MAP METHODS FOR MACHINE LEARNING

Mark A. Kon, Boston University
Leszek Plaskota, Warsaw University, Warsaw
Andrzej Przybyszewski, McGill University

Example: Control System

1. Paradigm: Machine learning of an unknown input-output function

Example: Control system

Seeking relationship between inputs and outputs in industrial chemical mixture

Example: Control System

Given: We control chemical mixture parameters, e.g.:

- temperature $= x_1$
- ambient humidity $= x_2$, along with other non-chemical parameters x_3, x_4, x_5
- proportions of various chemical components $= x_6, \ldots, x_{20}$

Goal: Control output variable y = ratio of strength to brittleness of resulting plastic

We want machine which predicts y from $\mathbf{x} = (x_1, \ldots, x_{20})$ based on data from finite number of experimental runs of equipment.

$\Rightarrow$ want "best" f so

$$y = f(x_1, \ldots, x_{20}) + \epsilon = f(\mathbf{x}) + \epsilon$$

with minimal error ϵ.

MAPN approach

2. Solution: MAPN (MAP for Nonparametric machine learning)

Maximum A Posteriori (MAP) methods common for parametric statistical problems (see below).

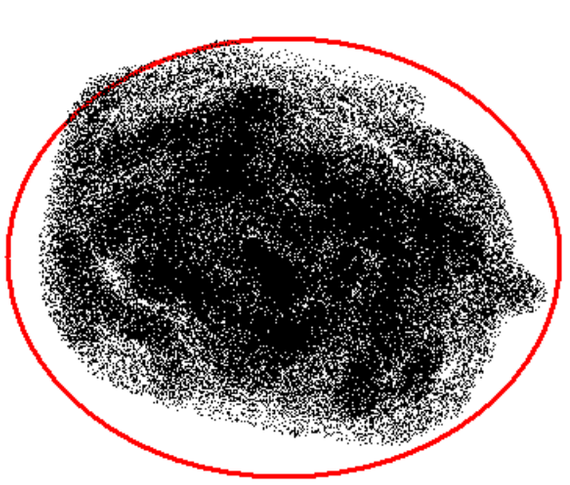
MAPN (MAP for Nonparametric machine learning) extends these methods to Nonparametric problems.

- Method simple, intuitively appealing (even in high dimension)
- Incorporation of prior knowledge explicit and transparent
- Results of method often coincide with standard methods, e.g., statistical learning theory (SLT), information-based complexity (IBC), etc..

MAPN approach

Bayesian machine learning strategy:

Use prior knowledge about f to choose reasonable probability distribution $dP(f)$ on set F of possible f.



Then combine prior knowledge with experimental data.

MAPN approach

Model:

Experimental data y_i satisfy

$$y_i = f(\mathbf{x}_i) + \epsilon_i,$$

with $\epsilon_i =$ Gaussian random error.

Equivalently,

$$\mathbf{y} = N(f) + \boldsymbol{\epsilon},$$

where

$$\mathbf{y} = (y_1, \ldots, y_n)$$

$$N(f) = \text{information vector } = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)).$$

MAPN approach

Standard strategy:

Compute conditional expectation

$$E(f|Nf = \mathbf{y}) \qquad (*)$$

as best guess of f.

Difficulties of the strategy:

- we have "infinite dimensional" parameter space F
- difficult to determine "reasonable" a priori probability measure P
- expectation (*) above hard to compute.

3. Sidebar: maximum a Posteriori (MAP) solutions yield a useful strategy in parametric statistics (finite dimension):

Finite dimensional method involves writing probability measure P for unknown parameter $\mathbf{z}$ as

$$dP(\mathbf{z}) = \rho(\mathbf{z})\, d\mathbf{z},$$

where $d\mathbf{z}$ = "uniform distribution" on $\mathbf{z}$, and

$$\frac{dP(\mathbf{z})}{d\mathbf{z}} = \rho(\mathbf{z})$$

is density function of $\mathbf{z}$.

MAP procedure finds

$$\widehat{\mathbf{z}} = \text{maximizer of } \rho(\mathbf{z})$$

(analogously to maximum likelihood estimation) as best estimate.

MAP Example: OCR

Example of MAP (discrete case):

Decide which letter ℓ_0 (a letter of the alphabet) we are looking at in an OCR program.

A priori information: overall probability distribution $P(\ell)$ on 26 letters ℓ in alphabet.

MAP Example: OCR

Data: Vector $\mathbf{z} = (z_1, \ldots, z_8)$ of 8 features of viewed letter.

Assume true letter is $\ell_0 = G$

$$\widehat{\ell}_0 = \text{MAP estimate of true letter } \ell_0 = \arg\max_{\ell} P(\ell|\mathbf{z}),$$

where ℓ ranges over alphabet. Here $P(\ell|\mathbf{z})$ is computed using Bayes' formula:

$$P(\ell|\mathbf{z}) = \frac{P(\mathbf{z}|\ell)P(\ell)}{\sum_j P(\mathbf{z}|j)P(j)} = CP(\mathbf{z}|\ell)P(\ell)$$

($C =$ denominator term is independent of ℓ).
Note $P(\mathbf{z}|\ell)$ is known, since we know which features occur in which letters.

MAP Example: OCR

Thus

$$\widehat{\ell} = \arg\min_{\ell} P(\mathbf{z}|\ell) P(\ell)$$

4. Extending MAP to non-parameteric statistics:

Can we use MAP to discover an entire function $f(\mathbf{x})$?

Recall: input data points are $\mathbf{x}_i$, output are y_i, model is

$$y_i = f(\mathbf{x}_i) + \epsilon_i = (Nf)_i.$$

Strategy:

Goal: Solve for unknown dependence $y = f(\mathbf{x})$ using above model $\mathbf{y} = Nf + \boldsymbol{\epsilon}$.

Prior knowledge: reflected in probability distribution $dP(f)$ on space $F =$ possible choices of f.

Extending MAP: Nonparametric case

Density function: Define a density function $\rho(f)$ for the distribution $dP(f)$, i.e., so that

$$dP(f) = \rho(f)df.$$

Algorithm: maximize $\rho(f|\mathbf{y})$, i.e., density conditioned on data $\mathbf{y} = Nf$.

Problem: There is no "uniform distribution" df on a function space such as F.

Remark: finding $\rho(f)$ is difficult part here - probability density in function space not easy to define!

Solution: **MAPN theorem** (below)

Extending MAP: Nonparametric case

Remark: The function $\rho(f)$ plays the role of a likelihood function in statistics -

- the larger $\rho(f)$ the more "likely" f is
- intuitively appealing
- easily interpretable
- very easily modifiable as prior information or intuition warrants
- nevertheless, $\rho(f)$ always corresponds to a genuine probability distribution $dP(f)$ on functions.

Algorithm: details

5. Details of the algorithm:

Note we have

$$\mathbf{y} = Nf + \boldsymbol{\epsilon} \;\Rightarrow \boldsymbol{\epsilon} = \mathbf{y} - Nf.$$

MAPN estimate is (using continuous Bayes formula)

$$\arg\max_f \rho(f|\mathbf{y}) = \arg\max_f \frac{\rho_{\mathbf{y}}(\mathbf{y}|f)\rho(f)}{\rho_{\mathbf{y}}(\mathbf{y})} = C_1 \rho_{\mathbf{y}}(\mathbf{y}|f)\rho(f).$$

(here $C_1 = \frac{1}{\rho_{\mathbf{y}}(\mathbf{y})}$ is independent of f).

Choose prior distribution $dP(f)$ for f to be Gaussian with covariance operator $C = (A^*A)^{-1}$, with A given below.

Algorithm: details

To define A: let

$$\mathbf{a} = (a_1, \ldots, a_{20}) = (.1, \ldots, .1, .2, \ldots, .2)$$

be a vector which determines strength of a priori information for each component x_i.

Here **a** has

$$\text{first 5 components} = .1$$

$$\text{last 15 components} = .2$$

reflects lower dependence of a priori assumptions on first 5 parameters (i.e., temperature, humidity, other non-chemical parameters);

greater on the last 15 (chemical parameters).

Algorithm: details

Choose covariance matrix of Gaussian to be the operator $C = (A^*A)^{-1}$, where

$$A = \text{"regularization operator"} = d(\mathbf{x})(1 + \mathbf{aD})d(\mathbf{x}),$$

with

$$\mathbf{aD} = \sum_{i=1}^{20} a_i \frac{\partial^{30}}{\partial x_i^{30}}.$$

Algorithm: details

Here $d(\mathbf{x})$ reflects density of sample points on $\mathbb{R}^{20}$ via

$$d(\mathbf{x}) = (1 + \delta(\mathbf{x}))^{-1},$$

where

$$\delta(\mathbf{x}) = \text{smoothed density of sample points}$$

$$= \left(1 + \sum_{k=1}^{n} e^{-(\mathbf{x}-\mathbf{x}_k)} \right)^{1/20}.$$

Note: order 30 above reflects smoothness level we expect of solution function f; note in 20 dimensions we need at least 10 derivatives for f to be continuous.

Thus distribution P concentrated on smooth solutions f; minimizing solution given by radial basis functions (see below)

Algorithm: details

Then **MAPN theorem** (below) shows probability distribution $dP(f)$ has density function $\rho(f) = C_2 e^{-\frac{1}{2}\|Af\|^2}$.

If we assume error vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ is Gaussian:

$$\rho_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) = C_3 e^{-\|B\boldsymbol{\epsilon}\|^2} = C_3 e^{-\|B(Nf-\mathbf{y})\|^2}.$$

(with $B = n \times n$ covariance matrix), then:

$$\rho(f|\mathbf{y}) = \frac{\rho_{\mathbf{y}}(\mathbf{y}|f)\rho(f)}{\rho_{\mathbf{y}}(\mathbf{y})} = C_3 \frac{e^{-\|B(N(f)-\mathbf{y})\|^2} e^{-\|Af\|^2}}{\rho_y(y)}$$

$$= C_4 e^{-\|B(Nf-\mathbf{y})\|^2} e^{-\|Af\|^2}$$

$$= C_4 e^{-(\|Af\|^2 + \|B(Nf-\mathbf{y})\|^2)}.$$

Algorithm: details

Thus maximizer of $\rho(f|\mathbf{y})$ is

$$\widehat{f} = \arg\min_f \|Af\|^2 + \|B(Nf - y)\|^2 \qquad (*)$$

$$= \sum_{k=1}^{n} c_k G(\mathbf{x}, \mathbf{x}_k) \qquad (**)$$

where

$G(\mathbf{x}, \mathbf{x}')$ = radial basis function = Green's function for operator A

$$= \mathcal{F}\left(\left(1 - \sum_{i=1}^{20} a_i^2 (i\xi_i)^{30}\right)^{-1}\right)(\mathbf{x} - \mathbf{x}'),$$

where ξ_i = Fourier variable dual to x_i.

Algorithm: details

Note: (*) is same functional appearing in regularization solutions in SLT, and the solution (**) is the same as spline solution in IBC.

Solution of problem:

Choose $\widehat{f}$ as in (**) for best approximation of i-o relationship $y = f(\mathbf{x})$

6. How does it work?

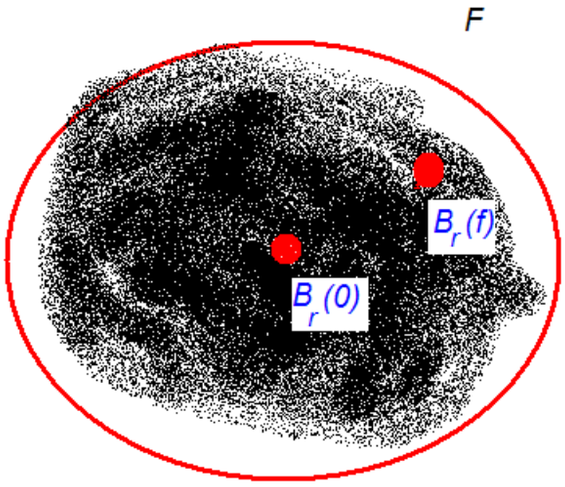
How do we define density $\rho(f)$ corresponding to prior probability distribution $dP(f)$ on the set of all functions?

Analogously to standard parametric MAP methods.

Use a definition of probability density $\rho(f)$ corresponding to probability distribution $dP(f)$ which works for both parametric (finite dimensional) and nonparametric (function) spaces.

Define $\rho(f)$ to be proportional to the ratio $\frac{P(B_r(f))}{P(B_r(0))}$ of probabilities of balls of radius r near f and 0, in the small r limit.

How does it work?



This limit exists for most generic measures in infinite dimension, including Gaussian measures:

How does it work?

MAPN Theorem: **(a)** The limit $\rho(f) = \lim_{r \to 0} \frac{P(B_r(f))}{P(B_r(0))}$ exists for a Gaussian measure P on a function space F with covariance operator C, defining its density function.

(b) If $C = (A^*A)^{-1}$ as above

$$\rho(f) = Ke^{-\frac{1}{2}\|Af\|^2}$$

In finite dimensions this reduces to ordinary density function of the Gaussian distribution.

http://math.bu.edu/people/mkon

Thank you!