Harman, G. and Kulkarni, S. **Reliable Reasoning: Induction and Statistical Learning Theory.** The MIT Press, London, England, 2007. x + 108 pp, \$30.00/£18.95, ISBN 9780262083607.

Inferring generalizations from data is one of the most interdisciplinary topics, studied of course in statistics, but also in branches of computer science, engineering, philosophy, and economics. The authors of “Reliable Reasoning” are two Princeton researchers from different disciplines Kalurkani (Electrical Engineering) and Harman (Philosophy). The book introduces learning theory, a mathematical framework for analyzing induction, with many applications in computer science and engineering. The main topic is *binary classification*: Consider an independent variable x ranging over a set X—the instances—and a classifier function l: $X \rightarrow \{0,1\}$ that labels instances as positive or negative. In typical applications, the argument x is a feature vector that describes an instance. For example, a linear separator is specified by a set of weights $w(i),..,w(n)$; it classfies a feature vector x as positive iff $\Sigma\, x(i)\, w(i) > 0$. Given a random sample of labelled instances, the problem is to infer the labeling function l that generates the data. This is an important problem with many applications, such as predicting whether an applicant will repay her loan, whether a patient will respond well to treatment, or whether an e-mail message is spam. Many other important inductive problems are not treated in the book; the advantage of the strict focus is a presentation that attains depth with little technicality.

The book focuses on an approach to classification problems pioneered by Vapnik and Chervonenkis [VC 1971]; in computer science, the subject is known as PAC learning [Valiant 1984]. The VC learning model is as follows. Suppose we fix a classifier space $\mathcal{H}$, a sampling distribution μ over the instance space X, and a classifier l from $\mathcal{H}$. The learner is presented with a random sample of correctly labelled points $<(x_1,l_1),\ldots,(x_k,l_k)>$, distributed according to μ. The error of a hypothesis l is the probability of misclassification according to the sampling distribution μ. The sample error of l is the error of l evaluated on the sample distribution. The weak law of large numbers implies that the sample error converges to the true error in probability. VC theory asks under what circumstances the convergence is *uniform* over the classifier space $\mathcal{H}$. In symbols, the goal is as follows: given a desired accuracy ε in (0,1/2) and confidence parameter δ in (0,1/2), find a bound $m_{\mathcal{H}}(\varepsilon,\delta)$, depending only on the desired accuracy and confidence, such that: for *every* sample distribution μ, for any classifier l in $\mathcal{H}$, for any sample of size at least $m_{\mathcal{H}}(\varepsilon,\delta)$, the probability is at least $1\text{-}\delta$ that the sample error provides an ε-close estimate of the true error.

Vapnik and Chernovenkis proved the deep result that uniform consistency depends on a finite combinatorial complexity measure called the VC dimension, defined as follows. Consider a set of instances $\{x_1,..,x_k\}$. If for each of the possible 2^k labellings of the instance set, there is some classifer l from $\mathcal{H}$ that agrees with that labeling on the k instances, the class $\mathcal{H}$ is said to *shatter* the instances $\{x_1,..,x_k\}$. The VC dimension of a classifier space $\mathcal{H}$ is the maximum size k such that some instance set of size k is shattered by $\mathcal{H}$. The VC theorem says that a hypothesis class $\mathcal{H}$ admits a uniform bound on the convergence of the empirical error to the true error if and only if $\mathcal{H}$ has a finite VC dimension.

Shattering and the VC dimension can be determined visually for the space of linear separations of two-dimensional points. Three collinear points cannot be shattered as no line separates both of the outer two points fom the interior one. If three points are not collinear, they form a triangle, and any one vertex of the triangle can be linearly separated from the other two. So the VC dimension is at least 3. No set of four points can be shattered: In the first case, one of the points is inside a triangle formed by the other three. Then no line separates the inside point from all the other three. In the second case, there are pairs (x_1,x_2) and (y_1,y_2) such that the two lines connecting x_1,x_2 and y_1,y_2 meet in the middle of the four points. No linear separation labels x_1 and x_2 as positive and y_1 and y_2 as negative. The fact that no set of four points can be shattered means that the VC dimension of the class of linear separators in two dimensions is at most 3. In general, the VC dimension of linear separators in n dimensions is n+1 [H & K, p.48].

About half of "Reliable Reasoning" introduces conceptual background, motivation, and examples for the VC dimension approach. The other half discusses extensions, alternatives, and applications. For example, an extension known as structural risk minimization proposes using the VC dimension as a parametrization-invariant definition of model complexity. The book compares structural risk minimization with traditional criteria such as the number of parameters in a model. Among the applications discussed in the book are support vector machines (SVM). SVMs use linear separators for classification, but in an extended instance space with extra dimensions added to the feature vectors. SVMs have been extensively applied in machine learning as their accuracy typically tops other classification learners [Taylor 2004].

"Reliable Reasoning" is a good introduction to statistical thinking in a focused setting. It presents several sophisticated mathematical ideas with a minimum of notation and a maximum of intuition. A reader with a background in statistics can read this book in a day and will learn much about how learning theory approaches fundamental issues in inferring classification rules.

References

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

V.N. Vapnik and A.Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264-280, 1971.

L.G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134-1142, 1984.

Oliver Schulte

School of Computing Science
Simon Fraser University
Burnaby, B.C., Canada,