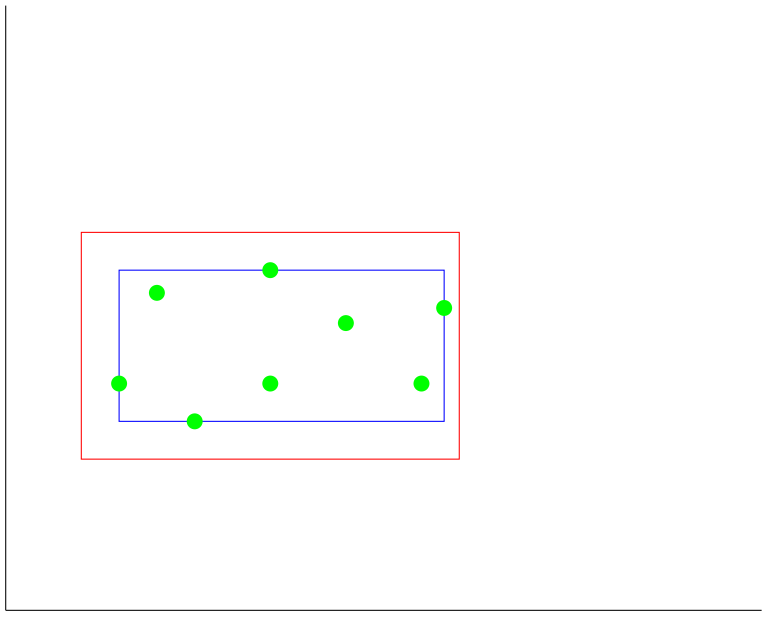
Rectangles

Prakash Panangaden

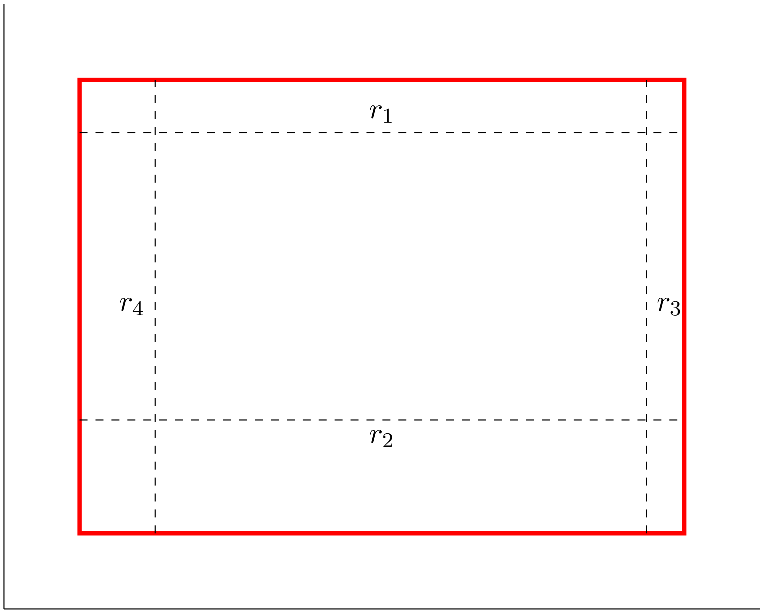
13th September 2020

Our input space is the positive quadrant of the plane and the hypotheses are closed rectangles aligned with the axes.



The green dots show the points in the sample. The red rectangle is the actual concept we want to learn and the blue rectangle is what we construct by using our rule of the tightest rectangle that contains all the sample points. The error region is between the blue and red rectangles.

For a measurable subset E we write $\mathsf{Pr}(E)$ for the probability that a sample point lies in E according to the distribution D that we are assuming. Now let us look closely at the big rectangle.



Let the big red rectangle be called R and the blue rectangle (not shown in the second picture) is $\hat{R}$. We are interested in finding out the probability δ, that the hypothesis $\hat{R}$ makes mistakes with probability more than ε given that we had m samples. We assume that $\mathsf{Pr}(R) \geq \varepsilon$ otherwise of course $\delta = 0$. We define 4 strips near the edges, labelled r_1, r_2, r_3, r_4 in the picture. These strips are defined as follows.

Let $R = [a, b] \times [c, d]$. We define r_4 to be the *smallest* rectangle such that $\mathsf{Pr}(r_4) \geq \varepsilon/4$. More precisely, $r_4 = [a, l] \times [c, d]$ where

$$l = \inf \{x | \mathsf{Pr}([a, x] \times [c, d]) \geq \varepsilon/4\}.$$

I claim that $\mathsf{Pr}([a, l) \times [c, d]) \leq \varepsilon/4$. To see this, note that by the definition of l we know that

$$\forall \gamma \in (0, l - a] \mathsf{Pr}([a, l - \gamma] \times [c, d]) < \varepsilon/4.$$

Now choose a countable sequence γ_n in $(0, l - a)$ converging to 0 from above. We have

$$\bigcup_n [a, l - \gamma_n] \times [c, d] = [a, l) \times [c, d], \text{ where } [a, l - \gamma_n] \times [c, d] \subset [a, l - \gamma_{n+1}] \times [c, d].$$

Hence by σ-additivity we have that the measure is continuous on countable nested families of sets so

$$\mathsf{Pr}([a,l) \times [c,d]) = \lim_{n \longrightarrow \infty} \mathsf{Pr}([a, l - \gamma_n] \times [c,d]) \leq \varepsilon/4.$$

I will call the strip $[a,l) \times [c,d]$ r'_4. Similarly for the other 3 strips.

Now if $\hat{R}$ meets (has non-empty intersection with) all the r_i (here I really mean r_i) then it has *all* its edges in these regions. So the set $R \setminus \hat{R}$ is contained in $\bigcup_{i=1,\ldots,4} r'_i$ (here I really mean r'_i) and the error region for this hypothesis cannot be more than ε. Contrapositively, if $\mathrm{err}(\hat{R}) > \varepsilon$ then $\hat{R}$ must miss one of the regions r_i completely (and here I mean r_i).

Thus we calculate

$[\mathsf{Pr}(\mathrm{err}(\hat{R})) > \varepsilon] \leq \mathsf{Pr}(\bigvee_i \hat{R} \cap r_i = \emptyset)$

$$\begin{aligned} &\leq \sum_{i=1}^{4} \mathsf{Pr}(\hat{R} \cap r_i = \emptyset) && \text{by the union bound} \\ &\leq 4(1 - \varepsilon/4)^m && \text{since } \mathsf{Pr}(r_i) \geq \varepsilon/4 \\ &\leq 4\exp(-m\varepsilon/4) && \text{since } 1 - x \leq e^{-x}. \end{aligned}$$

Thus, to ensure that $\mathsf{Pr}(\mathrm{err}(\hat{R}) > \varepsilon) \leq \delta$ we require that $\exp(-m\varepsilon/4) \leq (\frac{\delta}{4})$ or, writing it as a sample bound, we can say that to have error rate less than ε with probability at least $1 - \delta$ we should have

$$m \geq \frac{4}{\varepsilon} \log_e(\frac{4}{\delta}).$$