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

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

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 probability models, M_1 and M_2, to build it and some data, consisting in pairs of fMRI recordings and health conditions, to train it. Each model is a mixture of parametric *non-learning* models.

There are three possible sub-scenarios:

(S1) The software doesn't have the ability to take into account new data that the clinician acquires during its actual use, and we must choose one particular *non-learning* model among those constituting each of our two models.

[1] Porta Mana et al. 2018 § 1.

(S2) The software doesn't have the ability to take into account new data that 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 is able to take into account new data that 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 have background knowledge about schizophrenia and also a statistics of the subjects she examined in the past. She will order some tests, including an fMRI scan. Other implicit information in the problem is the set of training data used for our software and, possibly, recorded fMRI results from all or some of the subjects examined by the clinician in the past. Let's denote these pieces of information symbolically:

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

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

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

$$X_{\text{train}} \coloneqq \text{fMRI training data} \tag{2d}$$

$$X_{\text{past}} \coloneqq \text{clinician's fMRI past data} \tag{2e}$$

$$X \coloneqq \text{fMRI test result} \tag{2f}$$

$$H \coloneqq \text{'the subject is healthy'} \tag{2g}$$

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

Then, to make a decision about the subject, the clinician needs a utility matrix and the post-test probabilities

$$\begin{gathered}\mathrm{P}(H \mid X,\ X_{\text{past}},\ X_{\text{train}},\ I_{\text{tests}},\ I_{\text{subj}},\ I_{\text{gen}})\\ \mathrm{P}(S \mid X,\ X_{\text{past}},\ X_{\text{train}},\ I_{\text{tests}},\ I_{\text{subj}},\ I_{\text{gen}}).\end{gathered} \tag{3}$$

Each of these can be calculated with Bayes's theorem, for example

$$\begin{multlined}\mathrm{P}(H \mid X,\ X_{\text{past}},\ X_{\text{train}},\ I_{\text{tests}},\ I_{\text{subj}},\ I_{\text{gen}}) \propto\\ \mathrm{P}(X,\ I_{\text{tests}} \mid H,\ X_{\text{past}},\ X_{\text{train}},\ I_{\text{subj}},\ I_{\text{gen}}) \times\\ \mathrm{P}(H \mid X_{\text{past}},\ X_{\text{train}},\ I_{\text{subj}},\ I_{\text{gen}}).\end{multlined} \tag{4}$$

Now we make the reasonable assumption that the fMRI data from past subjects and training are irrelevant for the plausibility that this subject is healthy or not, if the subject's fMRI result isn't given. This follows from the reversed irrelevance via Bayes's theorem:

$$\begin{multlined}\mathrm{P}(H \mid X_{\text{past}},\ X_{\text{train}},\ I_{\text{subj}},\ I_{\text{gen}}) \propto\\ \mathrm{P}(X_{\text{past}},\ X_{\text{train}} \mid H,\ I_{\text{subj}},\ I_{\text{gen}}) \times \mathrm{P}(H \mid I_{\text{subj}},\ I_{\text{gen}}).\end{multlined} \tag{5}$$

The post-test probability (4) then becomes, separating the fMRI test from the others with the product rule,

$$
\begin{aligned}
\mathrm{P}(H \mid X,\ X_{\text{past}},\ X_{\text{train}},\ I_{\text{tests}},\ I_{\text{subj}},\ I_{\text{gen}}) \propto{}& \\
&\mathrm{P}(I_{\text{tests}} \mid X,\ H,\ X_{\text{past}},\ X_{\text{train}},\ I_{\text{subj}},\ I_{\text{gen}}) \times \\
&\mathrm{P}(X \mid H,\ X_{\text{past}},\ X_{\text{train}},\ I_{\text{subj}},\ I_{\text{gen}}) \times \\
&\mathrm{P}(H \mid I_{\text{subj}},\ I_{\text{gen}}).
\end{aligned}
\tag{6}
$$

The second probability on the right side,

$$
\boxed{\mathrm{P}(X \mid H,\ X_{\text{past}},\ X_{\text{train}},\ I_{\text{subj}},\ I_{\text{gen}})}
\tag{7}
$$

is what our software must provide, eq. (1). This expression answers the questions that we asked in our scenario:

- If our software can learn from data acquired during its use, they enter the expression above as X_{past}. If it can't learn, it means that X_{past} is unavailable and therefore drops out of the expression.
- If X_{past} past is unavailable, then the probability is the one that would be given at the end of the training: no non-learning model is chosen within the two learning models. The learning model has become a non-learning one, so to speak, because new data are discarded instead of being fed to it.
- No choice between the two probability models M_1 and M_2 is made either: our software should use their mixture, weighted by the probability that the training data (and past data, if available) give to each. In formulae,

$$
\begin{aligned}
\mathrm{P}(X \mid H,\ X_{\text{past}},\ X_{\text{train}},\ I_{\text{subj}},\ I_{\text{gen}}) ={}& \\
&\mathrm{P}(X \mid H,\ X_{\text{past}},\ X_{\text{train}},\ I_{\text{subj}},\ I_{\text{gen}},\ M_1) \times \\
&\mathrm{P}(M_1 \mid X_{\text{past}},\ X_{\text{train}},\ I_{\text{gen}}) + \\
&\mathrm{P}(X \mid H,\ X_{\text{past}},\ X_{\text{train}},\ I_{\text{subj}},\ I_{\text{gen}},\ M_2) \times \\
&\mathrm{P}(M_2 \mid X_{\text{past}},\ X_{\text{train}},\ I_{\text{gen}}).
\end{aligned}
\tag{8}
$$

But it may be the case that the software can only implement one of the two models, or only one of the parametric non-learning ones they're the mixtures of. In this case we must face a decision problem.

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.