Appendix: Classification by thresholding

In this appendix, we show how the bounds given in the first section of this monograph can be computed in practice on a simple example: the case when the classification is performed by comparing a series of measurements to threshold values. Let us mention that our description covers the case when the same measurement is compared to several thresholds, since it is enough to repeat a measurement in the list of measurements describing a pattern to cover this case.

5.1. Description of the model

Let us assume that the patterns we want to classify are described through h real valued measurements normalized in the range $(0,1)$. In this setting the pattern space can thus be defined as $\mathcal{X} = (0,1)^h$.

Consider the threshold set $\mathcal{T} = (0,1)^h$ and the response set $\mathcal{R} = \mathcal{Y}^{\{0,1\}^h}$. For any $t \in (0,1)^h$ and any $a : \{0,1\}^h \to \mathcal{Y}$, let

$$f_{(t,a)}(x) = a\Big\{ \big[\mathbb{1}(x^j \geq t_j) \big]_{j=1}^h \Big\}, \quad x \in \mathcal{X},$$

where x^j is the jth coordinate of $x \in \mathcal{X}$. Thus our parameter set here is $\Theta = \mathcal{T} \times \mathcal{R}$. Let us consider the Lebesgue measure L on $\mathcal{T}$ and the uniform probability distribution U on $\mathcal{R}$. Let our prior distribution be $\pi = L \otimes U$. Let us define for any threshold sequence $t \in \mathcal{T}$

$$\Delta_t = \Big\{ t' \in \mathcal{T} : \overline{(t'_j, t_j)} \cap \{X_i^j ; i = 1, \dots, N\} = \varnothing, j = 1, \dots, h \Big\},$$

where X_i^j is the jth coordinate of the sample pattern X_i, and where the interval $\overline{(t'_j, t_j)}$ of the real line is defined as the convex hull of the two point set $\{t'_j, t_j\}$, whether $t'_j \leq t_j$ or not. We see that Δ_t is the set of thresholds giving the same response as t on the training patterns. Let us consider for any $t \in \mathcal{T}$ the middle

$$m(\Delta_t) = \frac{\int_{\Delta_t} t' L(dt')}{L(\Delta_t)}$$

of Δ_t. The set Δ_t being a product of intervals, its middle is the point whose coordinates are the middle of these intervals. Let us introduce the finite set T composed of the middles of the cells Δ_t, which can be defined as

$$T = \{ t \in \mathcal{T} : t = m(\Delta_t) \}.$$

It is easy to see that $|T| \leq (N+1)^h$ and that $|\mathcal{R}| = |\mathcal{Y}|^{2^h}$.

5.2. Computation of inductive bounds

For any parameter $(t,a) \in \mathcal{T} \times \mathcal{R} = \Theta$, let us consider the posterior distribution defined by its density

$$\frac{d\rho_{(t,a)}}{d\pi}(t',a') = \frac{\mathbb{1}\bigl(t' \in \Delta_t\bigr)\mathbb{1}\bigl(a' = a\bigr)}{\pi\bigl(\Delta_t \times \{a\}\bigr)}.$$

In fact we are considering a finite number of posterior distributions, since $\rho_{(t,a)} = \rho_{(m(\Delta_t),a)}$, where $m(\Delta_t) \in T$. Moreover, for any exchangeable sample distribution $\mathbb{P} \in \mathcal{M}_+^1\bigl[(\mathcal{X} \times \mathcal{Y})^{N+1}\bigr]$ and any thresholds $t \in \mathcal{T}$,

$$\mathbb{P}\Bigl[\,\overline{(X_{N+1}^j, t_j)} \cap \{X_i^j, i = 1, \dots, N\} = \varnothing\Bigr] \leq \frac{2}{N+1}.$$

Thus, for any $(t,a) \in \Theta$,

$$\mathbb{P}\Bigl\{\rho_{(t,a)}\bigl[f_{\cdot}(X_{N+1})\bigr] \neq f_{(t,a)}(X_{N+1})\Bigr\} \leq \frac{2h}{N+1},$$

showing that the classification produced by $\rho_{(t,a)}$ on new examples is typically non-random; this result is only indicative, since it is concerned with a non-random choice of (t,a).

Let us compute the various quantities needed to apply the results of the first section, focussing our attention on Theorem 2.1.3 (page 54).

First note that $\rho_{(t,a)}(r) = r[(t,a)]$. The entropy term is such that

$$\mathcal{K}(\rho_{t,a}, \pi) = -\log\bigl[\pi\bigl(\Delta_t \times \{r\}\bigr)\bigr] = -\log\bigl[L(\Delta_t)\bigr] + 2^h \log\bigl(|\mathcal{Y}|\bigr).$$

Let us notice accordingly that

$$\min_{(t,a)\in\Theta} \mathcal{K}(\rho_{(t,a)}, \pi) \leq h \log(N+1) + 2^h \log\bigl(|\mathcal{Y}|\bigr).$$

Let us introduce the counters

$$b_y^t(c) = \frac{1}{N}\sum_{i=1}^N \mathbb{1}\Bigl\{Y_i = y \text{ and } \bigl[\mathbb{1}(X_i^j \geq t_j)\bigr]_{j=1}^h = c\Bigr\},$$
$$t \in T, c \in \{0,1\}^h, y \in \mathcal{Y},$$

$$b^t(c) = \sum_{y\in\mathcal{Y}} b_y^t(c) = \frac{1}{N}\sum_{i=1}^N \mathbb{1}\Bigl\{\bigl[\mathbb{1}(X_i^j \geq t_j)\bigr]_{j=1}^h = c\Bigr\}, \qquad t \in T, c \in \{0,1\}^h.$$

Since

$$r[(t,a)] = \sum_{c\in\{0,1\}^h} \bigl[b^t(c) - b_{a(c)}^t(c)\bigr],$$

the partition function of the Gibbs estimator can be computed as

$$\begin{aligned}\pi\bigl[\exp(-\lambda r)\bigr] &= \sum_{t\in T} L(\Delta_t) \sum_{a\in\mathcal{R}} \frac{1}{|\mathcal{Y}|^{2^h}} \exp\Biggl[-\lambda \sum_{i=1}^N \mathbb{1}\bigl[Y_i \neq f_{(t,a)}(X_i)\bigr]\Biggr] \\ &= \sum_{t\in T} L(\Delta_t) \sum_{a\in\mathcal{R}} \frac{1}{|\mathcal{Y}|^{2^h}} \exp\Biggl[-\lambda \sum_{c\in\{0,1\}^h} \bigl[b^t(c) - b_{a(c)}^t(c)\bigr]\Biggr]\end{aligned}$$

$$= \sum_{t \in T} L(\Delta_t) \prod_{c \in \{0,1\}^h} \left[\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \exp\Bigl(-\lambda \bigl[b^t(c) - b^t_y(c) \bigr] \Bigr) \right].$$

We see that the number of operations needed to compute $\pi\bigl[\exp(-\lambda r)\bigr]$ is proportional to $|T| \times 2^h \times |\mathcal{Y}| \leq (N+1)^h 2^h |\mathcal{Y}|$. An exact computation will therefore be feasible only for small values of N and h. For higher values, a Monte Carlo approximation of this sum will have to be performed instead.

If we want to compute the bound provided by Theorem 2.1.3 (page 54) or by Theorem 2.2.2 (page 68), we need also to compute, for any fixed parameter $\theta \in \Theta$, quantities of the type

$$\pi_{\exp(-\lambda r)} \Bigl\{ \exp\bigl[\xi m'(\cdot, \theta) \bigr] \Bigr\} = \pi_{\exp(-\lambda r)} \Bigl\{ \exp\bigl[\xi \rho_\theta(m') \bigr] \Bigr\}, \quad \lambda, \xi \in \mathbb{R}_+.$$

We need to introduce

$$\overline{b}^t_y(\theta, c) = \frac{1}{N} \sum_{i=1}^N \Bigl| \mathbb{1}\bigl[f_\theta(X_i) \neq Y_i \bigr] - \mathbb{1}(y \neq Y_i) \Bigr| \mathbb{1}\bigl\{ \bigl[\mathbb{1}(X_i^j \geq t_j) \bigr]_{j=1}^h = c \bigr\}.$$

Similarly to what has been done previously, we obtain

$$\pi\bigl\{ \exp\bigl[-\lambda r + \xi m'(\cdot, \theta) \bigr] \bigr\} \\ = \sum_{t \in T} L(\Delta_t) \prod_{c \in \{0,1\}^h} \left[\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \exp\Bigl(-\lambda \bigl[b^t(c) - b^t_y(c) \bigr] + \xi \overline{b}^t_y(\theta, c) \Bigr) \right].$$

We can then compute

$$\pi_{\exp(-\lambda r)}(r) = -\frac{\partial}{\partial \lambda} \log\bigl\{ \pi\bigl[\exp(-\lambda r) \bigr] \bigr\},$$

$$\pi_{\exp(-\lambda r)} \Bigl\{ \exp\bigl[\xi \rho_\theta(m') \bigr] \Bigr\} = \frac{\pi\bigl\{ \exp\bigl[-\lambda r + \xi m'(\cdot, \theta) \bigr] \bigr\}}{\pi\bigl[\exp(-\lambda r) \bigr]},$$

$$\pi_{\exp(-\lambda r)} \bigl[m'(\cdot, \theta) \bigr] = \frac{\partial}{\partial \xi}_{|\xi = 0} \log\Bigl[\pi\bigl\{ \exp\bigl[-\lambda r + \xi m'(\cdot, \theta) \bigr] \bigr\} \Bigr].$$

This is all we need to compute $B(\rho_\theta, \beta, \gamma)$ (and also $B(\pi_{\exp(-\lambda r)}, \beta, \gamma)$) in Theorem 2.1.3 (page 54), using the approximation

$$\log\Bigl\{ \pi_{\exp(-\lambda_1 r)} \Bigl[\exp\bigl\{ \xi \pi_{\exp(-\lambda_2 r)}(m') \bigr\} \Bigr] \Bigr\} \\ \leq \log\Bigl\{ \pi_{\exp(-\lambda_1 r)} \Bigl[\exp\bigl\{ \xi m'(\cdot, \theta) \bigr\} \Bigr] \Bigr\} + \xi \pi_{\exp(-\lambda_2 r)} \bigl[m'(\cdot, \theta) \bigr], \quad \xi \geq 0.$$

Let us also explain how to apply the posterior distribution $\rho_{(t,a)}$, in other words our randomized estimated classification rule, to a new pattern X_{N+1}:

$$\rho_{(t,a)} \bigl[f_\cdot(X_{N+1}) = y \bigr] = L(\Delta_t)^{-1} \int_{\Delta_t} \mathbb{1}\Bigl[a\bigl\{ \bigl[\mathbb{1}(X^j_{N+1} \geq t'_j) \bigr]_{j=1}^h \bigr\} = y \Bigr] L(dt') \\ = L(\Delta_t)^{-1} \sum_{c \in \{0,1\}^h} L\Bigl(\Bigl\{ t' \in \Delta_t : \bigl[\mathbb{1}(X^j_{N+1} \geq t'_j) \bigr]_{j=1}^h = c \Bigr\} \Bigr) \mathbb{1}\bigl[a(c) = y \bigr].$$

Let us define for short

$$\Delta_t(c) = \Bigl\{ t' \in \Delta_t : \bigl[\mathbb{1}(X^j_{N+1} \geq t'_j) \bigr]_{j=1}^h = c \Bigr\}, \qquad c \in \{0,1\}^h.$$

With this notation

$$\rho_{(t,a)}\bigl[f_{\cdot}(X_{N+1})=y\bigr]=L\bigl(\Delta_t\bigr)^{-1}\sum_{c\in\{0,1\}^h}L\bigl[\Delta_t(c)\bigr]\mathbb{1}\bigl[a(c)=y\bigr].$$

We can compute in the same way the probabilities for the label of the new pattern under the Gibbs posterior distribution:

$$\begin{aligned}&\pi_{\exp(-\lambda r)}\bigl[f_{\cdot}(X_{N+1})=y'\bigr]\\&=\Biggl\{\sum_{t\in T}\prod_{c\in\{0,1\}^h}\biggl[\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\exp\Bigl(-\lambda\bigl[b^t(c)-b^t_y(c)\bigr]\Bigr)\biggr]\\&\qquad\times\sum_{c\in\{0,1\}^h}L\bigl[\Delta_t(c)\bigr]\frac{\sum_{y\in\mathcal{Y}}\mathbb{1}(y=y')\exp\bigl\{-\lambda\bigl[b^t(c)-b^t_y(c)\bigr]\bigr\}}{\sum_{y\in\mathcal{Y}}\exp\bigl\{-\lambda\bigl[b^t(x)-b^t_y(c)\bigr]\bigr\}}\Biggr\}\\&\qquad\qquad\times\Biggl\{\sum_{t\in T}L(\Delta_t)\prod_{c\in\{0,1\}^h}\biggl[\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\exp\Bigl(-\lambda\bigl[b^t(c)-b^t_y(c)\bigr]\Bigr)\biggr]\Biggr\}^{-1}.\end{aligned}$$

5.3. Transductive bounds

In the case when we observe the patterns of a shadow sample $(X_i)_{i=N+1}^{(k+1)N}$ on top of the training sample $(X_i,Y_i)_{i=1}^N$, we can introduce the set of thresholds responding as t on the extended sample $(X_i)_{i=1}^{(k+1)N}$

$$\overline{\Delta}_t=\Bigl\{t'\in\mathcal{T}:\overline{(t'_j,t_j)}\cap\bigl\{X_i^j;i=1,\dots,(k+1)N\bigr\}=\varnothing,j=1,\dots,h\Bigr\},$$

consider the set

$$\overline{T}=\bigl\{t\in\mathcal{T}:t=m(\overline{\Delta}_t)\bigr\},$$

of the middle points of the cells $\overline{\Delta}_t$, $t\in\mathcal{T}$, and replace the Lebesgue measure $L\in\mathcal{M}^1_+\bigl[(0,1)^h\bigr]$ of the previous section with the uniform probability measure $\overline{L}$ on $\overline{T}$. We can then consider $\pi=\overline{L}\otimes U$, where U is as previously the uniform probability measure on $\mathcal{R}$. This gives obviously an exchangeable posterior distribution and therefore qualifies π for transductive bounds. Let us notice that $|\overline{T}|\le\bigl[(k+1)N+1\bigr]^h$, and therefore that $\pi(t,a)\ge\bigl[(k+1)N+1\bigr]^{-h}|\mathcal{Y}|^{-2^h}$, for any $(t,a)\in\overline{T}\times\mathcal{R}$.

For any $(t,a)\in\mathcal{T}\times\mathcal{R}$ we may similarly to the inductive case consider the posterior distribution $\rho_{(t,a)}$ defined by

$$\frac{d\rho_{(t,a)}}{d\pi}(t',a')=\frac{\mathbb{1}(t'\in\Delta_t)\mathbb{1}(a'=a)}{\pi\bigl(\Delta_t\times\{a\}\bigr)},$$

but we may also consider $\delta_{(m(\overline{\Delta}_t),a)}$, which is such that $r_i\{[m(\overline{\Delta}_t),a]\}=r_i[(t,a)]$, $i=1,2$, whereas only $\rho_{(t,a)}(r_1)=r_1[(t,a)]$, while

$$\rho_{(t,a)}(r_2)=\frac{1}{|\overline{T}\cap\Delta_t|}\sum_{t'\in\overline{T}\cap\Delta_t}r_2[(t',a)].$$

We get

$$\begin{aligned}\mathcal{K}(\rho_{(t,a)}, \pi) &= -\log\bigl[\overline{L}(\Delta_t)\bigr] + 2^h \log\bigl(|\mathcal{Y}|\bigr)\\ &\leq \log\bigl(|\overline{T}|\bigr) + 2^h \log\bigl(|\mathcal{Y}|\bigr) = \mathcal{K}(\delta_{[m(\overline{\Delta}_t),a]}, \pi)\\ &\leq h \log\bigl[(k+1)N + 1\bigr] + 2^h \log\bigl(|\mathcal{Y}|\bigr),\end{aligned}$$

whereas we had no such uniform bound in the inductive case. Similarly to the inductive case

$$\pi\bigl[\exp(-\lambda r_1)\bigr] = \sum_{t \in T} \overline{L}(\Delta_t) \prod_{c \in \{0,1\}^h} \left[\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \exp\Bigl(-\lambda\bigl[b^t(c) - b^t_y(c)\bigr]\Bigr)\right].$$

Moreover, for any $\theta \in \Theta$,

$$\begin{aligned}\pi\bigl\{\exp\bigl[-\lambda r_1 + \xi \rho_\theta(m')\bigr]\bigr\} &= \pi\bigl\{\exp\bigl[-\lambda r_1 + \xi m'(\cdot, \theta)\bigr]\bigr\}\\ &= \sum_{t \in T} \overline{L}(\Delta_t) \prod_{c \in \{0,1\}^h} \left[\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \exp\Bigl(-\lambda\bigl[b^t(c) - b^t_y(c)\bigr] + \xi \bar{b}(\theta, c)\Bigr)\right].\end{aligned}$$

The bound for the transductive counterpart to Theorems 2.1.3 (page 54) or 2.2.2 (page 68), obtained as explained page 115, can be computed as in the inductive case, from these two partition functions and the above entropy computation.

Let us mention finally that, using the same notation as in the inductive case,

$$\begin{aligned}&\pi_{\exp(-\lambda r_1)}\bigl[f_\cdot(X_{N+1}) = y'\bigr]\\ &= \Biggl\{\sum_{t \in T} \prod_{c \in \{0,1\}^h} \left[\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \exp\Bigl(-\lambda\bigl[b^t(c) - b^t_y(c)\bigr]\Bigr)\right]\\ &\qquad \times \sum_{c \in \{0,1\}^h} \overline{L}\bigl[\Delta_t(c)\bigr] \frac{\sum_{y \in \mathcal{Y}} \mathbb{1}(y = y') \exp\bigl\{-\lambda\bigl[b^t(c) - b^t_y(c)\bigr]\bigr\}}{\sum_{y \in \mathcal{Y}} \exp\bigl\{-\lambda\bigl[b^t(x) - b^t_y(c)\bigr]\bigr\}}\Biggr\}\\ &\qquad\qquad \times \left\{\sum_{t \in T} \overline{L}(\Delta_t) \prod_{c \in \{0,1\}^h} \left[\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \exp\Bigl(-\lambda\bigl[b^t(c) - b^t_y(c)\bigr]\Bigr)\right]\right\}^{-1}.\end{aligned}$$

To conclude this appendix on classification by thresholding, note that similar factorized computations are feasible in the important case of *classification trees.* This can be achieved using some variant of the *context tree weighting method* discovered by Willems et al. (1995) and successfully used in lossless compression theory. The interested reader can find a description of this algorithm applied to classification trees in Catoni (2004, page 62).