Report on results and technical details of AD study

for Ingrid, Alexandra, & co.

Luca <pgl@portamana.org>

17 December 2021; updated 17 December 2021

Here are the main results about the 'predictive power' of features. The technical details of the inference and method are explained in § 2.

Here I report the results for Alexandra's FAQ-based features. I'll report the results for Ingrid's study soon (some calculations are still ongoing).

1 Main results

Our general problem: to prognose the future onset of AD as opposed to stable MCI – predictand binary variate `Subgroup_num_` – given the set of twelve features: `AGE`, `RAVLT_immediate`, `AVDEL30MIN_neuro`, `AVDELTOT_neuro`, `TRAASCOR_neuro`, `TRABSCOR_neuro`, `CATANIMSC_neuro`, `GDTOTAL_gds`, `LRHHC_n_long`, `LRLV_n_long`, `FAQ`. In the following I use shorter names for them.

The 'predictive power' of a set of features is measured using the *mutual information* between that set of features and the predictand variate, measured in *bits*. The mutual information can in this case range from 0 bit, representing a complete lack of predictive power (features and predictand are independent, so we can just as well flip a coin to make our prognoses); to 1 bit, representing perfect, deterministic prediction.

The operational meaning of mutual information is explained in §***.

The results are reported for three different calculation set-ups, denoted `narrow`, `broad`, `all`. The first two use only the training dataset and have different prior smoothness preferences for the inference: `narrow` and `broad`. The third uses `all` datapoints, with a broad prior smoothness preference. More details are given in §***. It should be noted that the probability calculation should always use all data, and no division between training and test set is necessary (probability theory is implicitly already doing all such possible divisions).

1.1 Results for individual features

This is the mutual information for each feature if it were used *individually* to make the prognosis, ranked from highest to lowest:

feature	mutual information/bit		
	narrow	broad	all
AVDEL30MIN	0.13	0.09	0.09
RAVLT	0.12	0.09	0.08
FAQ	0.10	0.07	0.06
AVDELTOT	0.07	0.05	0.06
LRHHC	0.04	0.03	0.03
TRABSCOR	0.04	0.03	0.03
CATANIMSC	0.03	0.02	0.02
TRAASCOR	0.03	0.02	0.01
LRLV	0.01	0.01	0.01
AGE	0.00	0.00	0.00
GDTOTAL	0.00	0.00	0.00

The rank is the same in all computations set-ups; the values agree within the first significant digit.

1.2 Results for joint feature set

This is the mutual information for the set of features, used *jointly*:

narrow	broad	all
0.30 bit	0.19 bit	0.18 bit

These values indicate a low predictive power. A mutual information of 0.2 bit means that in a set of N new prognoses $(1 + 0.2)\,N/2$ will be correct (true positives + true negatives) on average, with a 95% variability of $\pm\sqrt{(1-0.2)\,N}$. So in 100 new prognoses 60 ± 9 will be correct. Analogously, a mutual information of 0.3 bit means 65 ± 8 correct prognoses in 100 new cases.

Note that this value is an implicit characteristic of this set of features: no predictive algorithm can do better than this by using these features as predictors. Algorithms that do less are not fully using the information in the predictors; algorithms that happen to do more have only had a stroke of luck.

1.3 'Importance' of individual features when used jointly

We must define what we mean by 'importance' of a feature in the joint prediction. Information is a very non-additive quantity: we cannot say "this feature contributes x to the information, that feature contributes y" and so on, with $x + y + \cdots$ adding up to the total information. Said otherwise, the 'importance' or 'contribution' of a feature is *context-dependent*.

For example, imagine that that we have three features that jointly give us a given amount of predictive power. It may happen that if we do not use the first feature, our predictions are just as good; its contribution seems to be zero. If we do not use the second feature, our predictions are also just as good: its contribution seems to be zero as well. And yet, if we do not use neither the first nor the second feature, our predictions become worse. So '$0 + 0 \neq 0$' in a manner of speaking. A figurative explanation of this phenomenon is as follow: the first feature provides pieces of information AB, the second feature provides BC, the third provides AC. Jointly they provide ABC. If we drop the first, we still have ABC from the remaining two. If we drop the second, we still have ABC. But if we drop the first and the second, we only have AB; our prediction become worse because of lack of the piece of information C.

From these considerations, my approach to somehow quantifying the 'importance' of feature X is as follow:

1. calculate the mutual info of all features, used jointly; denote the result by I
2. calculate the mutual info of all features but excluding X; denote the result by $I_{\setminus X}$
3. calculate the relative decrease of mutual info we suffer when we remove X from the features:

$$\Delta I_{\setminus X} \coloneqq \frac{I - I_{\setminus X}}{I} \equiv 1 - \frac{I_{\setminus X}}{I} \tag{1}$$

Note that by the properties of mutual info we always have $I \geqslant I_{\setminus X}$, so $0 \leqslant \Delta I_{\setminus X} \leqslant 1$.

I define the 'importance' of a feature X as the relative decrease in mutual information we suffer when X is removed from the whole set of

twelve features used for prediction. It seems a sensible intuitive definition to me, with values that have a concrete operational meaning (see below).

With this definition, here are the 'importances' of the individual features, ranked from highest to lowest according to the `all` setup:

feature	$\Delta I_{\setminus \text{feature}}$		
	`narrow`	`broad`	`all`
`TRABSCOR`	14%	12%	13%
`FAQ`	11%	11%	9%
`TRAASCOR`	6%	3%	4%
`AVDEL30MIN`	4%	5%	4%
`RAVLT`	2%	2%	3%
`AVDELTOT`	2%	2%	2%
`LRHHC`	1%	1%	1%
`CATANIMSC`	1.0%	1.1%	0.7%
`GDTOTAL`	0.7%	0.4%	0.5%
`LRLV`	0.1%	0.4%	0.3%
`AGE`	0.3%	0.4%	0.3%

Note that the percentages do not add up to 100%, nor should they, owing to the reasons given above.

We see that omitting either `TRABSCOR_neuro` or `FAQ` from the twelve features would reduce the mutual information by around 10%; that is, from a value of 0.2 bit to 0.18 bit. This also means that in 100 new prognoses we would drop from 60 ± 9 correct ones to around 59 ± 9. In clinical-importance terms this difference is not small: it means on average 10 000 additional incorrect predictions every million prognoses.

1.4 Further remarks

The mutual informations estimated above are calculated assuming that *we will not receive further training data*. It is also possible to calculate an estimate of what the mutual informations would be if we had a very large number of training data. I'm currently setting up the code to calculate this estimate.

Regarding 'explanation', I categorically refuse to say that the mutual info or some other score of a feature 'explains' anything. Statistical and probabilistic analyses and metrics do not 'explain', they only *fit*. Using 'explanation' in this context is an intentional or unintentional way of

making it sound as if we have discovered more than we actually have (the deceiving use of the word 'explanation' for what's just a *fit* has been a very bad habit of statisticians since the 1960s or 70s). Explanations can only be given by hardcore physics, chemistry, biology; but different levels of fit can of course suggest, or be evidence for, different explanations formulated within those disciplines.

2 Technical details

2.1 Technical terminology, intuitive understanding, and mathematical characteristics of the method

In the literature, the probabilistic calculations made here are usually called *Bayesian nonparametric density regression*. This terminology can be psychologically misleading because it sounds as if we are making peculiar assumptions, when in reality we are not: this is the most general probabilistic inference possible. Here's a simple explanation of the terminology.

Density regression stands opposed to *functional regression*. Functional regression makes the assumption that there exists a physical functional dependence between predictand and features, contaminated by (usually gaussian) noise. Examples of algorithms that do functional regression under the hood: linear regression, generalized linear models, support vector machines, gaussian processes, neural networks. Density regression does not assume that a functional relationship exists (it doesn't exclude it either). It simply evaluates the probabilities of observing different values of the predictand, given values of the features; that is, it evaluates probability *densities*, hence the name. Examples of algorithms that do density regression: logistic regression, random forests (I believe).

Nonparametric stands opposed to *parametric*. Parametric regression makes the assumption that the function or density has a specific form or shape (line, exponential; gaussian; and so on). Examples of parametric algorithms: linear regression, generalized linear models, support vector machines, logistic regression. Nonparametric regression does not make any such assumption; it considers all possible forms and shapes. Examples of nonparametric algorithms: gaussian processes, neural networks, random forests (I believe).

Nonparametric density regression is therefore the most general assumption-free inference we can possibly make.

2.2 Intuitive understanding

In simplified terms, probability theory in this inference first considers every possible frequency distribution of predictand given features, for *all future* patients. Such frequency distribution also tells us what we can guess about our next patient, given his/her features.

For example, suppose that by magic or some powerful technology we knew that, among *all* – or a billion of – future patients that have feature `FAQ` = 8, 68% of them will develop AD and 32% of them will stay MCI. If our next patient has `FAQ` = 8, then he/she must be either one of the 680 millions who will develop AD, or one of of the 320 millions who will not. So there's a 68% probability that he/she will develop AD. This consideration is made not only for `FAQ` but for all 12 features jointly, and all their possible values. This simplified understanding allows us to quickly interpret, for example, the solid red line in the plot of `FAQ` vs 'probability of AD' enclosed in this report.

Of course we do not know what the true frequency distribution of predictand and features will be. Probability theory considers each possible frequency distribution and attaches a probability – a weight, if you like – to it. This weight, called 'posterior', comes from two contributions:

'Likelihood' : How much the distribution fits the data we already have. This is simply the joint probability of our data, assuming that frequency distribution were the true one.

'Prior' : An initial weight based on biomedical considerations of how the true frequency could most likely look like.

In the present case we give a greater initial weight to distributions that are smooth, without lots of sudden jumps or a zigzag shape, although they may have as many large and small peaks as they like. This is biologically very plausible, owing to physical and chemical continuity. Note that we are not excluding discontinuities: we only saying that we at first we consider them less plausible.

The total probability or weight is closer to the prior one when we have very few data; but as the number of data increases, the preferences

expressed in the prior become less and less important. Below it's explained what kind of prior preferences correspond to the `narrow` and `broad` setups.

The first result of this inference is therefore a probability distribution over all possible frequency distributions for future patients. In other words, how their 'true' population distribution look like.

The second result is a probability for the next new patient we observe: given his/her features, the probability is an average over all candidate frequency distributions, weighted by their posteriors.

In summary, the probability calculation is guessing what the true distribution of predictand and features is, for the full population, and it also gives our uncertainty about such guess. It uses this guess to make predictions about the next and all subsequent patients.

In passing, it can be proven that the posterior mathematically includes an average of *all possible* k-fold cross-validations, for all k, and all possible divisions of the data into 'training' and 'test' sets[1]. This gives an idea of why probability calculations are so enormously computationally expensive.

[1] **portamana2019b**.