Groundwork for a new Approach to Knowledge Discovery:
Certainty upon Empirical Distributions

Joan Garriga

Dptmnt. de Llenguatges i Sistemes Informàtics,
Universitat Politècnica de Catalunya,
jgarriga@lsi.upc.edu
http://www.lsi.upc.edu/~jgarriga/

Abstract. We address the problem of assessing the information conveyed by a finite discrete probability distribution, within the context of knowledge discovery. Our approach is based on two main axiomatic intuitions: (i) the minimum information is given in the case of a uniform distribution, and (ii) knowledge is akin to a notion of richness, related to the dimension of the distribution. From this perspective, we define a statistic that has a clear interpretation in terms of a *measure of certainty*, and we build up a plausible hypothesis, that offers a comprehensible insight of knowledge, with a consistent algebraic structure. This includes a native value for the uncertainty related to unseen events. Our contributions are then faced up with entropy based measures. Finally, by implementing our measure in a decision tree induction algorithm, we show an empirical validation of the behavior of our measure with respect to entropy. Our conclusion is that the contributions of our measure are significant, and should lead to more robust models.

Key words: knowledge discovery, measures of information, entropy