Kolmogorov’s axioms of probability (1933) provided a suitable base for the theory of probability. In 1965 Kolmogorov proposed what was, in his opinion, a suitable base for the applications of probability: the algorithmic notion of randomness. Unfortunately, the latter is non-computable; in this note we describe a recent attempt to find practicable approximations to it.

Randomness

We are interested in the following standard problem of machine learning: given a training set \((x_1, y_1), \ldots, (x_l, y_l)\), consisting of \(l\) labelled examples \((x_i, y_i)\), and a new unlabelled example \(x_{l+1}\), we would like to say something about its label \(y_{l+1}\). We assume that the examples \((x_i, y_i)\) are produced from the same distribution independently of each other (the iid model); for simplicity, we will also assume that the labels \(y_i\) can only take two values, \(-1\) or \(1\).

The idea of using algorithmic randomness for predicting \(y_{l+1}\) is as follows: for every possible value \(y \in \{-1, 1\}\) for \(y_{l+1}\) we estimate the “randomness” (or “typicalness”) of the sequence \((x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, y)\) with respect to the iid model; we can make a confident prediction if and only if exactly one of these two sequences is typical.

The “universal” notion of randomness was defined by Kolmogorov, Martin-Löf and Levin. For finite sequences there is no clear-cut difference between typical and atypical sequences (all sequences are typical but to a different degree), so one needs “randomness deficiency” (see Li and Vitanyi, 1997) or “randomness level”. The randomness level of a sequence is a number between 0 and 1; when it is close to 0, the sequence is atypical. Traditionally, the minus logarithm of the randomness level (the randomness deficiency) is more widely used. There are two main definitions of randomness deficiency, which we call \(p\)-randomness (proposed by Martin-Löf) and \(i\)-randomness (proposed by Levin); the former is in the spirit of sampling theory and frequentist probability theory, and the latter is more in the spirit of likelihood theory and Bayesian statistics. The \(p\) and \(i\)-notions are close to each other, but the difference between them is important.

Imagine that we can compute the randomness level; let \(R_y\) be the randomness level of \((x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, y), y \in \{-1, 1\}\). Our prediction will be the \(y = \hat{y}\) with the largest \(R_y\); the confidence in this prediction will be \(1 - \max_{y \neq \hat{y}} R_y\); and the credibility of this prediction will be \(R_{\hat{y}}\).

Most of all we are interested in typicalness wr to the iid model and wr to the weaker exchangeability model. The former is the assumption we are willing to make, and the latter is the assumption we actually need for our efficient approximations (see below) to be valid. It can be shown that

\[
d_{\text{iid}}(z) = d_{\text{exch}}(z \mid d_{\text{exch}}(\text{conf}(z))) + d_{\text{exch}}(\text{conf}(z)),
\]

where \(d_{\text{iid}}\) and \(d_{\text{exch}}\) are the iid and exchangeability randomness deficiencies (in the sense of \(i\)-definitions), respectively, and \(\text{conf}(z)\) is the bag containing all elements of \(z\). This result is
closely connected with de Finetti’s theorem.

**Practical approximations**

As we already mentioned, the randomness level is non-computable. A powerful way to approximate it is to use Vapnik’s (1998) Support Vector machine. With every possible label $y \in \{-1, 1\}$ for $x_{l+1}$ we associate an optimization problem (see, eg, Vapnik, 1998, (10.10)– (10.12)); the solutions $\alpha_1, \ldots, \alpha_{l+1}$ to this problem reflect the “strangeness” of the examples ($\alpha_i$ being the strangeness of $(x_i, y_i)$, $i = 1, \ldots, l$, and $\alpha_{l+1}$ being the strangeness of $(x_{l+1}, y)$); all $\alpha_i$ are non-negative and, in practice, only few of them are different from zero (the corresponding examples are called *support vectors*). An easily computable approximation to the randomness level $R_y$ (wr to either iid or exchangeability models) is given by $\frac{\#\{i=1, \ldots, l+1: \alpha_i \geq \alpha_{l+1}\}}{l+1}$.

**Discussion**

In this note we only discussed the problem of pattern recognition (in Vapnik’s, 1998, terminology). The ideas of the algorithmic theory of randomness can be applied, in a similar way, to the problem of regression (in which the labels are allowed to take values in the whole real line). For the third fundamental problem studied by Vapnik, that of density estimation, confident predictions in high-dimensional cases (say, when all $x_i$ are different) are impossible.

Kolmogorov’s approach to randomness is closely connected with the notion of Kolmogorov complexity (see Li and Vitanyi, 1997). Both Kolmogorov complexity and randomness can be generalized to a wide range of important “games of prediction”: see (Vovk, 1999) and the CLRC web page given below.

The work on efficient approximations to the universal notion of randomness is currently under way at Computer Learning Research Centre, Royal Holloway, University of London; for further information, see its web site [http://www.clrc.rhbnc.ac.uk](http://www.clrc.rhbnc.ac.uk).

**REFERENCES**


**RÉSUMÉ**

Les axiomes de probabilité (1933) de Kolmogorov fournissent la base appropriée pour la théorie de probabilité. En 1965 Kolmogorov proposa ce que fut, en son opinion, une base appropriée pour les applications de probabilité: la notion algorithmique de séquence aléatoire. Malheureusement, le dernier est non-computable ; dans cette note nous décrivons une tentative récente de trouver approximations pratiques à cela.