ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI SCIENZE BIOLOGICHE, GEOLOGICHE E AMBIENTALI

DIPARTIMENTO DI SCIENZE STATISTICHE "P. FORTUNATI"

David Draper
Department of Applied Mathematics and Statistics, University of California, Santa Cruz,
and eBay Research Labs, San Jose CA

2 Dicembre 2013
AULA III, ORE 11:00-13:00
Dipartimento di Scienze Statistiche
Via Belle Arti, 41 - 40126 Bologna

Statistics Seminar 2013
Bayesian model specification: toward a Theory of Applied Statistics

Abstract: In the Bayesian statistical paradigm, when You (a person wishing to reason sensibly in the presence of uncertainty) are solving a problem $\mathcal{P}$ involving inference, prediction and/or decision-making, You begin with three ingredients: an unknown θ, a data set D, and a finite set of (true/false) propositions $\mathcal{B}$, all regarded by You as true, exhaustively describing the context of the problem $\mathcal{P}$ and the process that lead to the data set D.

A theorem of de Finetti and the American physicist RT Cox then says that, if You wish to quantify Your uncertainty about θ in a logically-internally-consistent manner, one way to accomplish this goal is to specify

(a) two probability distributions for inference and prediction, namely $p(\theta|\mathcal{B})$ and $p(D|\theta\,\mathcal{B})$, to quantify Your information about θ external and internal to D, respectively, and

(b) two additional ingredients for decision-making, namely Your action space $\mathcal{A}$ and Your utility function $U(a, \theta^*)$, which quantifies the costs and benefits arising from the choice of action a if the unknown θ took on the value θ^*.

Having specified these four ingredients, which collectively form Your *model* $M = \{p(\theta|\mathcal{B}),\, p(D|\theta\,\mathcal{B}), \mathcal{A}, U(a,\theta)\}$ for Your uncertainty about θ,

(1) the inference problem is solved with Bayes's Theorem,

$$p(\theta|D\,\mathcal{B}) \propto p(\theta|\mathcal{B})\,p(D|\theta\,\mathcal{B})\,,$$

(2) the prediction problem is solved with the equation

$$p(D^*|D\,\mathcal{B}) = \int_\Theta p(D^*|\theta\,D\,\mathcal{B})\,p(\theta|D\,\mathcal{B})\,d\theta\,,$$

in which D^* is future data and Θ is the set of all possible θ values, and

(3) the decision problem is solved with the equation

$$a^* = \text{argmax}_{a\in\mathcal{A}} \int_\Theta U(a,\theta)\,p(\theta|D\,\mathcal{B})\,d\theta\,,$$

in which a^* is the optimal action.

There are two principal challenges in implementing this program in real problems:

(I) (technical) the integrals in the three equations above can be difficult to approximate, and

(II) (substantive, and more serious) the de Finetti/RT Cox theorem gives little guidance on how to specify the four ingredients in Your model M.

It would be nice if the context of the problem $\mathcal{P}$ You're solving would uniquely identify M (this could be regarded as an instance of *optimal model specification*), but this is unfortunately rarely true; in practice You generally have to fall back on common sense and basic principles to aid You in the model specification process. Developing a logical progression from principles to axioms to model specification theorems could be said to constitute a *Theory of Applied Statistics*, which (as a profession) we need but do not yet have. In this talk I will (i) identify five basic principles that seem foundational to model specification, (ii) illustrate settings in which optimal model specification is possible, and (iii) discuss what to do when it is not.

Organizzatori
Prof.ssa Daniela Cocchi (Dipartimento Scienze Statistiche "P. Fortunati"),
Prof. Ezio Todini (Dipartimento di Scienze Biologiche, Geologiche e Ambientali)

La S.V. invitata