An Introduction to Computational Learning Theory

Michael J. Kearns
Umesh V. Vazirani

The MIT Press
Cambridge, Massachusetts
London, England

1

The Probably Approximately Correct Learning Model

1.1 A Rectangle Learning Game

Consider a simple one-player learning game. The object of the game is to learn an unknown axis-aligned rectangle R — that is, a rectangle in the Euclidean plane $\Re^2$ whose sides are parallel with the coordinate axes. We shall call R the **target** rectangle. The player receives information about R only through the following process: every so often, a random point p is chosen in the plane according to some fixed probability distribution $\mathcal{D}$. The player is given the point p together with a label indicating whether p is contained in R (a positive example) or not contained in R (a negative example). Figure 1.1 shows the unknown rectangular region R along with a sample of positive and negative examples.

The goal of the player is to use as few examples as possible, and as little computation as possible, to pick a **hypothesis** rectangle R' which is a close approximation to R. Informally, the player's knowledge of R is tested by picking a new point at random from the same probability distribution $\mathcal{D}$, and checking whether the player can correctly decide whether the point falls inside or outside of R. Formally, we measure the

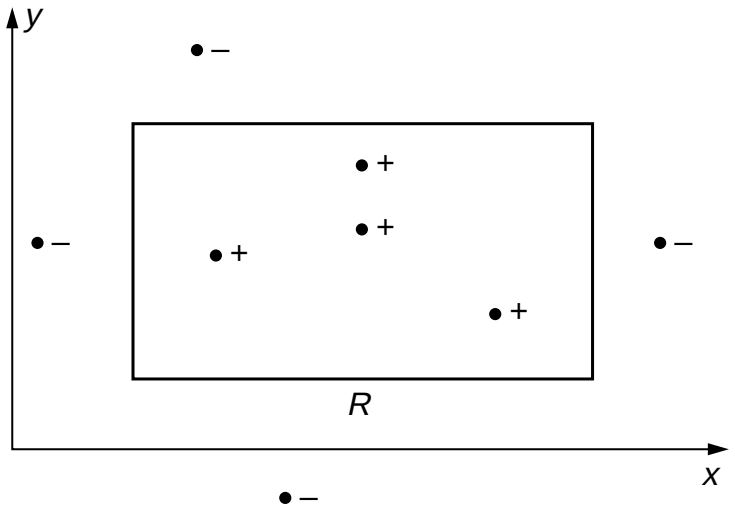

Figure 1.1: *The target rectangle R in the plane along with a sample of positive and negative examples.*

error of R' as the probability that a randomly chosen point from $\mathcal{D}$ falls in the region $R \Delta R'$, where $R \Delta R' = (R - R') \cup (R' - R)$.

To motivate the rectangle learning game, consider a slightly more concrete scenario that can be expressed as an instance of the game. Suppose that we wanted to learn the concept of "men of medium build". Assume that a man is of medium build if his height and weight both lie in some prescribed ranges — for instance, if his height is between five feet six inches and six feet, and his weight is between 150 pounds and 200 pounds. Then each man's build can be represented by a point in the Euclidean plane, and the concept of medium build is represented by an axis-aligned rectangular region of the plane. Thus, during an initial training phase, the learner is told for each new man he meets whether that man is of medium build or not. Over this period, the learner must form some model or hypothesis of the concept of medium build.

Now assume that the learner encounters every man in his city with

equal probability. Even under this assumption, the corresponding points in the plane may not be uniformly distributed (since not all heights and weights are equally likely, and height and weight may be highly dependent quantities), but will instead obey some fixed distribution $\mathcal{D}$ which may be quite difficult to characterize. For this reason, in our learning game, we allow the distribution $\mathcal{D}$ to be arbitrary, but we assume that it is fixed, and that each example is drawn independently from this distribution. (Note that once we allow $\mathcal{D}$ to be arbitrary, we no longer need to assume that the learner encounters every man in his city with equal probability.) To evaluate the hypothesis of the learner, we are simply evaluating its success in classifying the build of men in future encounters, still assuming that men are encountered according to the same probability distribution as during the training phase.

There is a simple and efficient strategy for the player of the rectangle learning game. The strategy is to request a "sufficiently large" number m of random examples, then choose as the hypothesis the axis-aligned rectangle R' which gives the tightest fit to the positive examples (that is, that rectangle with the smallest area that includes all of the positive examples and none of the negative examples). If no positive examples are drawn, then $R' = \emptyset$. Figure 1.2 shows the tightest-fit rectangle defined by the sample shown in Figure 1.1.

We will now show that for any target rectangle R and any distribution $\mathcal{D}$, and for any small values ϵ and δ $(0 < \epsilon, \delta \leq 1/2)$, for a suitably chosen value of the sample size m we can assert that with probability at least $1 - \delta$, the tightest-fit rectangle has error at most ϵ with respect to R and $\mathcal{D}$.

First observe that the tightest-fit rectangle R' is always contained in the target rectangle R (that is, $R' \subseteq R$ and so $R \Delta R' = R - R'$). We can express the difference $R - R'$ as the union of four rectangular strips. For instance, the topmost of these strips, which is shaded and denoted T' in Figure 1.3, is the region above the upper boundary of R' extended to the left and right, but below the upper boundary of R. Note that there is some overlap between these four rectangular strips at the corners. Now

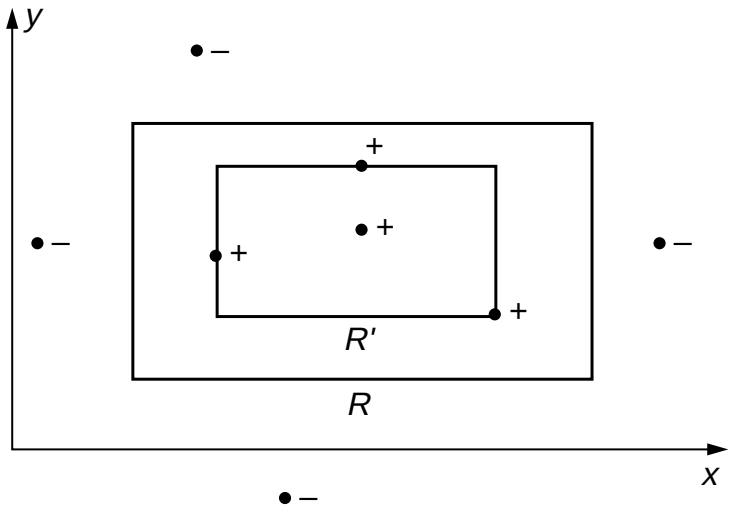

Figure 1.2: *The tightest-fit rectangle R' defined by the sample.*

if we can guarantee that the weight under $\mathcal{D}$ of each strip (that is, the probability with respect to $\mathcal{D}$ of falling in the strip) is at most $\epsilon/4$, then we can conclude that the error of R' is at most $4(\epsilon/4) = \epsilon$. (Here we have erred on the side of pessimism by counting each overlap region twice.)

Let us analyze the weight of the top strip T'. Define T to be the rectangular strip along the inside top of R which encloses *exactly* weight $\epsilon/4$ under $\mathcal{D}$ (thus, we sweep the top edge of R downwards until we have swept out weight $\epsilon/4$; see Figure 1.3). Clearly, T' has weight exceeding $\epsilon/4$ under $\mathcal{D}$ if and only if T' includes T (which it does not in Figure 1.3). Furthermore, T' includes T if and only if no point in T appears in the sample S — since if S does contain a point $p \in T$, this point has a positive label since it is contained in R, and then by definition of the tightest fit, the hypothesis rectangle R' must extend upwards into T to cover p.

By the definition of T, the probability that a single draw from the distribution $\mathcal{D}$ misses the region T is exactly $1 - \epsilon/4$. Therefore the

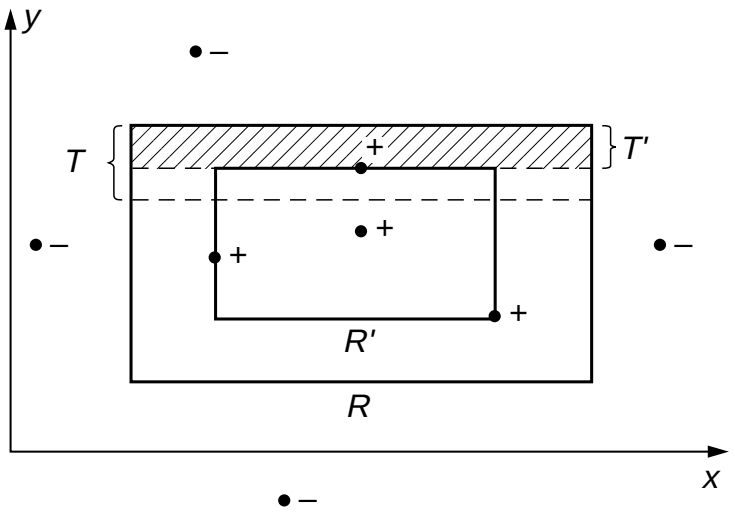

Figure 1.3: *Analysis of the error contributed by the top shaded strip T'. The strip T has weight exactly $\epsilon/4$ under $\mathcal{D}$.*

probability that m independent draws from $\mathcal{D}$ all miss the region T is exactly $(1-\epsilon/4)^m$. Here we are using the fact that the probability of a conjunction of independent events is simply the product of the probabilities of the individual events. The same analysis holds for the other three rectangular regions of $R - R'$, so by the **union bound**, the probability that any of the four strips of $R - R'$ has weight greater than $\epsilon/4$ is at most $4(1-\epsilon/4)^m$. By the union bound, we mean the fact that if A and B are any two events (that is, subsets of a probability space), then

$$\mathbf{Pr}[A \cup B] \leq \mathbf{Pr}[A] + \mathbf{Pr}[B].$$

Thus, the probability that one of the four error strips has weight exceeding $\epsilon/4$ is at most four times the probability that a fixed error strip has weight exceeding $\epsilon/4$.

Provided that we choose m to satisfy $4(1-\epsilon/4)^m \leq \delta$, then with probability $1-\delta$ over the m random examples, the weight of the error

region $R - R'$ will be bounded by ϵ, as claimed. Using the inequality

$$(1 - x) \leq e^{-x}$$

(which we shall appeal to frequently in our studies) we see that any value of m satisfying $4e^{-\epsilon m/4} \leq \delta$ also satisfies the previous condition. Dividing by 4 and taking natural logarithms of both sides gives $-\epsilon m/4 \leq \ln(\delta/4)$, or equivalently $m \geq (4/\epsilon)\ln(4/\delta)$.

In summary, provided our tightest-fit algorithm takes a sample of at least $(4/\epsilon)\ln(4/\delta)$ examples to form its hypothesis rectangle R', we can assert that with probability at least $1 - \delta$, R' will misclassify a new point (drawn according to the same distribution from which the sample was chosen) with probability at most ϵ.

A few brief comments are appropriate. First, note that the analysis really does hold for any fixed probability distribution. We only needed the independence of successive points to obtain our bound. Second, the sample size bound behaves as we might expect, in that as we increase our demands on the hypothesis rectangle — that is, as we ask for greater **accuracy** by decreasing ϵ or greater **confidence** by decreasing δ — our algorithm requires more examples to meet those demands. Finally, the algorithm we have analyzed is efficient: the required sample size is a slowly growing function of $1/\epsilon$ and $1/\delta$ (linear and logarithmic, respectively), and once the sample is given, the computation of the tightest-fit hypothesis can be carried out rapidly.

1.2 A General Model

In this section, we introduce the model of learning that will be the central object for most of our study: the **Probably Approximately Correct** or **PAC** model of learning. There are a number of features of the rectangle learning game and its solution that are essential to the PAC model, and bear highlighting before we dive into the general definitions.