Bayesian Learning

Peter J. Denning

6 Mar 89

RIACS Technical Report TR-89.12

NASA Cooperative Agreement Number NCC 2-387

RIACS

Research Institute for Advanced Computer Science

Bayesian Learning

Peter J. Denning

6 Mar 89

RIACS Technical Report TR-89.12

NASA Cooperative Agreement Number NCC 2-387

Bayesian Learning

Peter J. Denning

Research Institute for Advanced Computer Science
NASA Ames Research Center

RIACS Technical Report TR-89.12
6 Mar 89

IN 1983 and 1984, the Infrared Astronomical Satellite (IRAS) detected 5,425 stellar objects and measured their infrared spectra. In 1987 a program called AUTOCLASS used Bayesian inference methods to discover the classes present in these data and determine the most probable class of each object, revealing unknown phenomena in astronomy. AUTOCLASS has rekindled the old debate on the suitability of Bayesian methods, which are computationally intensive, interpret probabilities as plausibility measures rather than frequencies, and appear to depend on a subjective assessment of the probability of a hypothesis before the data were collected. Modern statistical methods have, however, recently been shown to also depend on subjective elements. These debates bring into question the whole tradition of scientific objectivity and offer scientists a new way to take responsibility for their findings and conclusions.

This is a preprint of the column *The Science of Computing* for
American Scientist 77, No. 3 (May-June 1989).

Work reported herein was supported in part by Cooperative Agreement NCC 2-387
between the National Aeronautics and Space Administration (NASA)
and the Universities Space Research Association (USRA).

Bayesian Learning

Peter J. Denning

Research Institute for Advanced Computer Science

6 Mar 89

In 1983, NASA launched the Infrared Astronomical Satellite (IRAS), a joint project of the United States, the Netherlands, and the United Kingdom. One of the instruments aboard IRAS scanned the skies for objects emitting in the infrared band from 7 to 23 microns and measured their spectra. These infrared wavelengths are not observable from earth because the atmosphere absorbs them and emits its own thermal radiation in this band. IRAS performed the first such survey of the skies.

To eliminate any possibility that thermal radiation from the telescope or the detectors might obliterate already-faint signals, the whole apparatus was cooled in liquid helium (to about 2°K). It functioned for a year until the helium supply was exhausted, scanning 96% of the sky before it ceased operation. For each stellar object detected, IRAS recorded two celestial coordinates and 94 spectral intensities at preselected wavelengths. The resulting 5,425 records make up what is now known as the IRAS low-resolution spectral database.

IRAS had been programmed to begin observations with the star Vega, which was to be the calibrator for the instrument, but unexpected excess energy at the longer wavelengths caused the selection of another star, Alpha Lyra, for this role. The anomaly was quickly interpreted as evidence of a dust disk and a possible planetary system around Vega, a discovery that received immediate attention in the media.

Over the next two years, the IRAS records were examined and grouped into classes already known to astronomers. No new classes were invented to explain the records. But this assay was inadequate to deal with a large database of complicated objects about which little was previously known. A few astronomers were keenly interested in whether any of the automatic learning systems under study by artificial intelligence (AI) researchers might help them understand the data better.

In 1987, Peter Cheeseman of the Research Institute for Advanced Computer Science, working with colleagues from the AI branch at the NASA Ames Research Center, completed a program called AUTOCLASS, which was designed for automatic classification of records in very large databases with many attributes *(1)*. AUTOCLASS calculates the most probable number of classes, the most probable parameters for each, and the most probable class of each record. It was well suited to the task of sifting through the IRAS data, and the IRAS data presented a challenging test case for Cheeseman's design. AUTOCLASS discovered new classes that differ significantly from those used for the earlier analysis and clearly represent unknown previously physical phenomena.

Some observers have suggested that, because AUTOCLASS is a product of AI research, it should be listed as one of the authors of the papers reporting the new

discoveries in astronomy. AUTOCLASS, however, deals only with statistics of the numbers in the database; its computation incorporates no information about astronomy. It is meaningless to say that a program having no knowledge of the field can make a discovery. It makes more sense to say that the program detects statistical patterns that humans interpret as discoveries in astronomy. The same can be said about other AI programs purported to have made discoveries in other disciplines.

AUTOCLASS is based on a principle of inference first enunciated by Thomas Bayes in 1763. Suppose that D is a set of data and $H_1, \ldots, H_n$ are distinct hypotheses. The law of conditional probability says that

$$p(D) = \sum_{i=1}^{n} p(D \mid H_i) p(H_i) . \tag{1}$$

Bayes's theorem says that the probability any one of the conditions, say H_k, occurs given D is the proportion of $p(D)$ contributed by the kth term:

$$p(H_k \mid D) = \frac{p(H_k) p(D \mid H_k)}{p(D)} . \tag{2}$$

This theorem is often stated in the form, "The posterior probability of the hypothesis given the data is proportional to the product of the prior probability of the hypothesis and the likelihood of the data given the hypothesis," where $1/p(D)$ is the constant of proportionality. Bayes's theorem shows how to calculate a backward inference, sometimes called "reversed conditioning" or "inverse probability."

Now suppose that we interpret the H_i as possible models (hypotheses) that explain given experimental data D. Given any model, one can calculate the likelihood that the data will be observed in that model, $p(D \mid H_i)$. If one also has a value for the prior

probability of each model, $p(H_i)$, one can use Bayes's theorem to calculate the probability of each model given the data. It is then reasonable to say that the "best" model is the most probable one according to this calculation. This is called Bayesian inference. In his book, Larry Bretthorst treats many estimation methods based on this principle *(2)*.

In this approach, probabilities are interpreted not as frequencies observable through experiments, but as degrees of plausibility one assigns to each hypothesis based on the data and on one's assessment of the plausibility of the hypotheses prior to seeing the data. The idea that probabilities must stand for something observable (i.e., frequencies) is closely related to deeply-rooted beliefs in Western scientific tradition. I will return to this point later.

According to Edwin Jaynes, Bayes's arguments about inverse probability were nearly incomprehensible *(3)*. But Laplace rediscovered the principle in 1774 and, for the next 40 years, applied it with great clarity to problems of astronomy, geodesy, meteorology, population statistics, and even jurisprudence. Jaynes cites the story of Laplace's estimate of the mass of Saturn to illustrate the power of the method. Using data on the mutual perturbations of Jupiter and Saturn, Laplace estimated that Saturn's mass is 1/3512 of the solar mass and gave a probability of 0.99991 that the true mass is within 1% of this estimate. The modern value for Saturn's mass is put at 0.63% higher, near the upper end of Laplace's range.

The AUTOCLASS program applies Bayesian inference to determine the most probable classification of given data. In the case of the IRAS data, it assumes that the spectral intensity at each wavelength is accounted for by a normal distribution whose

parameters come from one of N classes. Each class has a vector of parameters, the means and variances for each of the 94 intensities making up a spectrum. To specify a hypothesis, we need to state a value of N, a vector of 94 means and variances for each class, and a probability that each of the 5,425 records belongs to each particular class -- just over 10^6 numbers in all. Constructing a sample of several million hypotheses of this type from the astronomically large space of all possible hypotheses, and then applying Bayes's theorem to find the best, would far exceed the processing capacity of any existing supercomputer. Instead of an enumeration, AUTOCLASS uses a search procedure to modify a current hypothesis iteratively and obtain a maximum of equation (2). The search procedure contains extra steps to attempt escape from local maxima. When it completes its search, AUTOCLASS has constructed a locally most likely hypothesis that explains the data. AUTOCLASS takes about 36 hours on a Symbolics computer to process the IRAS data.

Even though it incorporates many approximations, AUTOCLASS has performed remarkably well on real databases and has outperformed other methods such as cluster analysis. In the IRAS database, it found classes agreeing with those determined previously by astronomers, and it produced several new discoveries. It found the previously known classes in databases of iris plants, soybean diseases, and horse colic cases *(1)*. When applied to data artificially generated from known class distributions, it found the parameters of those distributions. When applied to random data from a single distribution it found only one class, as it should.

Rather than being welcomed within the AI research community as a promising new approach to machine learning, the principles of Bayesian inference underlying the

AUTOCLASS program have been the subject of spirited debate. Should logic or probability be the basis of inference? A recent issue of a major journal published a lengthy debate on this question between Cheeseman and 23 of his critics *(4)*. Cheeseman goes well beyond asserting that Bayesian methods look promising in practice; he argues that probabilistic inference is fundamental to human reasoning. Having reminded his readers that he interprets probabilities as plausibility measures rather than relative frequencies, he notes a proof by R. T. Cox in 1946 that any system of plausibility measures that assigns a real number to each proposition and the same real number to logically equivalent propositions must satisfy the axioms of probability theory. According to Cheeseman, conditional probabilities behave the way people's beliefs do; new information can either increase or decrease one's belief in a proposition, and different people assign different plausibilities to the same proposition.

Cheeseman's critics say that Bayesian inference is far more demanding computationally than deductive logic, suggesting that a computation barrier will prevent wide application of Bayesian methods. They say that one can use "possibility theory" or "fuzzy logic," two systems that do not obey the sum and product rules of probability, to construct computationally feasible augmentations of deductive logic with plausibility measures. They also say that, to use Bayes's theorem, one must already have the prior probability of the hypothesis, $p(H)$, which can only be a subjective assessment before any data are taken; they say that this lack of objectivity offends the fundamental traditions of science.

The debate among AI researchers mirrors a much older debate among statisticians and philosophers about Bayesian inference. That debate also focuses on the feasibility of

the calculations and the apparent violation of scientific objectivity. As we have seen, modern high-speed computers are rapidly undermining the argument about computational feasibility, leaving objectivity as the main unresolved issue.

James Berger and Donald Berry have recently questioned the supposed objectivity of modern statistical methods. *(5)*. Using these methods, one searches for experimental outcomes that cast doubt on the negation of the desired hypothesis; the aggregate probability among all possible outcomes that cast at least as much doubt as the observed data is called the significance level. The significance level depends on the intended experimental procedure: two experiments with identical outcomes but different designs will have different significance levels. This happens because the significance level depends on data that might have been observed but were not, and that depends on the experimental design. An observer of the two experimenters would see no difference between their actions and, on hearing them declare different significance levels, would conclude that there were subjective elements (things known only to each experimenter) in the conclusions. Bayesian inference also has a subjective element -- the prior probability of the hypothesis -- but, say Berger and Berry, this is explicitly in the hands of the consumer rather than the producer of the results. They maintain that no known statistical inference method produces conclusions free of subjective influence.

That the scientific method is fundamentally objective is a notion deeply rooted in Western tradition. Philosopher and historian James Burke questions this notion, documenting the history of the idea that there are objective realities independent of human observation *(6)*. His examination of science through the ages shows that contemporary theories influence the types of investigations people undertake and that the

"structure of interpretation" within a given society affects which hypotheses and data are admissible. Each age believes its science is objective, and the next age refutes this. Following the example of our forebears, we also say, "Science is objective," and yet doubts about objective reality persist.

What if what we call objective reality is simply an interpretation of data agreed to by large numbers of people? What if the contribution of science is a way of defining "standard observers" that always produce the same data no matter who follows its rules? Burke says that such interpretations would actually lend power to science. By adopting them, we would acknowledge that much of what we call objectivity is an illusion created by agreements on standard observers. We would be able to accept responsibility for the influence of our own prejudices, biases, and interpretations on the results of our experiments.

References

1. P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D Freeman. 1988. "AUTOCLASS: A Bayesian Classification System." In *Proc. Fifth Machine Learning Workshop*. Morgan-Kaufmann. 54-64.

2. G. L. Bretthorst. 1988. *Bayesian Spectrum Analysis and Parameter Estimation.* Lecture Notes in Statistics. Springer-Verlag.

3. E. T. Jaynes. 1986. "Bayesian Methods: General Background." In *Maximum entropy and Bayesian Methods in Applied Statistics* (J. H. Justice, ed.), Cambridge University Press. 1-25.

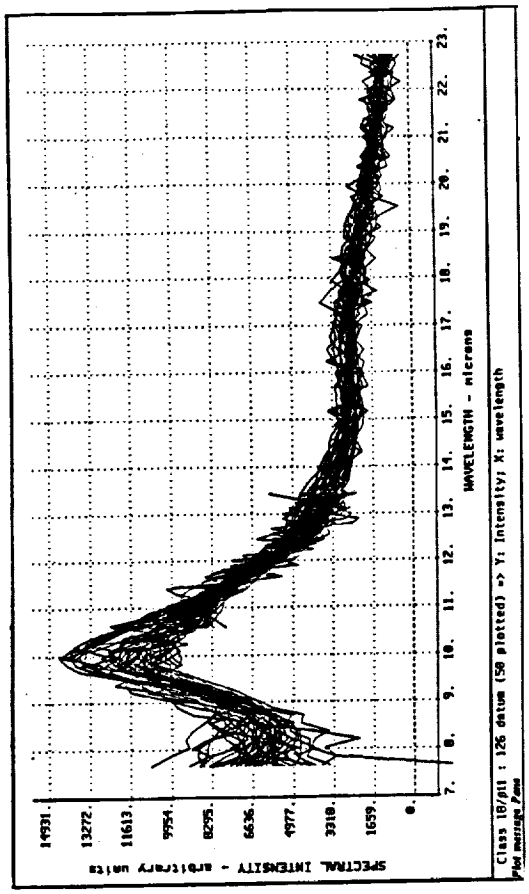
4. M. McLeish, ed. 1988. *Computer Intelligence* **4**. *Special issue on an inquiry into computer understanding*.

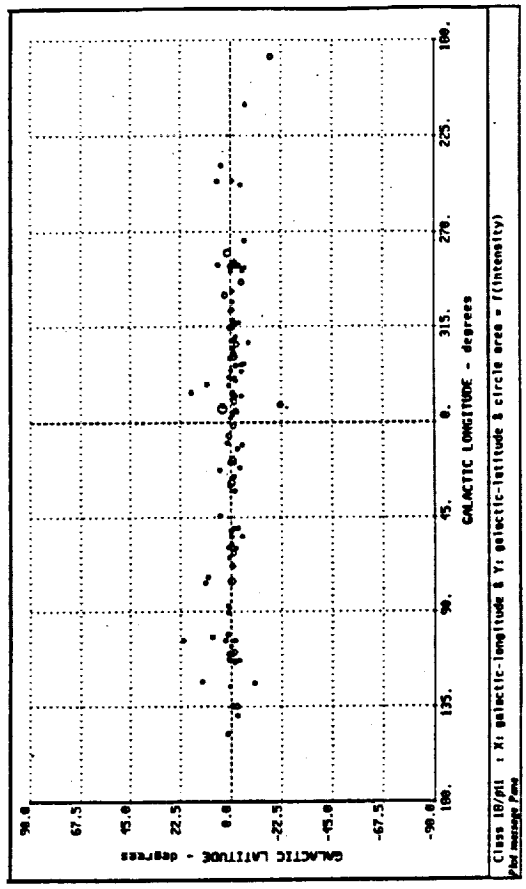
5. J. O. Berger and D. A. Berry. 1988. "Statistical analysis and the illusion of objectivity." *American Scientist* **76**, 2 (March- April), 159-165.

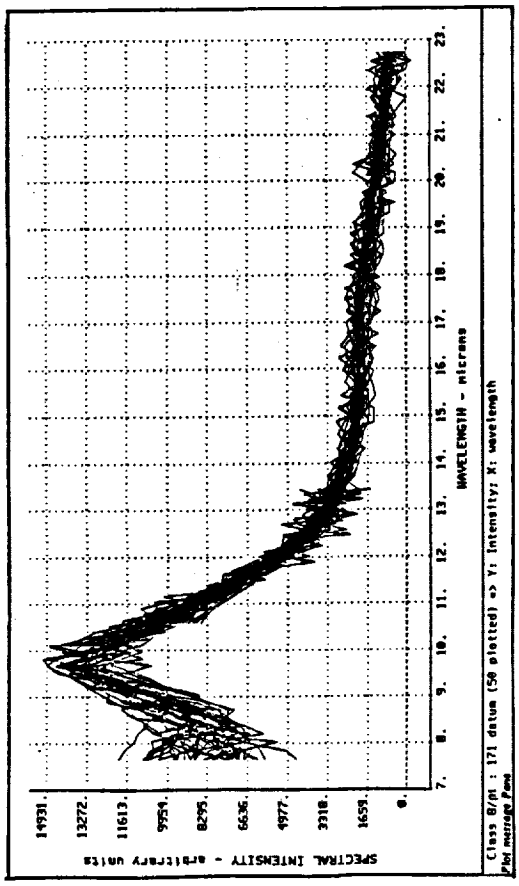
6. J. Burke. 1985. *The day the universe changed*. Little, Brown.

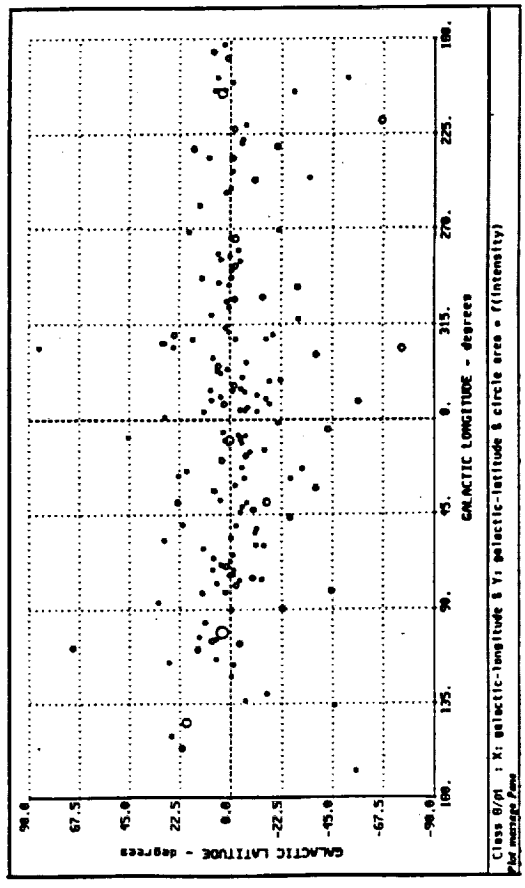
AUTOCLASS discoveries

In 1983 and 1984, the Infrared Astronomical Satellite (IRAS) detected 5,425 stellar objects and measured their infrared spectra. A program called AUTOCLASS used Bayesian inference methods to discover the classes present in the data and determine the most probable class of each object. It discovered some classes that were significantly different from those previously known to astronomers. One such discovery is illustrated in the accompanying picture. Previous analysis had identified a set of 297 objects with strong silicate spectra. AUTOCLASS partitioned this set into two parts (*top*). The class on the left (171 objects) has a peak at 9.7 microns and the class on the right (126 objects) a peak at 10.0 microns. When the objects are plotted on a star map by their celestial coordinates (*bottom*), the right set shows a marked tendency to cluster around the galactic plane, confirming that the classification represents real differences between the classes of objects. AUTOCLASS did not use the celestial coordinates in its estimates of classes. Astronomers are studying the phenomenon further to determine the cause.









ORIGINAL PAGE IS
OF POOR QUALITY