Simple Bayesian Classifiers Do Not Assume Independence

Pedro Domingos* **Michael Pazzani**

Department of Information and Computer Science
University of California, Irvine
Irvine, California 92717, U.S.A.
{pedrod, pazzani}@ics.uci.edu
http://www.ics.uci.edu/~pedrod

Bayes' theorem tells us how to optimally predict the class of a previously unseen example, given a training sample. The chosen class should be the one which maximizes $P(C_i|E) = P(C_i)\,P(E|C_i)\,/\,P(E)$, where C_i is the ith class, E is the test example, $P(Y|X)$ denotes the conditional probability of Y given X, and probabilities are estimated from the training sample. Let an example be a vector of a attributes. If the attributes are *independent* given the class, $P(E|C_i)$ can be decomposed into the product $P(v_1|C_i)\ldots P(v_a|C_i)$, where v_j is the value of the jth attribute in the example E. Therefore we should predict the class that maximizes:

$$P(C_i|E) = \frac{P(C_i)}{P(E)} \prod_{j=1}^{a} P(v_j|C_i) \qquad (1)$$

This procedure is often called the *naive Bayesian classifier*. Here we will prefer the term *simple*, and abbreviate to *SBC*. The SBC is commonly thought to be optimal, in the sense of achieving the best possible accuracy, only when the "independence assumption" above holds, and perhaps close to optimal when the attributes are only slightly dependent. However, this very restrictive condition seems to be contradicted by the SBC's surprisingly good performance in a wide variety of domains, including many where there are clear dependencies between the attributes. In a study on 28 datasets from the UCI repository, we found the SBC to be more accurate than C4.5 in 16 domains, and similarly for CN2 and PEBLS. Other authors have made similar observations, but no interpretation has been proposed so far. Here we shed some light on the matter by showing that the SBC is in fact optimal even when the independence assumption is grossly violated, and thus applicable to a far broader range of domains than previously thought.

The key to this result lies in the distinction between classification and probability estimation. Equation 1 yields a correct estimate of the class probabilities only when the independence assumption holds; but for purposes of classification, the class probability estimates can diverge widely from the true values, as long as the maximum estimate still corresponds to the maximum true probability. For example, suppose there are two classes + and −, and let $P(+|E) = 0.51$ and $P(-|E) = 0.49$ be the true class probabilities given example E. The optimal decision is then to assign E to class +. Suppose also that Equation 1 gives the estimates $\hat{P}(+|E) = 0.99$ and $\hat{P}(-|E) = 0.01$. The independence assumption is violated by a wide margin, and yet the SBC still makes the optimal decision.

Consider the general two-class case. Let the classes be + and −, $p = P(+|E)$, $r = \frac{P(+)}{P(E)} \prod_{j=1}^{a} P(v_j|+)$, and $s = \frac{P(-)}{P(E)} \prod_{j=1}^{a} P(v_j|-)$. The SBC is optimal iff:

$$(p \geq \tfrac{1}{2} \wedge r \geq s) \vee (p < \tfrac{1}{2} \wedge r < s) \qquad (2)$$

The space U of values of (p, r, s) that correspond to valid probability combinations is a subspace of the unit cube $[0, 1]^3$, and its projection on all planes $p = k$ is the same. It is easily shown that Condition 2 holds in exactly *half* the total volume of U. In contrast, by the independence assumption the SBC would be optimal only on the line where the planes $r = p$ and $s = 1-p$ intersect. Thus the previously assumed region of optimality of the SBC is a second-order infinitesimal fraction of the actual one.

The SBC will be the optimal classifier in the entire example space iff Condition 2 holds for every possible combination of attribute values. For this reason, the fraction of all possible concepts on a attributes for which the SBC is optimal everywhere decreases exponentially with a, starting at 100% for $a = 1$. However, a similar statement is true for other learners, given a fixed training set size.

Testing Condition 2 directly for all combinations of values will generally be infeasible; see (Domingos & Pazzani 1996) for a number of more easily tested conditions. In summary, the work reported here demonstrates that the SBC has a far greater range of applicability than previously thought, and suggests that its use should be considered more often.

References

Domingos, P., and Pazzani, M. 1996. Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. In *Proceedings of the Thirteenth International Conference on Machine Learning.* Bari, Italy: Morgan Kaufmann. Forthcoming.

*Partly supported by a PRAXIS XXI scholarship.