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 $t \in \mathcal{T}$

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

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 $|\mathcal{R}| = |\mathcal{Y}|^{2^h}$.