hat{\theta} \text{ for some } x^n \in \mathcal{X}^n \}.$$

The new definition becomes

$$P_{\text{new-nml}}(x^n) := \frac{P_{\hat{\theta}(x^n)}(x^n) W(\hat{\theta}(x^n))}{\sum_{x^n \in \mathcal{X}^n} P_{\hat{\theta}(x^n)}(x^n) W(\hat{\theta}(x^n))}.$$

Show that $\sum_{x^n \in \mathcal{X}^n} P_{\hat{\theta}(x^n)}(x^n) W(\hat{\theta}(x^n)) \leq 1$ and hence finite, so that $P_{\text{new-nml}}$ is always well-defined [HINT: first relate, for every fixed $x^n \in \mathcal{X}^n$ $P_{\hat{\theta}(x^n)}(x^n) W(\hat{\theta}(x^n))$ to $P_{\text{Bayes}}(x^n)$, where $P_{\text{Bayes}}(x^n)$ is the Bayesian marginal distribution defined relative to the same prior W on $\hat{\Theta}_n$].

2. Entropy vs. Variance

Both entropy and variance are often used as measures of the 'inherent uncertainty' in a distribution, so it is interesting to find out how similar they are. Consider sample space $\mathcal{X} = \{1, 2, \ldots, N\}$ for some $N \geq 2$.

(a:1/4) What distribution P_{maxe} on $\mathcal{X}$ maximizes the entropy, and what is the entropy of P_{maxe}?

(b:1/4) What distribution P_{maxv} on $\mathcal{X}$ maximizes the variance, and what is the variance of P_{maxv}?

(c:1/8) What distribution(s) on $\mathcal{X}$ minimize the variance, and what distribution(s) on $\mathcal{X}$ minimize the entropy?

(d:3/4) Now let $\mathcal{X}$ be the positive natural numbers. Show that for every $\epsilon > 0$, no matter how small, and for every finite C, no matter how large, there exists a distribution on $\mathcal{X}$ that has entropy smaller than ϵ and variance greater than C.

(e:3/8) Is the converse of the previous question also true, i.e. does there exist, for arbitrarily large C and arbitrarily small ϵ, a distribution on the natural numbers with variance $< \epsilon$ but entropy $> C$? If so, what distribution is this? If not, why not?

3. Fat Tails

(a:1) For the following three distributions on the positive natural numbers, work out whether or not the mean is finite, and whether or not the entropy is finite.

- A geometric distribution: $P(n) = 2^{-n}$
- A heavy tailed distribution: $P(n) = 1/(n(n+1))$
- An even heavier-tailed distribution: $P(n) \propto 1/(n(\log n)^2)$

(b:1) This is the bonus question! Do not spend time on it before you have completed the rest of the exam. Prove that any distribution on the positive natural numbers with finite mean must have finite entropy. Hint: split the terms of the entropy into two groups: terms with $n \geq -\log P(n)$ and terms with $n < -\log P(n)$. Show that for both groups, the sum cannot diverge if the mean is finite.

4. Is it Real?

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, $P_\theta(x^n) := \theta^{n_1}(1-\theta)^{n_0}$.

In this question we compare the rational Bernoulli model to the ordinary Bernoulli model.

(a:1/4) Which model is larger?

CONTINUED ON NEXT PAGE!

(b:$\frac{1}{2}$) 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:1) 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:$\frac{1}{2}$) Does the NML code satisfy (1)?