MLT Homework 13

Ana Borovac
Bas Haver

December 11, 2018

Question 1

Finite Θ.

Subquestion 1.1

Verify that $\bar{p}$ *is a probability mass function on* $\{0,1\}^m$.

Solution

$$\begin{aligned}
\sum_{(z_1,\ldots,z_m)\in\{0,1\}^m} \bar{p}(z_1,\ldots,z_m) &= \sum_{(z_1,\ldots,z_m)\in\{0,1\}^m} \sum_{\theta\in\Theta} w(\theta) p_\theta(z^m) \\
&= \sum_{(z_1,\ldots,z_m)\in\{0,1\}^m} \sum_{\theta\in\Theta} \frac{1}{N} p_\theta(z^m) \\
&= \frac{1}{N} \sum_{\theta\in\Theta} \sum_{(z_1,\ldots,z_m)\in\{0,1\}^m} p_\theta(z^m) \\
&= \frac{1}{N} \sum_{\theta\in\Theta} 1 \\
&= \frac{1}{N} N \\
&= 1
\end{aligned}$$

Subquestion 1.2

Show that the worst-case regret of $\bar{p}$ *is also bounded by* $\log N$.

Solution

We first show that $\sum_{i=1}^{m} -\log \bar{p}(z_i|z^{i-1}) = -\log \bar{p}(z^m)$ as in the hint:

$$\begin{aligned}
\sum_{i=1}^{m} -\log \bar{p}(z_i|z^{i-1}) &= -\log \bar{p}(z_1|\epsilon) - \cdots - \log \bar{p}(z_m|z^{m-1}) \\
&= -(\log \bar{p}(z_1|\epsilon) + \cdots + \log \bar{p}(z_m|z^{m-1})) \\
&= -\left(\log \frac{\bar{p}(z^1)}{\bar{p}(\epsilon)} + \cdots + \log \frac{\bar{p}(z^m)}{\bar{p}(z^{m-1})}\right) \\
&= -\left(\log \frac{\bar{p}(z^1)}{\bar{p}(\epsilon)} \cdot \cdots \cdot \frac{\bar{p}(z^m)}{\bar{p}(z^{m-1})}\right) \\
&= -\left(\log \frac{\bar{p}(z^m)}{\bar{p}(\epsilon)}\right) \\
&= -\log \bar{p}(z^m)
\end{aligned}$$

Now we have that the regret is given by:

$$\begin{aligned}
-\log \bar{p}(z^m) - \inf_{\theta\in\Theta} -\log p_\theta(z^m) &= \sum_{i=1}^{m} -\log \bar{p}(z_i|z^{i-1}) - \inf_{\theta\in\Theta} \sum_{i=1}^{m} -\log p_\theta(z_i|z^{i-1}) \\
&= -\log(\Pi_{i=1}^{m} \frac{\bar{p}(z^i)}{\bar{p}(z^{i-1})}) + \sup_{theta\in Theta} \log(\Pi_{i=1}^{m} \frac{p_\theta(z^i)}{p_\theta)z^{i-1})}) \\
&= -\log \frac{\bar{p}(z^m)}{\bar{p}(z^0)} + \sup_{\theta\in\Theta} \log \frac{p_\theta(z^m)}{p_\theta(z^0)} \\
&= -\log \bar{p}(z^m) + \sup_{\theta\in\Theta} \log p_\theta(z^m) \\
&= -\log(\frac{1}{N} \sum_{\theta\in\Theta} p_\theta(z^m)) + \log p_{\hat{\theta}}(z^m) \\
&= -\log \frac{1}{N} - \log \sum_{\theta\in\Theta} p_\theta(z^m) + \log p_{\hat{\theta}}(z^m) \\
&\leq -\log \frac{1}{N} - \log p_{\hat{\theta}}(z^m) + \log p_{\hat{\theta}}(z^m) \\
&= \log N
\end{aligned}$$

Subquestion 1.3

L et us fix $N = |\Theta|$ and set $\Theta = \{1/(N+1), ..., N/(N+1)\}$. For example, if we set $N = 4$ then $\Theta = \{0.2, 0.4, 0.6, 0.8\}$. Use the law of large numbers to argue that

$$\lim_{m\to\infty} S_m = \log N.$$

Solution

Once again we will follow the hint to write $S_m = \log \sum_{\theta' \in \Theta} \sum_{z^m : \hat{\theta}(z^m) = \theta'} P_{\theta'}(z^m)$. We obtain this by looking at the definition and then note that

$$\begin{aligned} S_m &:= \log \sum_{z^m \in \{0,1\}^m} p_{\hat{\theta}(z^m)}(z^m) \\ &= \log \sum_{\theta' \in \Theta} \sum_{z^m : \hat{\theta}(z^m) = \theta'} p_{\hat{\theta}(z^m)}(z^m) \\ &= \log \sum_{\theta' \in \Theta} \sum_{z^m : \hat{\theta}(z^m) = \theta'} p_{\theta'(z^m)}(z^m) \end{aligned}$$

Now we can rewrite S_m in a useful way:

$$\begin{aligned} S_m &:= \log \sum_{z^m \in \{0,1\}^m} p_{\hat{\theta}(z^m)}(z^m) \\ &= \log \sum_{\theta' \in \Theta} \sum_{z^m : \hat{\theta}(z^m) = \theta'} p_{\theta'(z^m)}(z^m) \\ &= \log \sum_{\theta' \in \{1/N+1, \ldots, N/N+1\}} \sum_{z^m : th\hat{e}ta(z^m) = \theta'} p_{\theta'}(z^m) \\ &= \log \sum_{i=1}^{N} \sum_{z^m : \hat{\theta}(z^m) = \frac{i}{N+1}} p_{\frac{i}{N+1}}(z^m) \end{aligned}$$

But now for $m \to \infty$ we know by the law of large numbers that the MLE will be almost surely be the same for every z^m (in case of an odd N) and therefore the part $\sum_{z^m : \hat{\theta}(z^m) = \frac{i}{N+1}} p_{\frac{i}{N+1}}(z^m)$ will almost surely be equal to 1. Now by the continuity of the logarithm and the finiteness of the summation over N terms, we obtain

$$\lim_{m \to \infty} S_m \to \log \sum_{i=1}^{N} 1 = \log N.$$

We now still need to consider the case for which N is an even number. Then we have the two values in Θ which are closest to 0.5 also are equally far away from 0.5. Then half of the z^m will be determined under the smaller MLE $\hat{\theta}_L$ and the other half under the larger MLE $\hat{\theta}_R$. But then we find

$$\sum_{z^m : \hat{\theta}_L(z^m) = \frac{i}{N+1}} p_{\frac{i}{N+1}}(z^m) = \frac{1}{2}$$

and also

$$\sum_{z^m : \hat{\theta}_R(z^m) = \frac{i}{N+1}} p_{\frac{i}{N+1}}(z^m) = \frac{1}{2},$$

such that we obtain the same result.

Subquestion 1.4

Informally explain why, for small sample sizes m, the Shtarkov sum for $\Theta = \{0.47, 0.49, 0.51, 0.53\}$ is significantly smaller than the Shtarkov sum for $\Theta = \{0.2, 0.4, 0.6, 0.8\}$.

Solution

For relatively small values of m, it is more likely to find values for the MLE that are further away from 0.5. Now when the possible values of θ are still farily close to 0.5, such as in $\Theta = \{0.47, 0.49, 0.51, 0.53\}$, we obtain as a contribution of $p_{\hat{\theta}(z^m)}(z^m)$ bigger values than when the possible values of θ are further away from 0.5. (Compare for instance when finding for $m = 10$, z^m having two ones and eight zeros. We find $(0.2)^2(0.8)^8 > (0.47)^2(0.53)^8$ since $(0.2)^2(0.8)^8 \approx 0.0067$ and $(0.47)^2(0.53)^8 \approx 0.0014$.) In the previous question, these cases would not occur almost surely, but for small m, this of course does not hold.

Subquestion 1.5

Suppose that P consists of a finite number of black-box experts, which given each history Z_1, Z_{i-1} provide us a distribution on Z_i, which may depend on the past — we don't know how they come up with their predictions, we just observe the predictions they make on the sample. Explain why we'd rather use the Bayesian than the Shtarkov predictor in such a setting.

Solution

If we repeatedly apply the Bayesian predictor and supply it with the last guess as the prior distribution, it will adapt to fit the distributions better in the case that Z_i depends on the other Z's. The Shtarkov does not do this and will therefore be worse in case of dependence. If the distributions are independent, we only have equal regret bounds of $\log N$, so we do not have an indication which one to use and therefore the Bayesian predictor can be chosen just as well as the Shtarkoc predictor. So without any risk, the Bayesian predictor may attain better results, which is when Z_i is dependent of $Z_1, \ldots, Z_{i-1}$.

Question 2

(Uncountable Θ, [4+1/3 pt]) Now consider the full Bernoulli model, $\Theta = [0, 1]$, with the sum in (1) repladed by an integral, for the uniform prior probability densiwty $w(\theta) := 1$ for all $\theta \in [0, 1]$.

Subquestion 2.1

For each of the follwoing statements, indicate wheter it is true or false, and prove your answer.

1. [3 + 1/3*pt*] The NML predictor p^* achieves the same regret for every sequence $x^m - x_1, \ldots, x_m \in \{0,1\}^m$.

Solution

True.

$$p_n^*(x^n) = \frac{\sup_{\theta\in\Theta} \theta_n(x^n)}{\sum_{z^n\in Z^n} \sup_{\theta\in\Theta} \theta_n(z^n)}$$

$$\begin{aligned}\hat{L}(x^n) - \inf_{\theta\in\Theta} L_\theta(x^n) &= \log \frac{\sup_{\theta\in\Theta} \theta_n(x^n)}{p_n^*(x^n)} \\ &= \log \sum_{z^n\in Z^n} \sup_{\theta\in\Theta} \theta_n(z^n) \\ &\Rightarrow \text{independant of } x^n \\ &\Rightarrow \text{regret is the same for } \forall x^n\end{aligned}$$

2. same as (i) with 'regret' replaced by 'cumulative loss'.

Solution

False.

$$p_n^*(x^n) = \frac{\sup_{\theta\in\Theta} \theta_n(x^n)}{\sum_{z^n\in Z^n} \sup_{\theta\in\Theta} \theta_n(z^n)}$$

$$\begin{aligned}\hat{L}(x^n) &= \log \frac{1}{p_n^*(x^n)} \\ &= \log \frac{\sum_{z^n\in Z^n} \sup_{\theta\in\Theta} \theta_n(z^n)}{\sup_{\theta\in\Theta} \theta_n(x^n)} \\ &\Rightarrow \text{dependant on } x^n \\ &\Rightarrow \text{cumulative loss is not the same for } \forall x^n\end{aligned}$$

3. the Bayesian predictor $\bar{p}$ achieves the same regret for every sequence $x^m \in \{0,1\}^m$.

Solution

False.

$$\bar{p}_n(x^n) = \int w(\theta) p_\theta(x^n)\, d\theta = \int p_\theta(x^n)\, d\theta$$

$$\begin{aligned}\hat{L}(x^n) - \inf_{\theta\in\Theta} L_\theta(x^n) &= \log \frac{\sup_{\theta\in\Theta} \theta_n(x^n)}{\bar{p}_n(x^n)} \\ &= \log \frac{\sup_{\theta\in\Theta} \theta_n(x^n)}{\int p_\theta(x^n)\, d\theta} \\ &\Rightarrow \text{dependant on } x^n \\ &\Rightarrow \text{regret is the not the same for } \forall x^n\end{aligned}$$

4. same as (iii) with 'regret' replaced by 'cumulative loss'.

Solution

False.

$$\bar{p}_n(x^n) = \int w(\theta) p_\theta(x^n)\, d\theta = \int p_\theta(x^n)\, d\theta$$

$$\begin{aligned}\hat{L}(x^n) &= \log \frac{1}{\bar{p}_n(x^n)} \\ &= \log \frac{1}{\int p_\theta(x^n)\, d\theta} \\ &\Rightarrow \text{dependant on } x^n \\ &\Rightarrow \text{not the same for } \forall x^n\end{aligned}$$

5. for every fixed $x^m \in \{0,1\}^m$, the NML predictor p^* achieves the same regret on every y^m that is a permutation of x^m.

Solution

True.
Since the regret is the same for every x^n, it is also the same for the permutation of x^n.

6. same as (vii) with 'regret' replaced by 'cumulative loss'.

Solution

True.

$$\begin{aligned}\hat{L}(x^n) &= \log \frac{1}{p_n^*(x^n)} \\ &= \log \frac{\sum_{z^n\in Z^n} \sup_{\theta\in\Theta} \theta_n(z^n)}{\sup_{\theta\in\Theta} \theta_n(x^n)} \\ &\Rightarrow \sup_{\theta\in\Theta} \theta_n(x^n) \text{ is independant on the order of the elements of } x^n \\ &\Rightarrow \text{cumulative loss is the same for every permutation of } x^n\end{aligned}$$

7. for every fixed $x^m \in \{0,1\}^m$, every $n < m$, the NML predictor p^* achieves the same loss on $y_1, \ldots, y_n$, for every $y_1, \ldots, y_n$ that are the initial segment of an $y_1, \ldots, y_m$ that is a permutation of x^m.

   Solution

   False.

$$\hat{L}(y^n) = \log \frac{\sum_{z^n \in Z^n} \sup_{\theta \in \Theta} \theta_n(z^n)}{\sup_{\theta \in \Theta} \theta_n(y^n)}$$

   $\sup_{\theta \in \Theta} \theta_n(y^n)$ can be different for every y^n, therefore the loss is not the same for every y^n.

8. same as (ix) but with p^* replaced by $\bar{p}$.

   Solution

   False.

$$\hat{L}(y^n) = \log \frac{1}{\int p_\theta(y^n)\, d\theta}$$

   $\int p_\theta(y^n)\, d\theta$ can be different for every y^n, therefore the loss is not the same for every y^n.

Subquestion 2.2

[1 pt] Suppose we do not know in advance how many predictions we have to make, i.e. we do not know the horizon n. Explain why this is a problem for prediction with the NML p^* but not for prediction with $\bar{p}$.

Solution

Prediction with NML p^* contains a part $\sum_{z^n \in Z^n} \sup_{\theta \in \Theta} \theta_n(z^n)$ which is hard to compute.