Introduction to Bayesian Statistics - 7

PhD Physics course (XXVIII ciclo)

Università di Trieste

Edoardo Milotti

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

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

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:

$$P(y_k|\mathbf{x}) = \frac{P(\mathbf{x}|y_k)}{P(\mathbf{x})}P(y_k) = \frac{P(\mathbf{x}|y_k)}{\sum_j P(\mathbf{x}|y_j)P(y_j)}P(y_k)$$

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

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

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

More general discrete inputs

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

Continuous inputs and discrete classes – the Gaussian case

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

here we must estimate 2*NK* 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):

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

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

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

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

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

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

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.

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;

$$0 \le \theta_i \le 1; \qquad \sum_{i=1}^{6} \theta_i = 1; \qquad 0 \le n_i \le N; \qquad \sum_{i=1}^{6} n_i = N$$

likelihood is given by the multinomial probability

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

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

$$p\left(\boldsymbol{\theta} \middle| \{n_1,\ldots,n_6\}, N, I\right) = \frac{\prod_{i=1}^{6} \theta_i^{n_i} \delta\left(\sum_{j=1}^{6} \theta_j - 1\right)}{\int_0^1 \prod_{i=1}^{6} \theta_i^{n_i} \delta\left(\sum_{j=1}^{6} \theta_j - 1\right) d\theta_i}$$

$$= \frac{\Gamma(N+6)}{\prod_{j=1}^{6} \Gamma(n_j + 1)} \prod_{i=1}^{6} \theta_i^{n_i} \delta\left(\sum_{j=1}^{6} \theta_j - 1\right)$$

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:

$$B(m,n)=\int_0^1 t^{m-1}(1-t)^{n-1}\,dt=\frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$$

relationship between Beta and Gamma function

$$\int_{0\le\theta_i\le1}\theta_1^{n_1}\theta_2^{n_2}\theta_3^{n_3}\delta(\theta_1+\theta_2+\theta_3-1)\,d\theta_1\,d\theta_2\,d\theta_3=\int_{0\le\theta_i\le1}\theta_1^{n_1}\,d\theta_1\int_0^{1-\theta_1}p^{n_2}\left[(1-\theta_1)-p\right]^{n_3}dp$$

$$=\int_{0\le\theta_i\le1}\theta_1^{n_1}\,d\theta_1(1-\theta_1)^{n_2+n_3+1}\int_0^1 x^{n_2}(1-x)^{n_3}\,dx$$

$$=B(n_2+1,n_3+1)\int_0^1\theta_1^{n_1}(1-\theta_1)^{n_2+n_3+1}\,d\theta_1=B(n_2+1,n_3+1)B(n_1+1,n_2+n_3+2)$$

$$=\frac{\Gamma(n_2+1)\Gamma(n_3+1)}{\Gamma(n_2+n_3+2)}\cdot\frac{\Gamma(n_1+1)\Gamma(n_2+n_3+2)}{\Gamma(n_1+n_2+n_3+3)}=\frac{\Gamma(n_2+1)\Gamma(n_3+1)\Gamma(n_1+1)}{\Gamma(n_1+n_2+n_3+3)}$$

$$\int_{0\le\theta_i\le1}\prod_{i=1}^{M}\theta_i^{n_i}\,d\theta_i\delta\left(\sum_{j=1}^{M}\theta_j-1\right)=\frac{\prod_{i=1}^{M}\Gamma(n_i+1)}{\Gamma(N+M)}$$

normalization factor

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

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

and obtain the posterior distribution

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

$$=\frac{N!}{\prod_{j=1}^{6}n_j!}\cdot\frac{\Gamma(W)}{\prod_{j=1}^{6}\Gamma(w_j)}\prod_{i=1}^{6}\theta_i^{n_i+w_i-1}\delta\left(\sum_{j=1}^{6}\theta_j-1\right)$$

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

$$P\left(\mathbf{r}\middle|\mathbf{n},N,R,\mathbf{w},I\right)=\int_{\boldsymbol{\theta}}P\left(\mathbf{r}\middle|\boldsymbol{\theta},N,R,I\right)p\left(\boldsymbol{\theta}\middle|\mathbf{n},N,\mathbf{w},I\right)d\boldsymbol{\theta}=$$

$$=\int_{\boldsymbol{\theta}}\frac{R!}{\prod_{j=1}^{6}r_j!}\prod_{i=1}^{6}\theta_i^{r_i}\frac{\Gamma\left(N+W\right)}{\prod_{j=1}^{6}\Gamma\left(n_j+w_j\right)}\prod_{i=1}^{6}\theta_i^{n_i+w_i-1}\delta\left(\sum_{j=1}^{6}\theta_j-1\right)d\boldsymbol{\theta}$$

$$=\frac{R!}{\prod_{j=1}^{6}r_j!}\cdot\frac{\Gamma\left(N+W\right)}{\prod_{j=1}^{6}\Gamma\left(n_j+w_j\right)}\cdot\frac{\prod_{j=1}^{6}\Gamma\left(n_j+r_j+w_j\right)}{\Gamma\left(N+R+W\right)}$$

so that we find, e.g.,

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

$$= \frac{n_1 + w_1}{N+W}$$

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

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

now notice that

$$P\left(\boldsymbol{\theta}\middle|M,N,\mathbf{w},I\right)=\frac{P\left(M\middle|\boldsymbol{\theta},N,I\right)P\left(\boldsymbol{\theta}\middle|N,\mathbf{w},I\right)}{P\left(M\middle|N,I\right)}$$

$$=\frac{\sum_{\langle\mathbf{n}\rangle_{NM}}P\left(\mathbf{n}\middle|\boldsymbol{\theta},N,I\right)P\left(\boldsymbol{\theta}\middle|N,\mathbf{w},I\right)}{\sum_{\langle\mathbf{n}\rangle_{NM}}\int_{\boldsymbol{\theta}}P\left(\mathbf{n}\middle|\boldsymbol{\theta},N,I\right)P\left(\boldsymbol{\theta}\middle|N,\mathbf{w},I\right)d\boldsymbol{\theta}}$$

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

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

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

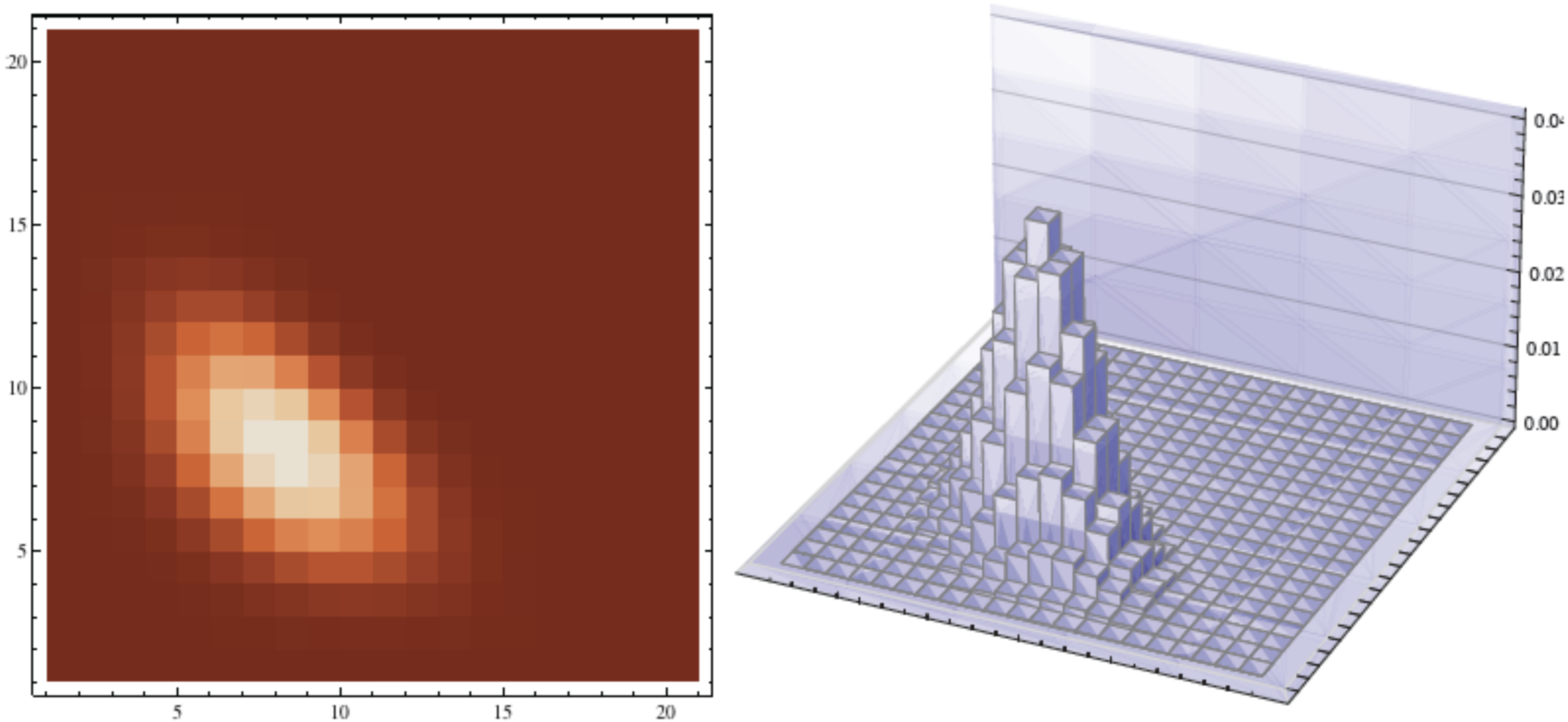


Figure 3: Multinomial distribution with $n = 20$, $k = 3$ and $p_1 = p_2 = p_3 = 1/3$, plotted as a function of the independent values n_1 and n_2. Density plot (left panel) and lego plot (right panel). As an exercise, explain why in this symmetrical case the distribution is not centered in the n_1,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

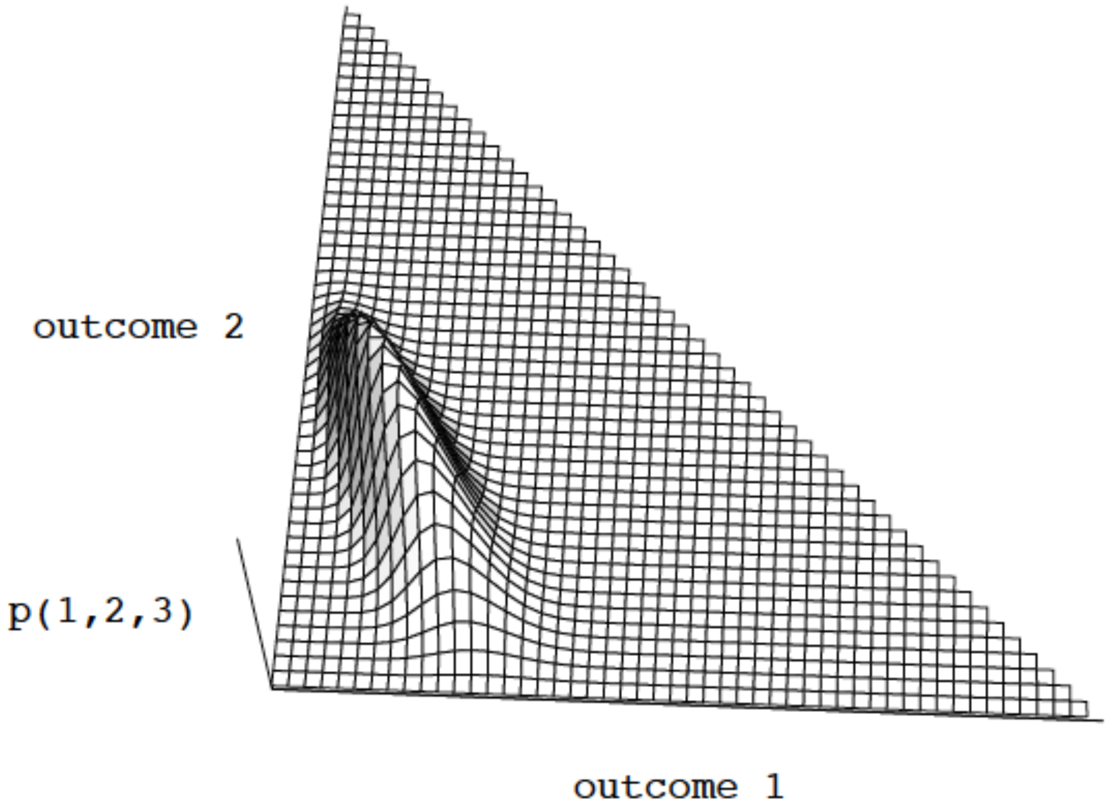


FIGURE 2. The posterior density for the 3-faces die example with a mean spot count of 2.5, N = 60, and prior weights of (1,1,1). Because of the normalization constraint, the third variable (not shown) is given by $\theta_3 = 1 - \theta_1 - \theta_2$.

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.The only information inside the subspace comes from priors.

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 2x2x7x3 = 84 entries

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

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

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

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

with the likelihood

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

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

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

in the original kangaroo problem

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

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

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

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

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

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

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

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

this solution coincides with the independence hypothesis

In the extended kangaroo problem we find

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

$$\Rightarrow \quad \begin{aligned} \theta_{1,j,k,l} &= \exp\left(\lambda_0 + \lambda_1 - 1\right) \\ \theta_{2,j,k,l} &= \exp\left(\lambda_0 + \lambda_2 - 1\right) \end{aligned}$$

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”.

AutoClass

+ Home
+ AutoClass C
+ 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

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

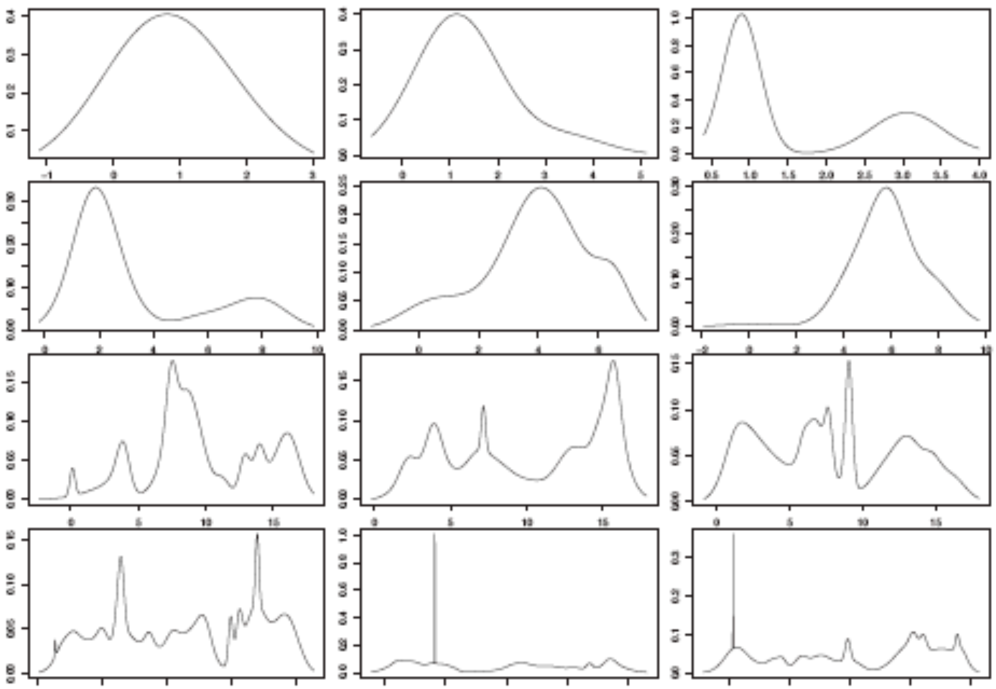


FIGURE 1. Some normal mixture densities for $K = 2$ *(first row)*, $K = 5$ *(second row)*, $K = 25$ *(third row)* and $K = 50$ *(last row)*.

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

there is a variable number of classes

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

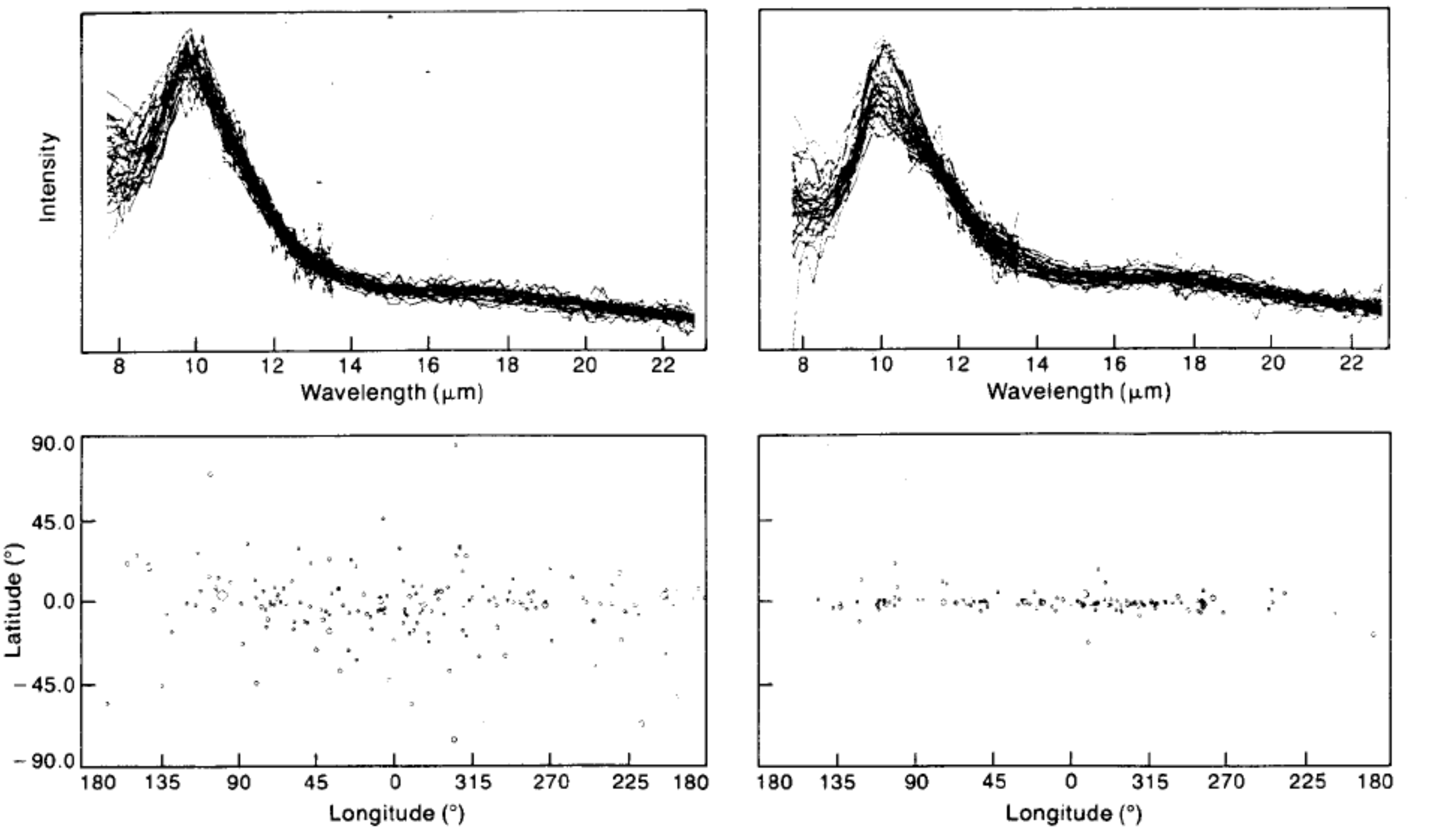
the component distributions are taken from a set of predefined distributions

the parameters define the shape of the component distribution

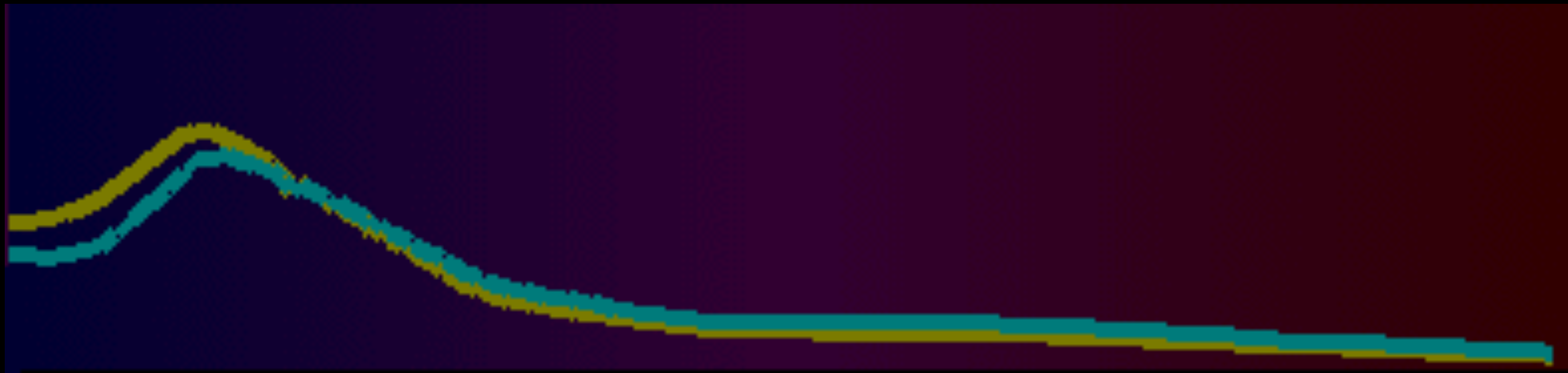
AUTOCLASS chooses a distribution and a parameter set for each class. Every data set determines a likelihood, and therefore a posterior distribution.

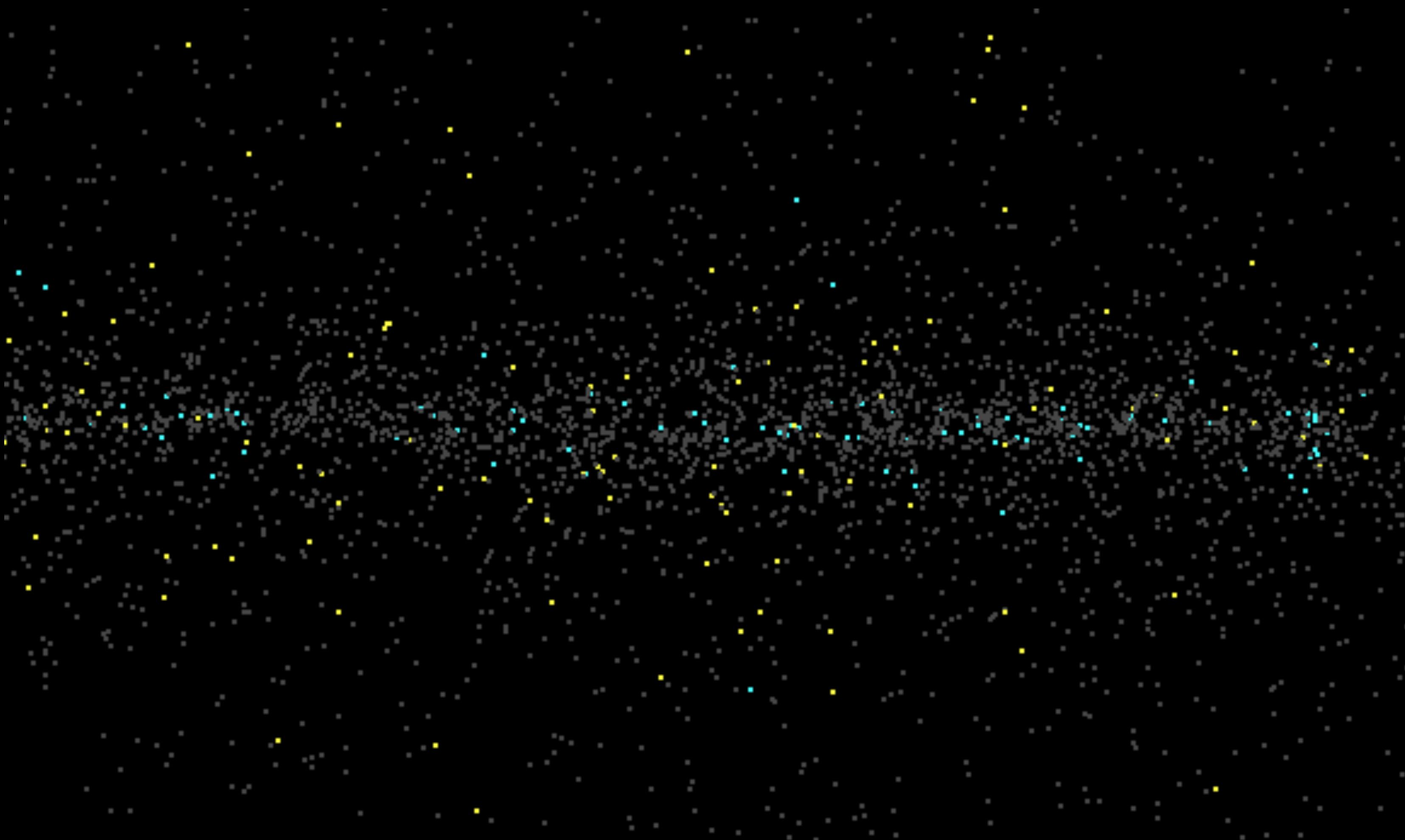
The class is selected by maximizing the posterior probability (MAP class estimate).

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

Welcome to AutoClass@IJM
the webserver for AutoClass Bayesian clustering system.

Developped by F. Achcar[1,2] and D. Mestivier[1] in collaboration with J.M. Camadro[2]

We kindly ask users to cite this paper when publishing results derived of the use of AutoClass@IJM.

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: draft of new chapter for "Machine Learning", http://www.cs.cmu.edu/%7Etom/NewChapters.html
- AUTOCLASS @ NASA: http://ti.arc.nasa.gov/tech/rse/synthesis-projects-applications/autoclass/
- AUTOCLASS @ IJM: http://ytat2.ijm.univ-paris-diderot.fr/