# MDL exam, 26 May 2015

You start off with one point, and can earn up to 10 points. Don't spend too much time on questions that you find very difficult! — it is wiser to look ahead and see if you can solve an easier question first.

## 1. Normalized Maximum Likelihood

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 $\bar{P}$ — do not give an asymptotic approximation, but directly calculate the model complexity $\text{COMP}^{(n)}(\bar{P}) = \log \sum_{x^n \in \{0,1\}^n} P_{\hat{\theta}(x^n)}(x^n)$ and argue that it is equal to the minimax regret. What is $\bar{P}(x^n)$ for $x^n$ a sequence consisting of $n_1$ ones? (here $n_1 \in \{0, 1, \ldots, n\}$).

(b:1) We now impose the constraint that we will use a two-part code for $\mathcal{M}_1$, i.e. with codelengths of the form $L(x^n) = -\log P_\theta(x^n) + L'(\theta)$ for some codelength function $L'$ on $\theta \in \{0, 1/2, 1\}$. 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?

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 $\pi$ is the well-known constant, 3.14.... For example, for sufficiently small $\alpha$, we have $P_\alpha(1^n) = 1$ (because for any $0 < z < 1$, in particular for $z = \pi^{-1}$, we have $z^\alpha$ is decreasing in $\alpha$ and $\lim_{\alpha \downarrow 0} z^\alpha = 1$. For this second model, we will ask roughly the same questions:

(c:1) First, calculate the maximum likelihood for data $x^n$, i.e. $ML(x^n) := \max_{0<\alpha<\infty} P_\alpha(x^n)$, 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 $P$ (HINT: even though the ML estimator $\hat{\alpha}$ is not uniquely defined, the model complexity $\text{COMP}^{(n)}$ is still well-defined and you can use it to calculate minimax regret —see page 180 of the book). What distribution $P$ achieves this minimax regret? 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. 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:$\frac{1}{2}$) Which model is larger?

(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:

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

HINT: note that the constant $C_P$ does *not* depend on $n$. So this code must be different from the standard two-part code based on discretization of the model parameters (which asymptotically has a term that depends on $n$ but not on $P$). The code $L$ is not based on discretization — you really have to use that each $P$ has a parameter in $\mathbb{Q}$.

## 3. Pareto

The *Pareto distribution* with parameter $\alpha$ is the distribution on the natural numbers $\mathbb{N} = \{1, 2, \ldots\}$ with $P_\alpha(x) = x^{-\alpha}/C$, where $C = \sum_{x \in \mathbb{N}} x^{-\alpha}$. The *Pareto family* is the set of all Pareto distributions $P_\alpha$ with parameter $\alpha > 1$.

(a:1) Let $\mathcal{X} = \mathbb{N}$ and consider the set of distributions on $\mathcal{X}$ satisfying the constraint $E_{X \sim P}[\ln X] = t$, where ln denotes natural logarithm. Show that, if $t$ is some value for which a distribution satisfying the constraint exists, then the maximum entropy distribution, given the constraint, is a member of the Pareto family.

(b:1) As $t$ varies, the corresponding maximum entropy distributions form an exponential family which is thus equal to the Pareto family. How is the parameter $\alpha$ for Pareto distribution related to the parameter $\beta$ for the corresponding exponential family? Does the mean $t$ decrease or increase with $\alpha$?

(c:1) Now let the sample space $\mathcal{X} = \mathbb{Z}$ also include the negative integers. Consider the set $\mathcal{P}_0$ of distributions on $\mathcal{X}$ satisfying the constraint $E_{X \sim P}[X] = 0$. Show that $\sup_{P \in \mathcal{P}_0} H(P) = \infty$, i.e. the set $\mathcal{P}_0$ contains no maximum entropy distribution.