Investigations on model comparison and selection [draft]

P.G.L. Porta Mana
Kavli Institute, Trondheim, Norway
<pgl@portamana.org>

***; updated 2 September 2019

Some questions and investigations on model comparison and selection, based on previous work on fMRI data and health conditions (Porta Mana et al. 2018).
Note: Dear Reader & Peer, this manuscript is being peer-reviewed by you. Thank you.

1 Probability models

Consider the following scenario:

New software will be put into use in several clinical centres, for use in diagnosis of schizophrenia (or some other brain disease or condition). The software must be designed to give clinicians the *likelihood* that a subject has one of two health conditions, given fMRI data of some kind recorded from the subject. In other words, the software must calculate the numerical value of

$$\mathrm{p}(\text{fMRI data} \mid \text{health condition}, \text{pre-test info}) \tag{1}$$

which will be used by the clinician together with the likelihoods from other tests and her pre-test probabilities for the health conditions, to arrive at a final probability for the health conditions, to be used to decide upon treatment, dismissal, or other actions[1].

We have to prepare and deliver such software, using two possible parametric learning models to build it and some data, consisting in pairs of fMRI recordings and health conditions, to train it.

There are three possible sub-scenarios:

(S1) The software doesn't have the ability to update its decisions based on new data the clinician acquires during its actual use, and we must choose one particular *non-learning* model among those constituting our two learning models (the latter are a convex combinations of the former).

[1] Porta Mana et al. 2018 § 1.

(S2) The software doesn't have the ability to update its decisions based on new data the clinician acquires during its actual use, and we must choose one of the two learning models, which will be used in its final state of training.

(S3) The software will be able to update its decisions based on new data the clinician acquires during its use. We must still choose one of the two learning models.

For each sub-scenario we ask two questions: which model must we choose for the software? And how to make such a decision?

And if we had the possibility of choosing one of the three sub-scenarios, which should we choose?

This problem will be approached using Bayesian probability theory and decision theory, but some ad hoc approaches (for example cross-validation) will also be examined.

The purpose of this investigation is not to give a final answer to the questions above, but rather to bring to light all the different factors that enter this complex problem. I'll first also show the answer that's straightforwardly given by the probability calculus – and which is the simplest – to compare it to the other answers we obtained.

1.1 Solution from the probability calculus

The direct application of the probability calculus and decision theory to our scenario gives a straightforward answer to our problem.

Consider the clinician being visited by an actual subject in the future. She will make a general assessment of the subject's brain health, taking also into account gender, age, family history, lifestyle, environment, and similar factors. She will order some tests, including the fMRI scan. She will also have background knowledge about schizophrenia and also a statistics of the subjects she examined in the past. Let's denote groups of

these pieces of information symbolically:

$$I_{\text{gen}} \coloneqq \text{general background information, including statistics of past schizophrenic and healthy subjects,} \tag{2a}$$

$$I_{\text{subj}} \coloneqq \text{information about the subject gathered before clinical tests} \tag{2b}$$

$$D_{\text{tests}} \coloneqq \text{results from tests, excluding fMRI} \tag{2c}$$

$$x \coloneqq \text{fMRI test result} \tag{2d}$$

$$S \coloneqq \text{'the subject suffers from schizophrenia'} \tag{2e}$$

Bibliography

('de *X*' is listed under D, 'van *X*' under V, and so on, regardless of national conventions.)

Porta Mana, P. G. L., Bachmann, C., Morrison, A. (2018): *Inferring health conditions from fMRI-graph data*. Open Science Framework doi:10.17605/osf.io/r2huz, bioRxiv doi:10.1101/295113, arXiv:1803.02626.