

Introduction to Bayesian Statistics - 7

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

A few more applications of Bayesian methods

- The incidence of lung cancer
- Bayes classifiers
- The case of the AUTOCLASS unsupervised classifier
- The nature of learning in Bayesian and MaxEnt methods

Cornfield, Jerome

Born: October 30, 1912, in New York City, New York.

Died: September 17, 1979, in Herndon, Virginia.



A METHOD OF ESTIMATING COMPARATIVE RATES FROM CLINICAL DATA. APPLICATIONS TO CANCER OF THE LUNG, BREAST, AND CERVIX ¹

JEROME CORNFIELD, *National Cancer Institute, National Institutes of Health, U. S. Public Health Service, Bethesda, Md.*

¹ Received for publication February 23, 1951.



FIGURE 1. Passport photograph of Ronald Aylmer Fisher at age 34. Reprinted from Box JF. RA Fisher: the life of a scientist. New York: John Wiley & Sons, Inc., 1978.

Fisher developed four lines of argument in questioning the causal relation of lung cancer to smoking.

- 1) If A is associated with B, then not only is it possible that A causes B, but it is also possible that B is the cause of A. In other words, smoking may cause lung cancer, but it is a logical possibility that lung cancer causes smoking.
- 2) There may be a genetic predisposition to smoke (and that genetic predisposition is presumably also linked to lung cancer).
- 3) Smoking is unlikely to cause lung cancer because secular trend and other ecologic data do not support this relation.
- 4) 4) Smoking does not cause lung cancer because inhalers are less likely to develop lung cancer than are noninhalers

Lung cancer and cigarette smoking

Consider the following data for fractions of the population (Cornfield, 1951)

	Having cancer of the lung	Healthy	Total
Smokers	$0.119 \cdot 10^{-3}$	0.579910	0.580025
Nonsmokers	$0.036 \cdot 10^{-3}$	0.419935	0.419971
Total	$0.155 \cdot 10^{-3}$	0.999845	1.000000

what is the proportion having cancer of the lung in each population?

Smokers: $0.119 \cdot 10^{-3} / 0.580025 = 2.05164 \cdot 10^{-4}$

Nonsmokers: $0.036 \cdot 10^{-3} / 0.419971 = 8.57202 \cdot 10^{-5}$

And the prevalence of lung cancer in smokers with respect to nonsmokers is

$$\text{Smokers/Nonsmokers} \approx 2.4$$

We can also write an easy Bayesian equation that leads to some information as to the causation of cancer of the lung

$$P(\text{Cancer}|\text{Smoker}) = \frac{P(\text{Smoker}|\text{Cancer})P(\text{Smoker})}{P(\text{Cancer})}$$

$$P(\text{Cancer}|\text{Nonsmoker}) = \frac{P(\text{Nonsmoker}|\text{Cancer})P(\text{Nonsmoker})}{P(\text{Cancer})}$$

Therefore

$$\frac{P(\text{Cancer}|\text{Smoker})}{P(\text{Cancer}|\text{Nonsmoker})} = \frac{P(\text{Smoker}|\text{Cancer})P(\text{Smoker})}{P(\text{Nonsmoker}|\text{Cancer})P(\text{Nonsmoker})}$$

and with the numbers in the table one finds that this ratio is about 3.5.

According to Jeffreys, a Bayes ratio of 3.5 is already substantial support in favor of the hypothesis that smoking does cause lung cancer.

$\log_{10}(B)$	B	Evidence support
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

In 1954 Richard Doll and Bradford Hill published evidence in the British Medical Journal showing a strong link between smoking and lung cancer. They published further evidence in 1956.

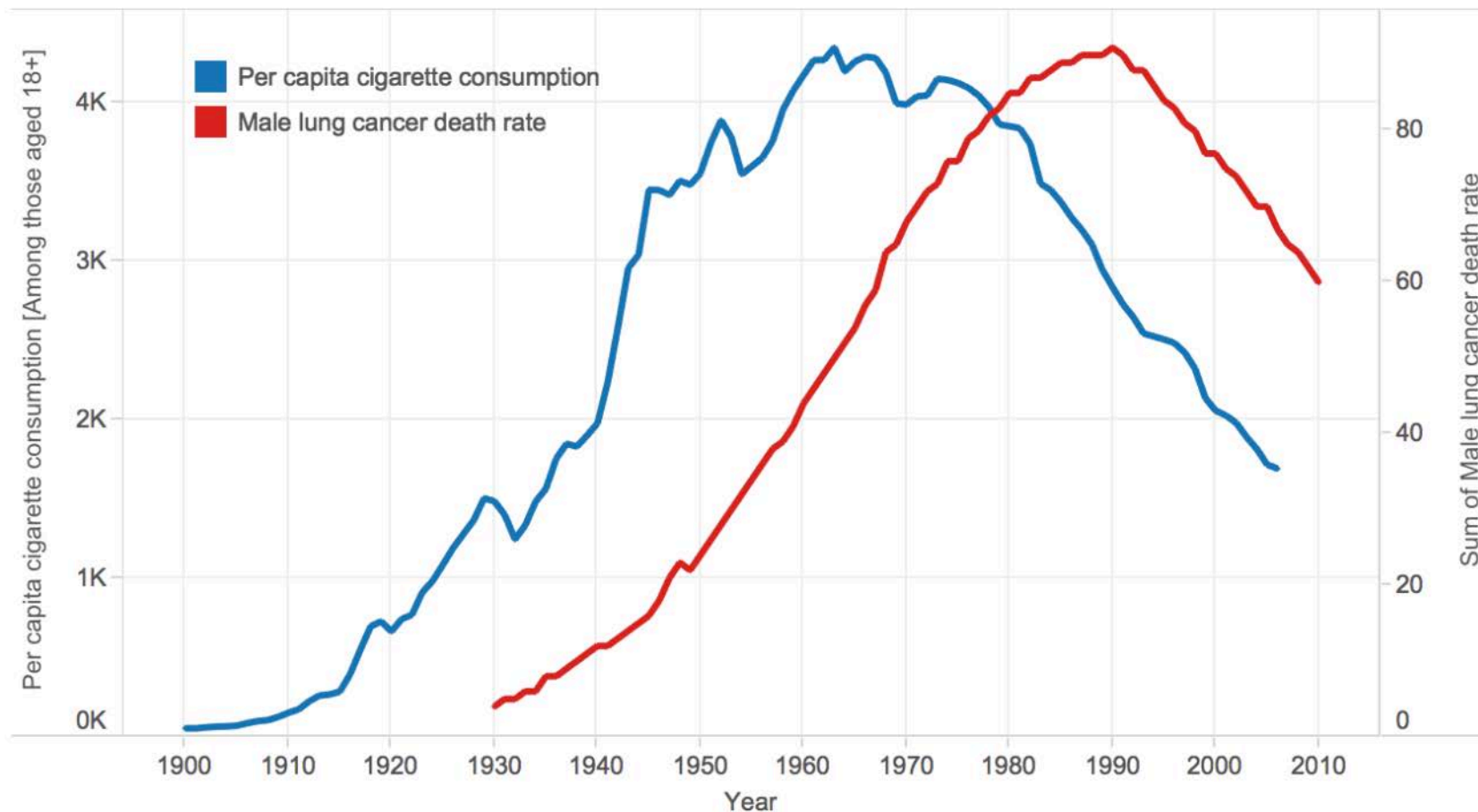
Fisher was a paid tobacco industry consultant and a devoted pipe smoker. He did not think the statistical evidence for a link was convincing.

Ronald Fisher died aged 72 on July 29, 1962, in Adelaide, Australia following an operation for colon cancer.

With bitter irony, we now know that the likelihood of getting this disease increases in smokers. Ronald Fisher was cremated and his ashes interred in St. Peter's Cathedral, Adelaide.

(from "Ronald Fisher." Famous Scientists. famousscientists.org. 17 Sep. 2015. Web. 5/30/2017 <www.famousscientists.org/ronald-fisher/>.)

Trends in Tobacco Use and Lung Cancer Death Rates in the U.S.



Death rates source: US Mortality Data, 1960-2010, US Mortality Volumes, 1930-1959, National Center for Health Statistics, Centers for Disease Control and Prevention.

Cigarette consumption source: US Department of Agriculture, 1900-2007.

TOBACCO THREATENS US ALL



**SAY NO TO
TOBACCO**

**PROTECT HEALTH,
REDUCE POVERTY AND
PROMOTE DEVELOPMENT**

31MAY:WORLDNOTOBACCODAY

#NoTobacco

Bayesian classification

data X , classes C

this likelihood is defined by training data

$$P(C|X) = \frac{P(X|C)}{P(X)} P(C)$$

the prior is also defined by training data

we can use the prior learning to assign a class to new data

$$C_k = \arg \max_{C_k} \frac{P(X|C_k)}{P(X)} P(C_k) = \arg \max_{C_k} P(X|C_k) P(C_k)$$

Consider a vector of N attributes given as Boolean variables $\mathbf{x} = \{x_i\}$ and classify the data vectors with a single Boolean variable.

The learning procedure must yield:

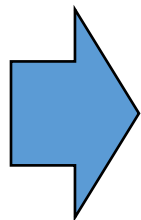
$P(y)$

it is easy to obtain it as an empirical distribution from an histogram of training class data: y is Boolean, the histogram has just two bins, and a hundred examples suffice to determine the empirical distribution to better than 10%.

$P(\mathbf{x}|y)$

there is a bigger problem here: the arguments have 2^{N+1} different values, and we must estimate $2(2^N-1)$ parameters ... for instance, with $N = 30$ there are more than 2 billion parameters!

How can we reduce the huge complexity of learning?



we assume the conditional independence of the x_n 's:
naive Bayesian learning

for instance, with just two attributes

$$P(x_1, x_2 | y) = P(x_1 | x_2, y) P(x_2 | y) = P(x_1 | y) P(x_2 | y)$$



conditional independence assumption

with more than 2 attributes

$$P(\mathbf{x} | y) \approx \prod_{k=1}^N P(x_k | y)$$

Therefore:

$$P(y_k | \mathbf{x}) = \frac{P(\mathbf{x} | y_k) P(y_k)}{P(\mathbf{x})} = \frac{P(\mathbf{x} | y_k)}{\sum_j P(\mathbf{x} | y_j) P(y_j)} P(y_k)$$
$$\approx \frac{\prod_{n=1}^N P(x_n | y_k)}{\sum_j P(y_j) \prod_{n=1}^N P(x_n | y_j)} P(y_k)$$

and we assign the class according to the rule (MAP)

$$y = \arg \max_{y_k} \frac{\prod_{n=1}^N P(x_n | y_k)}{\sum_j P(y_j) \prod_{n=1}^N P(x_n | y_j)} P(y_k)$$

More general discrete inputs

If any of the N variables has J different values, and if there are K classes, then we must estimate in all $NK(J-1)$ free parameters with the Naive Bayes Classifier (this includes normalization) (compare this with the $K(J^N-1)$ parameters needed by a complete classifier)

Continuous inputs and discrete classes – the Gaussian case

$$P(x_n | y_k) = \frac{1}{\sqrt{2\pi\sigma_{nk}^2}} \exp\left[-\frac{(x_n - \mu_{nk})^2}{2\sigma_{nk}^2}\right]$$

here we must estimate $2NK$ parameters + the shape of the distribution $P(y)$ (this adds up to another $K-1$ parameters)

Gaussian special case with class-independent variance and Boolean classification (two classes only):

$$P(y = 0 | \mathbf{x}) = \frac{P(\mathbf{x} | y = 0)P(y = 0)}{P(\mathbf{x} | y = 0)P(y = 0) + P(\mathbf{x} | y = 1)P(y = 1)}$$

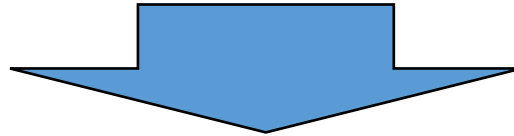
$$P(x_n | y = 0) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{(x_n - \mu_{n0})^2}{2\sigma_n^2}\right]$$

$$P(x_n | y = 1) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{(x_n - \mu_{n1})^2}{2\sigma_n^2}\right]$$

$$\begin{aligned}
P(y = 0 | \mathbf{x}) &= \frac{P(\mathbf{x} | y = 0) P(y = 0)}{P(\mathbf{x} | y = 0) P(y = 0) + P(\mathbf{x} | y = 1) P(y = 1)} \\
&= \frac{1}{1 + \frac{P(\mathbf{x} | y = 1) P(y = 1)}{P(\mathbf{x} | y = 0) P(y = 0)}} \\
&= \frac{1}{1 + \frac{P(y = 1)}{P(y = 0)} \prod_{n=1}^N \exp \left[-\frac{(x_n - \mu_{n1})^2}{2\sigma_n^2} + \frac{(x_n - \mu_{n0})^2}{2\sigma_n^2} \right]} \\
&= \frac{1}{1 + \exp \left\{ \ln \left(\frac{P(y = 1)}{P(y = 0)} \right) + \sum_{n=1}^N \left[\frac{(\mu_{n1} - \mu_{n0}) x_n}{\sigma_n^2} + \frac{\mu_{n0}^2 - \mu_{n1}^2}{2\sigma_n^2} \right] \right\}}
\end{aligned}$$

$$w_0 = \ln \left(\frac{P(y=1)}{P(y=0)} \right) + \sum_{n=1}^N \left[\frac{\mu_{n0}^2 - \mu_{n1}^2}{2\sigma_n^2} \right]$$

$$w_n = \frac{(\mu_{n1} - \mu_{n0})}{\sigma_n^2}$$



logistic shape

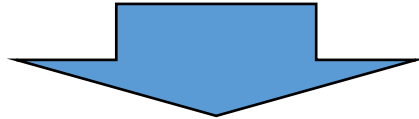
$$P(y=0|\mathbf{x}) = \frac{1}{1 + \exp \left(w_0 + \sum_{n=1}^N w_n x_n \right)}$$



$$P(y=1|\mathbf{x}) = 1 - P(y=0|\mathbf{x}) = \frac{\exp \left(w_0 + \sum_{n=1}^N w_n x_n \right)}{1 + \exp \left(w_0 + \sum_{n=1}^N w_n x_n \right)}$$

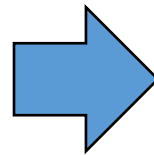
Finally an input vector belongs to class $y = 0$ if

$$\frac{P(y = 0|\mathbf{x})}{P(y = 1|\mathbf{x})} > 1$$

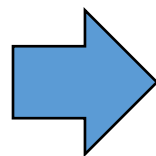


$$P(y = 0|\mathbf{x}) = \frac{1}{1 + \exp\left(w_0 + \sum_{n=1}^N w_n x_n\right)}$$

$$P(y = 1|\mathbf{x}) = \frac{\exp\left(w_0 + \sum_{n=1}^N w_n x_n\right)}{1 + \exp\left(w_0 + \sum_{n=1}^N w_n x_n\right)}$$



$$\exp\left(w_0 + \sum_{n=1}^N w_n x_n\right) < 1$$



$$w_0 + \sum_{n=1}^N w_n x_n < 0$$

Naive Bayesian learning is an example of supervised learning, however there are also unsupervised Bayesian learning methods, such as the AUTOCLASS program and similar such projects.



- + NASA Home
- + Ames Home
- + Intelligent Systems Division
- + Robust Software Engineering
- + Synthesis Projects & Applications

AutoClass

+ Home

+ AutoClass C

+ Web-based Interface

+ References



Introduction

In previous years, the Bayes group at Ames Research Center developed the basic theory and associated algorithms for various kinds of general data analysis techniques. Our earliest efforts were applied to the problem of automatic classification of data. We implemented this theory in the Autoclass series of programs. AutoClass takes a database of cases described by a combination of real and discrete valued attributes, and automatically finds the natural classes in that data. It does not need to be told how many classes are present or what they look like -- it extracts this information from the data itself. The classes are described probabilistically, so that an object can have partial membership in the different classes, and the class definitions can overlap. AutoClass generates reports on the classes it has found at the end of its search. AutoClass has been used and tested on many data sets, both within NASA and by industry, academia and other agencies. These applications typically find surprising classifications that show patterns in the data unknown to the user. Examples include: discovery of new classes of infra-red stars in the IRAS Low Resolution Spectral catalogue (see figure below; and see [here](#) and [here](#) for more information), new classes of airports in a database of all USA airports, discovery of classes of proteins, introns and other patterns in DNA/protein sequence data, and others.

The starting point of AUTOCLASS is a **mixture model**

$$dP(x) = \sum_k p_k dP_k(x|\theta); \quad \sum_k p_k = 1$$

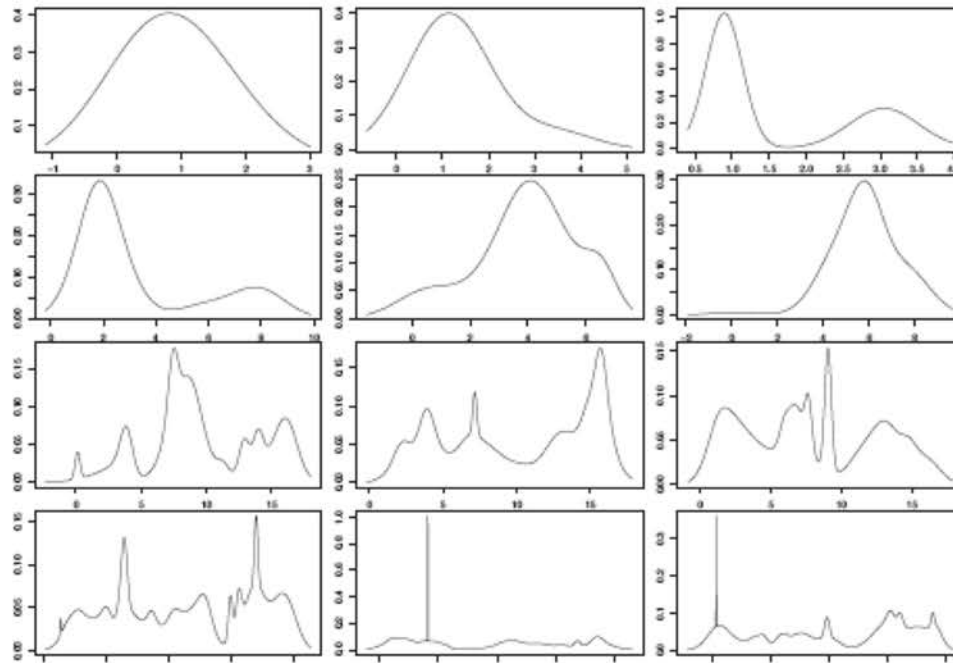


FIGURE 1. Some normal mixture densities for $K = 2$ (first row), $K = 5$ (second row), $K = 25$ (third row) and $K = 50$ (last row).

$$dP(x) = \sum_k p_k dP_k(x|\theta)$$

there is a variable number of classes

the probabilities of belonging to a given class are drawn from a multinomial distribution

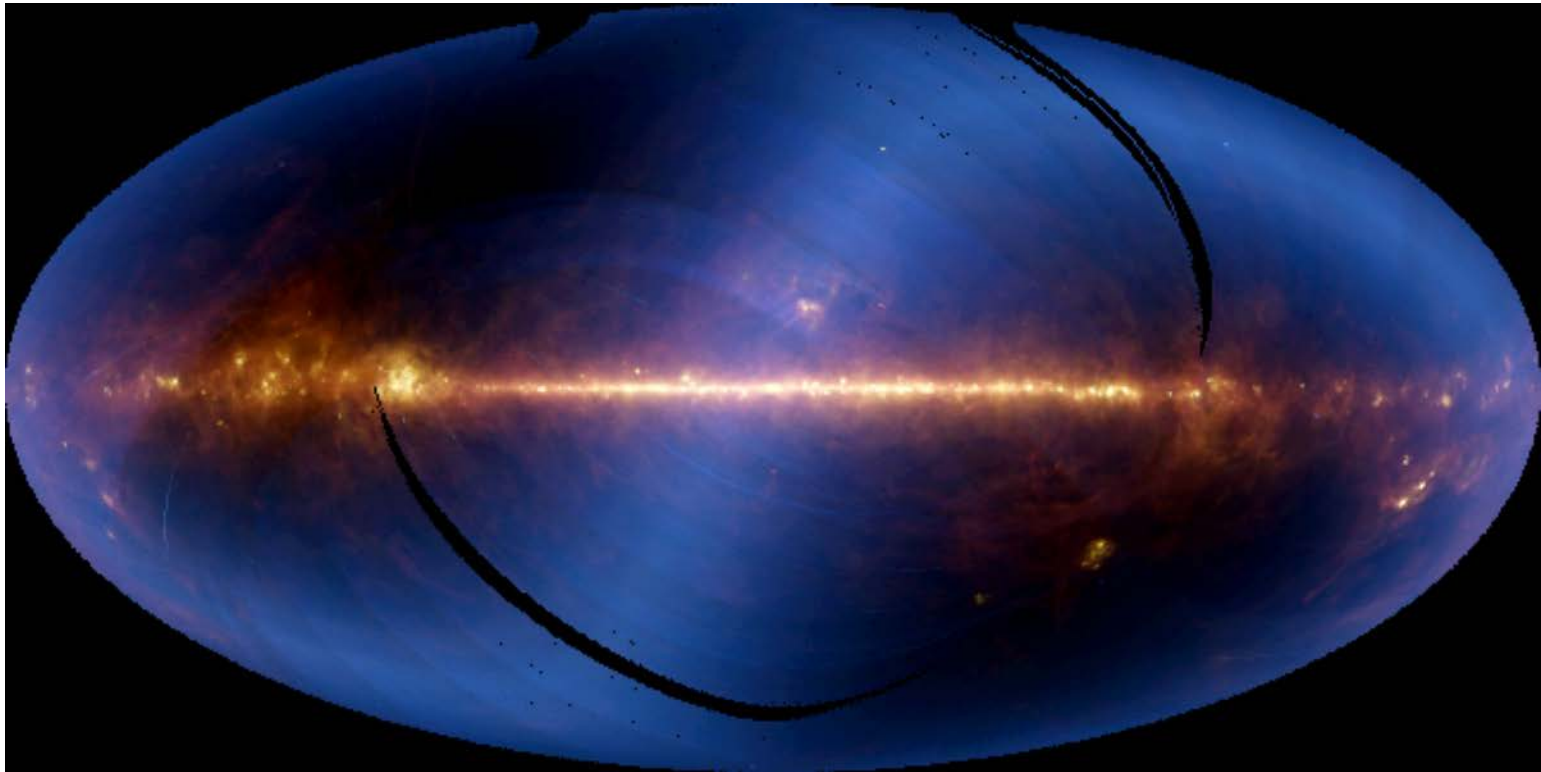
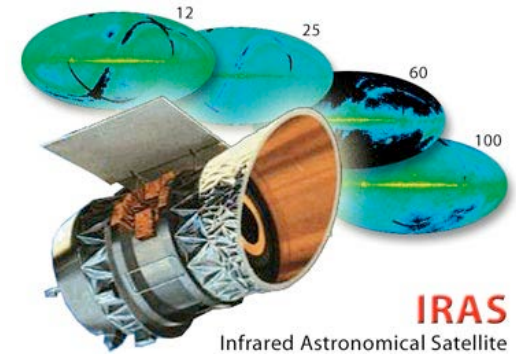
the component distributions are taken from a set of predefined distributions

the parameters define the shape of the component distribution

AUTOCLASS chooses a distribution and a parameter set for each class. Every data set determines a likelihood, and therefore a posterior distribution.

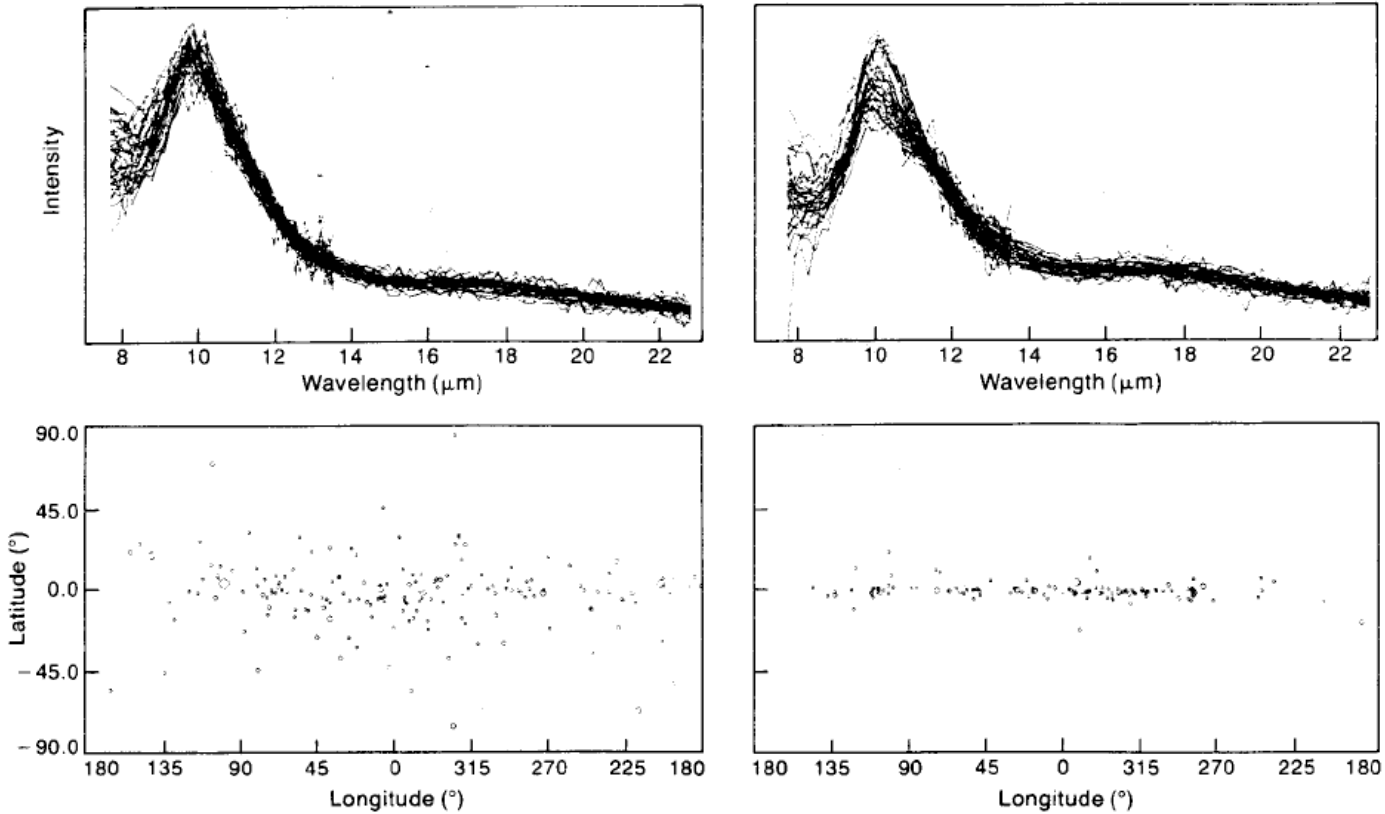
The class is selected by maximizing the posterior probability (MAP class estimate).

The Infrared Astronomical Satellite (IRAS) was the first-ever space telescope to perform a survey of the entire night sky at infrared wavelengths (launch date: 25 January 1983; mission end date: November 21, 1983)



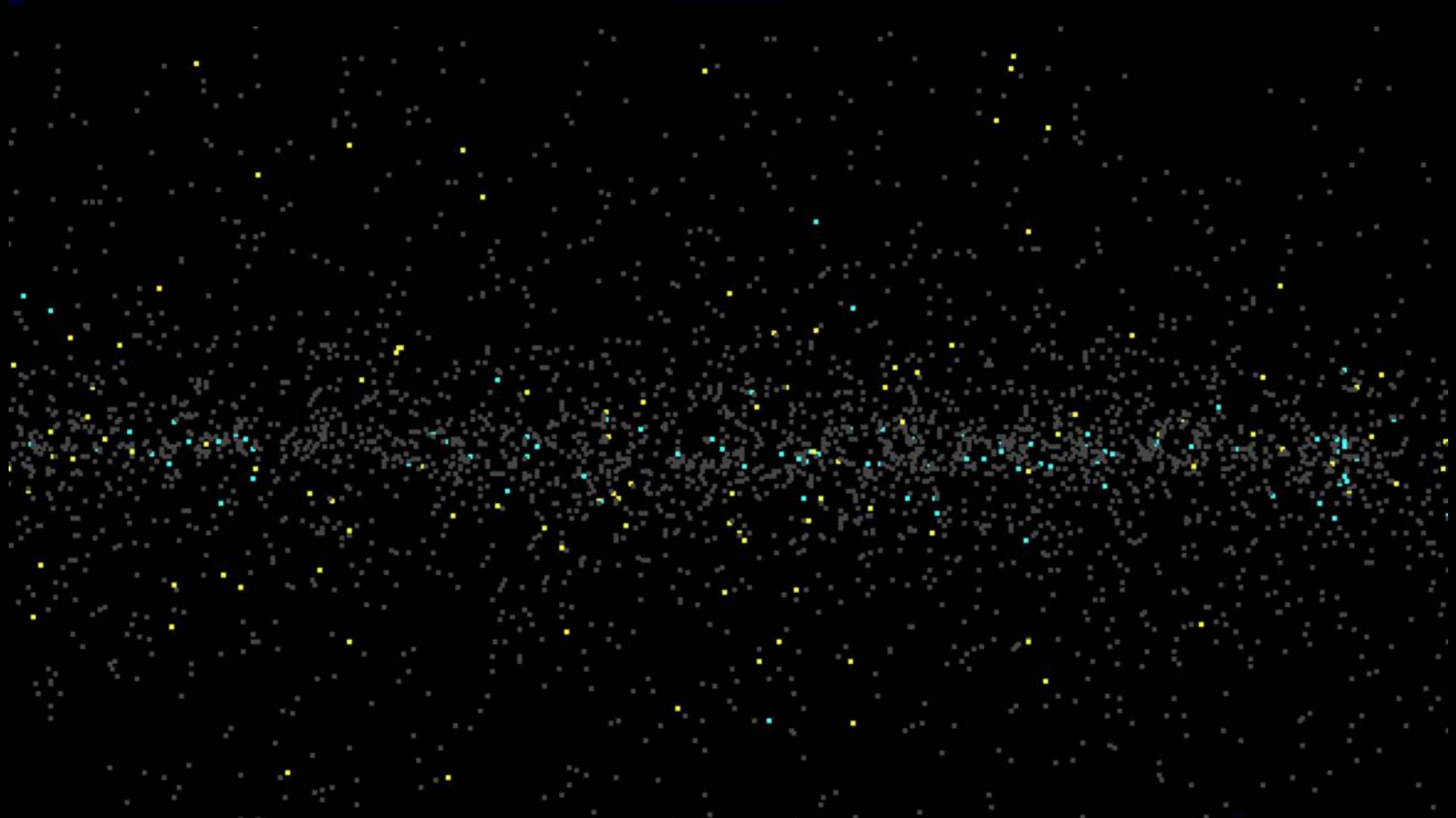
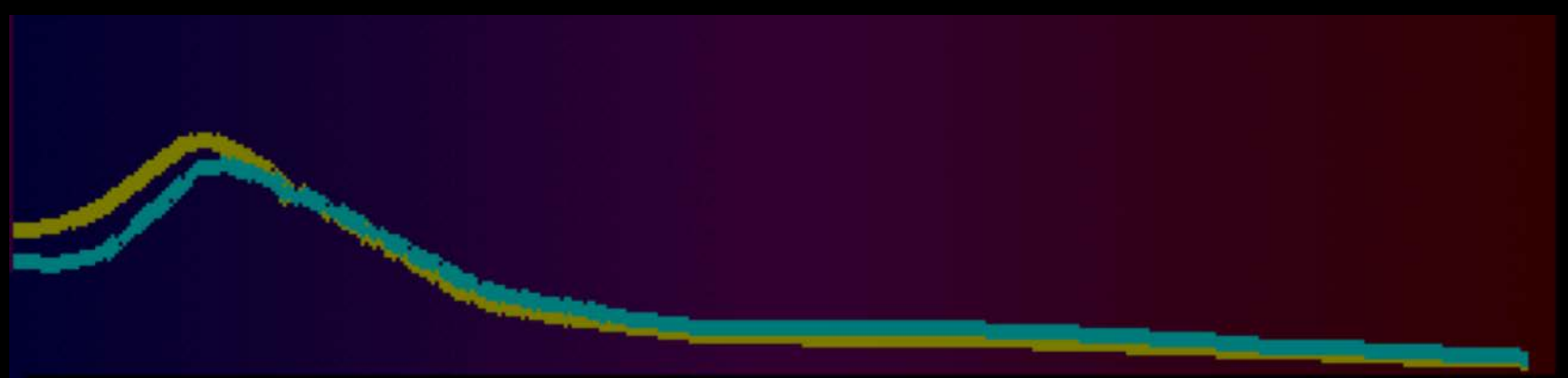
Infrared all-sky survey by IRAS (<http://irsa.ipac.caltech.edu/Missions/iras.html>)

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 above. 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.




AutoClass@IJM: a powerful tool for Bayesian classification of heterogeneous data in biology

Fiona Achcar^{1,2}, Jean-Michel Camadro² and Denis Mestivier^{1,*}

¹'Modeling in Integrative Biology' Group and ²'Protein Engineering and Metabolic Control' Group, Jacques Monod Institute, UMR7592 CNRS and Univ Paris-Diderot, Bâtiment Buffon, 15 rue Héléne Brion, 75205 Paris Cedex 13, France

Received January 31, 2009; Revised April 23, 2009; Accepted May 11, 2009

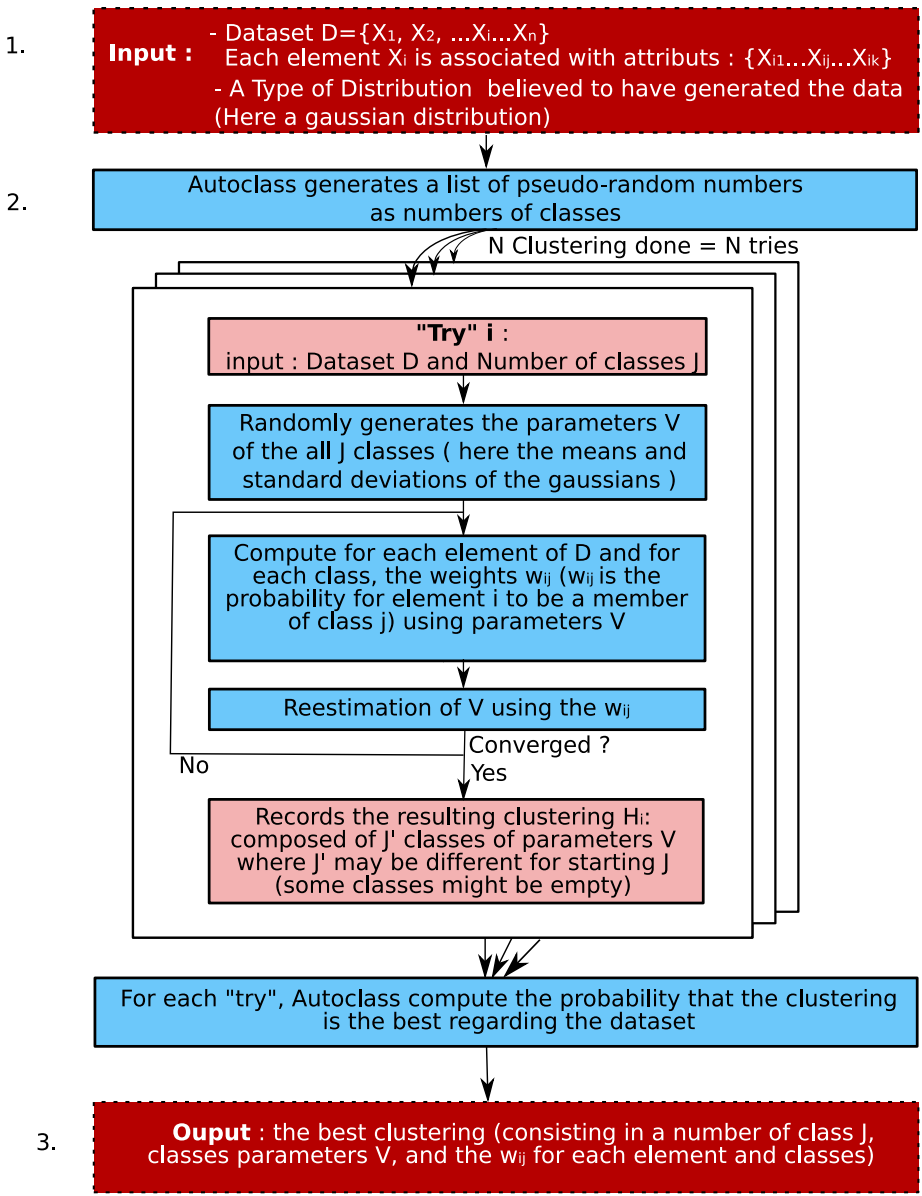


AutoClass@IJM

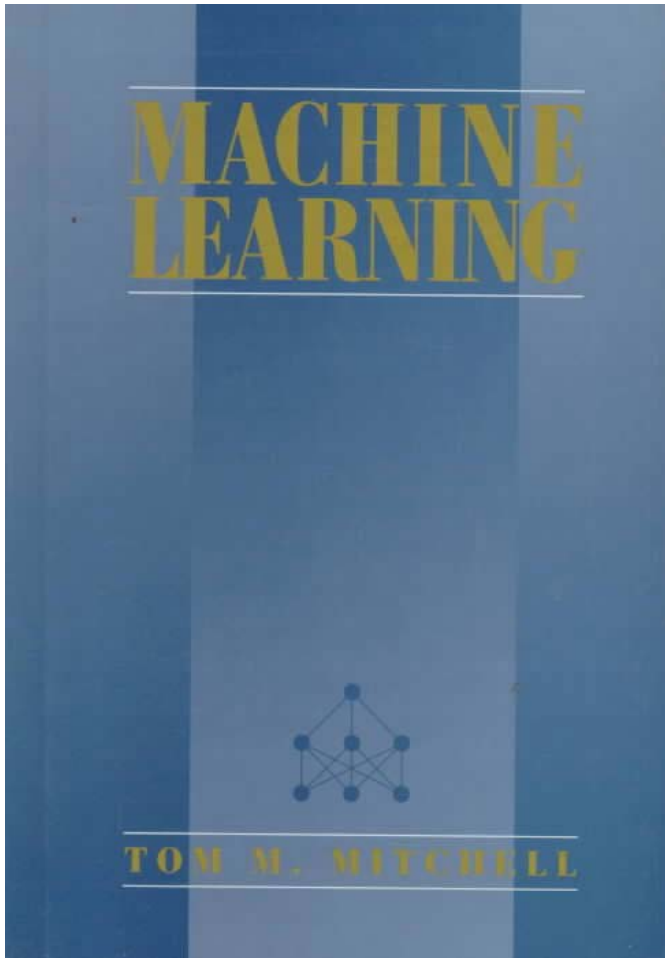
the webservice for [AutoClass Bayesian clustering system](#).

Developped by F. Achcar^{1,2} and D. Mestivier¹ in collaboration with J.M. Camadro²

We kindly ask users to cite [this paper](#) when publishing results derived of the use of AutoClass@IJM.



Naive Bayesian classifiers are part of the current toolbox of **machine learning** (see, e.g., Tom Mitchell's introductory book "Machine Learning", <http://www.cs.cmu.edu/~tom/>)



1. Introduction
2. Concept Learning and the General-to-Specific Ordering
3. Decision Tree Learning
4. Artificial Neural Networks
5. Evaluating Hypotheses
6. **Bayesian Learning**
7. Computational Learning Theory
8. Instance-Based Learning
9. Genetic Algorithms
10. Learning Sets of Rules
11. Analytical Learning
12. Combining Inductive and Analytical Learning
13. Reinforcement Learning

On the nature of learning in Bayesian and MaxEnt Inference

(from Cheeseman & Stutz, 2004)

here we consider these three problems:

1. find the probabilities θ_i of getting face i in a throw of a possibly biased die, given the frequencies n_i of each face in a total of N throws;
2. find the probabilities when only the mean $M = \sum_{i=1}^6 i n_i$ and the total number of throws N , are given;
3. analyze the kangaroo problem with a more complex contingency table

1. Find the probabilities θ_i of getting face i in a throw of a possibly biased die, given the frequencies n_i of each face in a total of N throws;

$$0 \leq \theta_i \leq 1; \quad \sum_{i=1}^6 \theta_i = 1; \quad 0 \leq n_i \leq N; \quad \sum_{i=1}^6 n_i = N$$

likelihood is given by the multinomial probability

$$L(\{n_1, \dots, n_6\} | \boldsymbol{\theta}, N, I) = \frac{N!}{\prod_{j=1}^6 n_j!} \prod_{i=1}^6 \theta_i^{n_i}$$

if, initially, we take a uniform prior, the posterior distribution from Bayes' theorem is

$$\begin{aligned} p(\boldsymbol{\theta} | \{n_1, \dots, n_6\}, N, I) &= \frac{\prod_{i=1}^6 \theta_i^{n_i} \delta\left(\sum_{j=1}^6 \theta_j - 1\right)}{\int_0^1 \prod_{i=1}^6 \theta_i^{n_i} \delta\left(\sum_{j=1}^6 \theta_j - 1\right) d\boldsymbol{\theta}_i} \\ &= \frac{\Gamma(N + 6)}{\prod_{j=1}^6 \Gamma(n_j + 1)} \prod_{i=1}^6 \theta_i^{n_i} \delta\left(\sum_{j=1}^6 \theta_j - 1\right) \end{aligned}$$

and we obtain a Dirichlet distribution (conjugate posterior of the multinomial distribution, just as the Beta distribution is the conjugate posterior of the binomial distribution).

Mathematical note on the normalization of the Dirichlet distribution:

$$B(m, n) = \int_0^1 t^{m-1} (1-t)^{n-1} dt = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)} \quad \text{relationship between Beta and Gamma function}$$

$$\begin{aligned} \int_{0 \leq \theta_i \leq 1} \theta_1^{n_1} \theta_2^{n_2} \theta_3^{n_3} \delta(\theta_1 + \theta_2 + \theta_3 - 1) d\theta_1 d\theta_2 d\theta_3 &= \int_{0 \leq \theta_i \leq 1} \theta_1^{n_1} d\theta_1 \int_0^{1-\theta_1} p^{n_2} [(1-\theta_1) - p]^{n_3} dp \\ &= \int_{0 \leq \theta_i \leq 1} \theta_1^{n_1} d\theta_1 (1-\theta_1)^{n_2+n_3+1} \int_0^1 x^{n_2} (1-x)^{n_3} dx \\ &= B(n_2+1, n_3+1) \int_0^1 \theta_1^{n_1} (1-\theta_1)^{n_2+n_3+1} d\theta_1 = B(n_2+1, n_3+1) B(n_1+1, n_2+n_3+2) \\ &= \frac{\Gamma(n_2+1)\Gamma(n_3+1)}{\Gamma(n_2+n_3+2)} \cdot \frac{\Gamma(n_1+1)\Gamma(n_2+n_3+2)}{\Gamma(n_1+n_2+n_3+3)} = \frac{\Gamma(n_2+1)\Gamma(n_3+1)\Gamma(n_1+1)}{\Gamma(n_1+n_2+n_3+3)} \end{aligned}$$

$$\int_{0 \leq \theta_i \leq 1} \prod_{i=1}^M \theta_i^{n_i} d\theta_i \delta\left(\sum_{j=1}^M \theta_j - 1\right) = \frac{\prod_{i=1}^M \Gamma(n_i+1)}{\Gamma(N+M)} \quad \text{normalization factor}$$

thus, if we assume some prior information, we can start with a Dirichlet prior

$$p(\boldsymbol{\theta}|\mathbf{w}, I) = \frac{\Gamma(W)}{\prod_{j=1}^6 \Gamma(w_j)} \prod_{i=1}^6 \theta_i^{w_j-1} \delta\left(\sum_{j=1}^6 \theta_j - 1\right) \quad \text{with} \quad W = \sum_{j=1}^6 w_j$$

and obtain the posterior distribution

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{n}, \mathbf{w}, N, I) &= \frac{\prod_{i=1}^6 \theta_i^{n_i+w_i-1} \delta\left(\sum_{j=1}^6 \theta_j - 1\right)}{\int_0^1 \prod_{i=1}^6 \theta_i^{n_i+w_i-1} \delta\left(\sum_{j=1}^6 \theta_j - 1\right) d\boldsymbol{\theta}_i} = \frac{\Gamma(N+W)}{\prod_{j=1}^6 \Gamma(n_j+w_j)} \prod_{i=1}^6 \theta_i^{n_i+w_i-1} \delta\left(\sum_{j=1}^6 \theta_j - 1\right) \\ &= \frac{N!}{\prod_{j=1}^6 n_j!} \cdot \frac{\Gamma(W)}{\prod_{j=1}^6 \Gamma(w_j)} \prod_{i=1}^6 \theta_i^{n_i+w_i-1} \delta\left(\sum_{j=1}^6 \theta_j - 1\right) \end{aligned}$$

The inferred distribution can be used to compute averages, and also for prediction.

Indeed, the probability of observing r_i occurrences of the i -th face in the future is

$$\begin{aligned}
 P(\mathbf{r}|\mathbf{n}, N, R, \mathbf{w}, I) &= \int_{\boldsymbol{\theta}} P(\mathbf{r}|\boldsymbol{\theta}, N, R, I) p(\boldsymbol{\theta}|\mathbf{n}, N, \mathbf{w}, I) d\boldsymbol{\theta} = \\
 &= \int_{\boldsymbol{\theta}} \frac{R!}{\prod_{j=1}^6 r_j!} \prod_{i=1}^6 \theta_i^{r_i} \frac{\Gamma(N+W)}{\prod_{j=1}^6 \Gamma(n_j+w_j)} \prod_{i=1}^6 \theta_i^{n_i+w_i-1} \delta\left(\sum_{j=1}^6 \theta_j - 1\right) d\boldsymbol{\theta} \\
 &= \frac{R!}{\prod_{j=1}^6 r_j!} \cdot \frac{\Gamma(N+W)}{\prod_{j=1}^6 \Gamma(n_j+w_j)} \cdot \frac{\prod_{j=1}^6 \Gamma(n_j+r_j+w_j)}{\Gamma(N+R+W)}
 \end{aligned}$$

so that we find, e.g.,

$$P(r_1 = 1 | \mathbf{n}, N, R = 1, \mathbf{w}, I) = \frac{\Gamma(N + W)}{\prod_{j=1}^6 \Gamma(n_j + w_j)} \cdot \frac{\prod_{j=1}^6 \Gamma(n_j + w_j + \delta_{1j})}{\Gamma(N + W + 1)}$$
$$= \frac{n_1 + w_1}{N + W}$$

2. Find the probabilities when only the total $M = \sum_{i=1}^6 i n_i$, and the total of throws N , are given

Let $\langle \mathbf{n} \rangle_{NM}$ be the set of vectors that satisfy the conditions,

$$N = \sum_{i=1}^6 n_i; \quad M = \sum_{i=1}^6 i n_i$$

then the likelihood is

$$P(M | \boldsymbol{\theta}, N, I) = \sum_{\langle \mathbf{n} \rangle_{NM}} P(\mathbf{n} | \boldsymbol{\theta}, N, I) = \sum_{\langle \mathbf{n} \rangle_{NM}} \frac{N!}{\prod_{j=1}^6 n_j!} \prod_{i=1}^6 \theta_i^{n_i}$$

now notice that

$$\begin{aligned}
 P(\boldsymbol{\theta}|M, N, \mathbf{w}, I) &= \frac{P(M|\boldsymbol{\theta}, N, I)P(\boldsymbol{\theta}|N, \mathbf{w}, I)}{P(M|N, I)} \\
 &= \frac{\sum_{\langle \mathbf{n} \rangle_{NM}} P(\mathbf{n}|\boldsymbol{\theta}, N, I)P(\boldsymbol{\theta}|N, \mathbf{w}, I)}{\sum_{\langle \mathbf{n} \rangle_{NM}} \int_{\boldsymbol{\theta}} P(\mathbf{n}|\boldsymbol{\theta}, N, I)P(\boldsymbol{\theta}|N, \mathbf{w}, I)d\boldsymbol{\theta}}
 \end{aligned}$$

$$\sum_{\langle \mathbf{n} \rangle_{NM}} P(\mathbf{n}|\boldsymbol{\theta}, N, I)P(\boldsymbol{\theta}|N, \mathbf{w}, I) = \sum_{\langle \mathbf{n} \rangle_{NM}} \frac{N!}{\prod_{j=1}^6 n_j!} \frac{\Gamma(W)}{\prod_{j=1}^6 \Gamma(w_j)} \prod_{i=1}^6 \theta_i^{n_i+w_i-1}$$

$$\sum_{\langle \mathbf{n} \rangle_{NM}} \int_{\boldsymbol{\theta}} P(\mathbf{n}|\boldsymbol{\theta}, N, I)P(\boldsymbol{\theta}|N, \mathbf{w}, I)d\boldsymbol{\theta} = \sum_{\langle \mathbf{n} \rangle_{NM}} \frac{N!}{\prod_{j=1}^6 n_j!} \frac{\Gamma(W)}{\prod_{j=1}^6 \Gamma(w_j)} \frac{\prod_{i=1}^6 \Gamma(n_i + w_i)}{\Gamma(N + W)}$$

from these formulas we can calculate all marginals and any expectation, although it is quite difficult to manipulate them

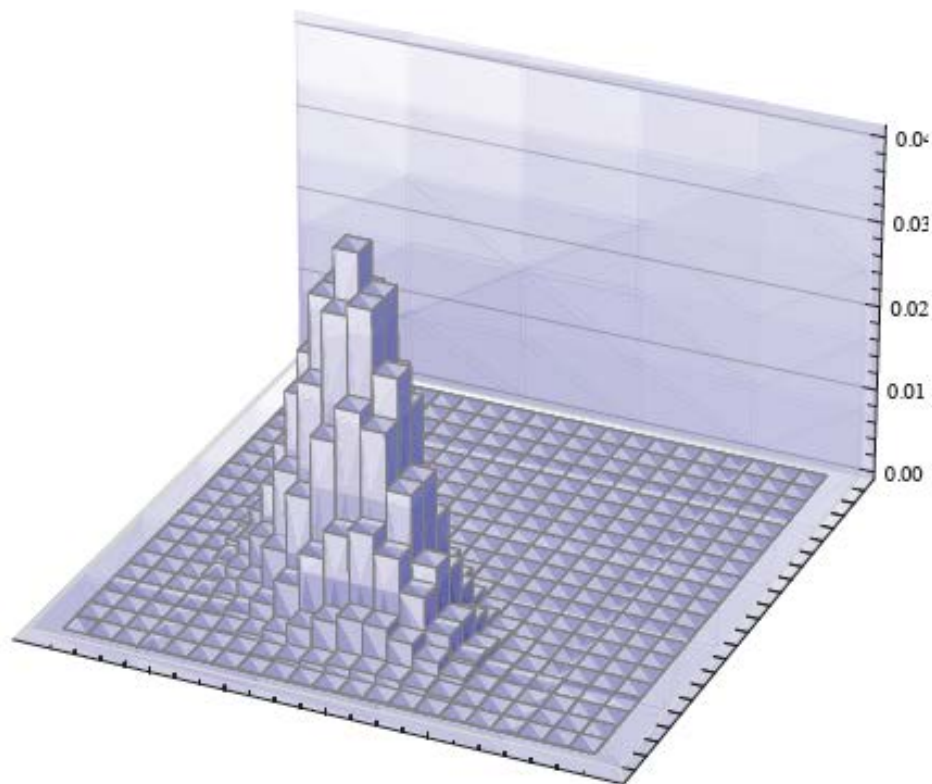
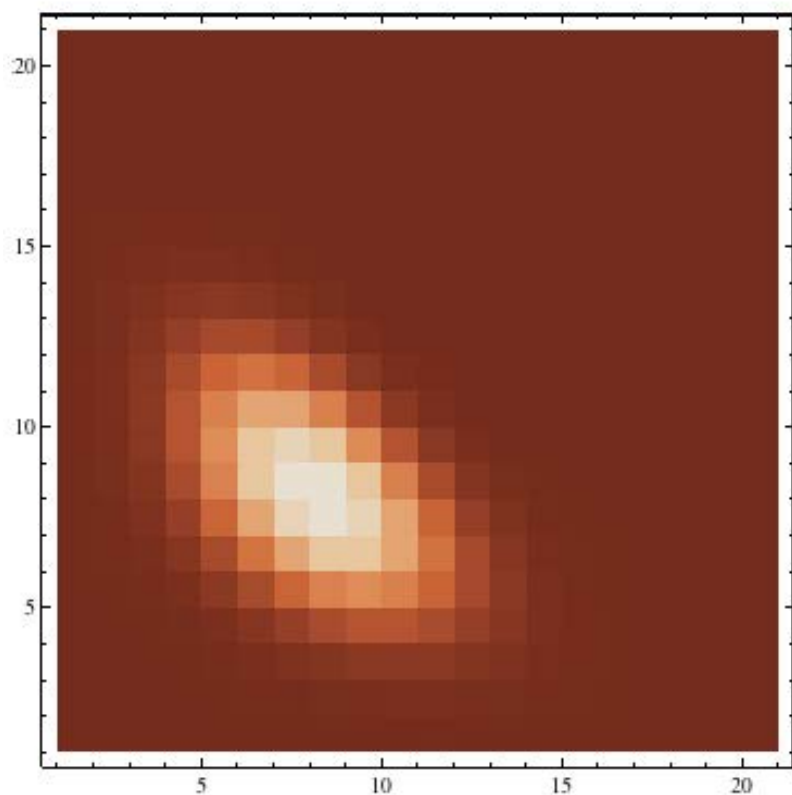


Figure 3: Multinomial distribution with $n = 20$, $k = 3$ and $p_1 = p_2 = p_3 = 1/3$, plotted as a function of the independent values n_1 and n_2 . Density plot (left panel) and lego plot (right panel). As an exercise, explain why in this symmetrical case the distribution is not centered in the n_1, n_2 domain, and consider ways to represent multinomial distributions with $k > 3$.

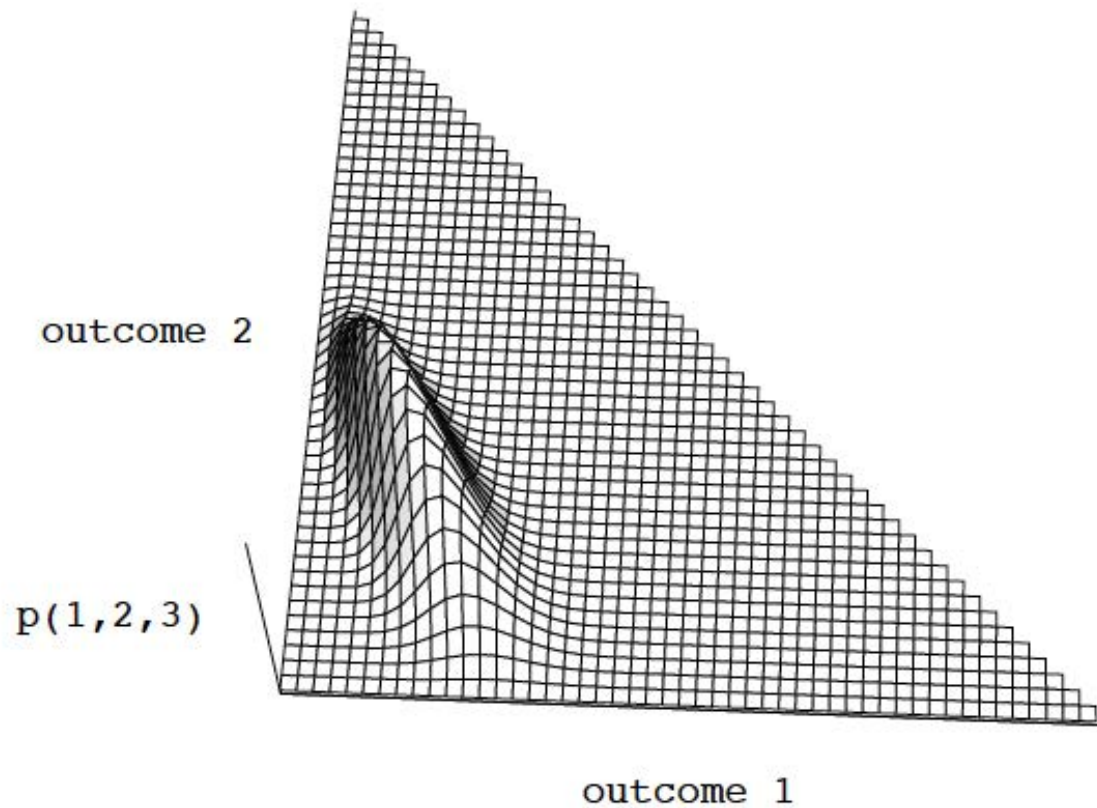


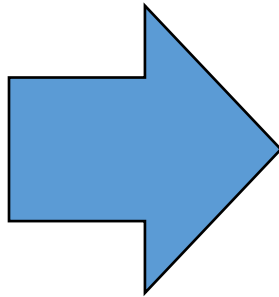
FIGURE 2. The posterior density for the 3-faces die example with a mean spot count of 2.5, $N = 60$, and prior weights of (1,1,1). Because of the normalization constraint, the third variable (not shown) is given by $\theta_3 = 1 - \theta_1 - \theta_2$.

The figure, from C&S, shows that the probability mass is concentrated close to the subspace defined by constraints, and becomes increasingly so as N increases. Bayesian inference tells us nothing on the distribution inside the subspace.

3. The kangaroo problem with an extended contingency table

attributes (number of values):

- handedness (2)
- beer-drinking (2)
- state-of-origin (7)
- color (3)



4-dimensional contingency table
with $2 \times 2 \times 7 \times 3 = 84$ entries

The size of the contingency table increases exponentially as the number of attributes grows

If we are given the number of occurrences $n_{i,j,k,l}$ for each position in the contingency table, we fall back on the first example of dice throw

$$0 \leq \theta_{i,j,k,l} \leq 1; \quad \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^7 \sum_{l=1}^3 \theta_{i,j,k,l} = 1$$

$$0 \leq n_{i,j,k,l} \leq N; \quad \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^7 \sum_{l=1}^3 n_{i,j,k,l} = N$$

with the likelihood

$$L(\mathbf{n} | \boldsymbol{\theta}, N, I) = \frac{N!}{\prod_{i,j,k,l} n_{i,j,k,l}} \prod_{i,j,k,l} \theta_{i,j,k,l}^{n_{i,j,k,l}}$$

The $n_{i,j,k,l}$'s are sufficient statistics and we can estimate all the corresponding probabilities as in the first example.

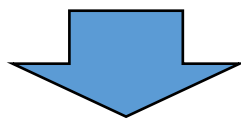
However if we are only given a set of marginals, i.e., of constraints, we are in the same situation as example 2, the marginals define a subspace of the whole parameter space, and in this subspace the distribution is eventually determined by the prior information only.

With enough attributes, the contingency table becomes VERY large, and it becomes impossible to collect sufficient statistics, we are mostly limited to marginals.

The situation is very different if we assume independence: then the marginals are sufficient statistics. E.g., if probabilities factorize, then kangaroos have only $(2+2+7+3)-(1+1+1+1) = 10$ independent values (using normalization) instead of 84.

Maximum entropy approach to the kangaroo problem, given marginals

$$\sum_{j,k,l} n_{i,j,k,l} = n_i; \quad \sum_i n_i = N$$



$$\sum_{i,j,k,l} \theta_{i,j,k,l} = 1; \quad \sum_{j,k,l} \theta_{i,j,k,l} = \frac{n_i}{N}$$

Example with two marginals: we maximize the constrained entropy

$$S = - \sum_{i,j,k,l} \theta_{i,j,k,l} \log \theta_{i,j,k,l} + \lambda_0 \left(\sum_{i,j,k,l} \theta_{i,j,k,l} - 1 \right) + \lambda_1 \left(\sum_{j,k,l} \theta_{1,j,k,l} - \frac{n_1}{N} \right) + \lambda_2 \left(\sum_{i,k,l} \theta_{2,i,k,l} - \frac{n_2}{N} \right)$$

in the original kangaroo problem

$$S_V = \left(p_{bl} \log \frac{1}{p_{bl}} + p_{\bar{bl}} \log \frac{1}{p_{\bar{bl}}} + p_{b\bar{l}} \log \frac{1}{p_{b\bar{l}}} + p_{\bar{b}\bar{l}} \log \frac{1}{p_{\bar{b}\bar{l}}} \right) \\ + \lambda_1 (p_{bl} + p_{\bar{bl}} + p_{b\bar{l}} + p_{\bar{b}\bar{l}} - 1) + \lambda_2 (p_{bl} + p_{b\bar{l}} - 1/3) + \lambda_3 (p_{bl} + p_{\bar{bl}} - 1/3)$$

$$\frac{\partial S_V}{\partial p_{bl}} = -\log p_{bl} - 1 + \lambda_1 + \lambda_2 + \lambda_3 = 0$$

$$\frac{\partial S_V}{\partial p_{\bar{bl}}} = -\log p_{\bar{bl}} - 1 + \lambda_1 + \lambda_3 = 0$$

$$\frac{\partial S_V}{\partial p_{b\bar{l}}} = -\log p_{b\bar{l}} - 1 + \lambda_1 + \lambda_2 = 0$$

$$\frac{\partial S_V}{\partial p_{\bar{b}\bar{l}}} = -\log p_{\bar{b}\bar{l}} - 1 + \lambda_1 = 0$$

$$\begin{cases} p_{b\bar{l}} = p_{b\bar{l}} \exp(\lambda_3) \\ p_{b\bar{l}} = p_{b\bar{l}} \exp(\lambda_2) \\ p_{bl} = p_{b\bar{l}} \exp(\lambda_2 + \lambda_3) \end{cases} \Rightarrow p_{b\bar{l}} p_{b\bar{l}} = p_{bl} p_{b\bar{l}}$$

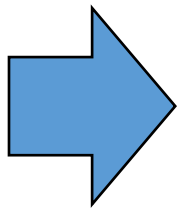
$$\begin{cases} p_{bl} + p_{b\bar{l}} + p_{b\bar{l}} + p_{b\bar{l}} = 1 \\ p_{bl} + p_{b\bar{l}} = 1/3 \\ p_{bl} + p_{b\bar{l}} = 1/3 \\ p_{b\bar{l}} p_{b\bar{l}} = p_{bl} p_{b\bar{l}} \end{cases} \Rightarrow \begin{cases} p_{b\bar{l}} = p_{b\bar{l}} = 1/3 - p_{bl} \\ p_{b\bar{l}} = 1/3 + p_{bl} \\ (1/3 - p_{bl})^2 = p_{bl}/3 + p_{bl}^2 \\ 1/9 - 2p_{bl}/3 + p_{bl}^2 = p_{bl}/3 + p_{bl}^2 \end{cases}$$

$$\Rightarrow p_{bl} = \frac{1}{9}; \quad p_{b\bar{l}} = p_{b\bar{l}} = \frac{2}{9}; \quad p_{b\bar{l}} = \frac{4}{9}$$

this solution coincides with the independence hypothesis

In the extended kangaroo problem we find

$$\frac{\partial S}{\partial \theta_{m,j,k,l}} = -(\log \theta_{m,j,k,l} + 1) + \lambda_0 + \lambda_m = 0$$



$$\theta_{1,j,k,l} = \exp(\lambda_0 + \lambda_1 - 1)$$

$$\theta_{2,j,k,l} = \exp(\lambda_0 + \lambda_2 - 1)$$

thus we obtain again a multiplicative structure.

Whatever the choice of marginals, probabilities factorize, and the MaxEnt solution corresponds to a set of independent probabilities.

Thus independence is built-in the MaxEnt method, which is a sort of “generalized independence method”.

the theory



that would



not die



how bayes' rule cracked



the enigma code,

hunted down russian

submarines & emerged

triumphant from two



centuries of controversy

sharon bertsch mcgrayne

A nice account of the history of Bayesian ideas by Sharon Bertsch (Yale Univ. Press, 2011)

(strongly bent towards history, no math)

References:

- P. Cheeseman and J. Stutz, “On the Relationship Between Bayesian and Maximum Entropy Inference”, in AIP Conf. Proc. , Volume 735, pp. 445-461, BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (2004)
- C. Elkan: “Naive Bayesian Learning”, CS97-557 tech. rep. UCSD
- Tom Mitchell: “Machine Learning” https://www.amazon.it/Machine-Learning/dp/1259096955/ref=sr_1_1?ie=UTF8&qid=1496165079&sr=8-1&keywords=tom+mitchell+machine+learning
- AUTOCLASS @ NASA:
<http://ti.arc.nasa.gov/tech/rse/synthesis-projects-applications/autoclass/>