Report on results and technical details of AD study

for Ingrid, Alexandra, & co.

Luca <pgl@portamana.org>

17 December 2021; updated 3 January 2022

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).

All material is available at https://github.com/pglpm/ADBayes.

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 quickly explained in § 1.2 and somewhat more extensively and with references in § 2.4.

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 §§ 2.2–2.3. 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).

This document is designed for screen reading and two-up printing on A4 or Letter paper

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	0.1	0.09
RAVLT	0.12	0.09	0.08
FAQ	0.10	0.08	0.06
AVDELTOT	0.07	0.05	0.06
LRHHC	0.04	0.03	0.03
TRABSCOR	0.03	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. These values do not only give us a ranking, but also an estimate of the predictive power. See the next section for what these numbers actually mean.

1.2 Results for joint feature set

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

narrow	broad	all
0.31 bit	0.20 bit	0.18 bit

What does a value such as ‘0.2 bit’ actually mean?

A mutual information $I = 0.2$ bit between the predictand and a set of features gives us a range of the number of correct prognoses we can expect to do about the predictand by using that set of features.

In a set of N new prognoses, we have an approximate lower bound of $(1 + I)\,N/2 \pm \sqrt{(1 - I)\,N}$ correct ones (true positives + true negatives) on average, and an upper bound of $C\,N \pm 2\sqrt{(1-C)\,C\,N}$ correct ones (the interval is for a 95% certainty) on average, with C given by the formula

$$C = \frac{1}{2}\left[1 + \sqrt{1-(1-I)^{4/3}}\right] . \tag{1}$$

So, with a mutual information of 0.2 bit, in 100 new prognoses we can expect that between 60 ± 9 and 75 ± 9 will be correct. Analogously, a mutual information of 0.3 bit means from 65 ± 8 to 81 ± 8 correct prognoses in 100 new cases.

Note that this value is an implicit characteristic of this set of features: it is a mathematical result from information theory that 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:

$$\Delta I_{\setminus X} \coloneqq \frac{I - I_{\setminus X}}{I} \equiv 1 - \frac{I_{\setminus X}}{I} \tag{2}$$

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`	13%	10%	9%
`TRAASCOR`	7%	4%	4%
`AVDEL30MIN`	4%	4%	4%
`RAVLT`	3%	1%	3%
`AVDELTOT`	2%	2%	2%
`LRHHC`	1%	1%	1%
`CATANIMSC`	1%	1%	0.7%
`GDTOTAL`	0.7%	0.1%	0.5%
`LRLV`	0.0%	0.3%	0.3%
`AGE`	0.4%	0.5%	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 roughly 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 a best case of 75 ± 9 correct ones to a best case of around 74 ± 9. In clinical-importance terms this difference is not small: it means on average more than 10 000 additional incorrect predictions (more exactly around 12 500) 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.

• It is possible to calculate a more precise forecast of how many correct predictions are possible in new prognoses. Let me know if this is of interest.

• As shown in the enclosed plots (all coming from the `all` setup), many features have two clearly different population distributions for the future AD or MCI patients. This means that something is already at work affecting those features. However, the two population distributions have very large overlaps, so the feature does not help very much in making predictions for a single individual, as clear also from the mutual informations for the individual features.

• A further analysis of direct and inverse conditional frequencies in the population suggest that the dataset might have some biases for some features. These biases, however, could be overcome by appropriately combining inverse and direct predictions from the features. This point and topic would require a very extensive discussion, so I won't write any further details here.

• As discussed in § 2.3, the calculation also allows us to count the number of peaks in the joint distribution of predictand and features, possibly hinting at the presence of subpopulations. A very rough estimate gives around 10 main peaks. If this is of interest, a more exact calculation can be done.

• Regarding 'explanation', I categorically refuse to say that the mutual info or some other score of a feature 'explains' anything. For me, statistical and probabilistic analyses and metrics do not ‘explain’, they only *fit* or correlate. 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

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 some powerful technology we knew that, among *all* – or a billion of – future patients that have feature `FAQ` = 8, a percentage of 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 at the end of the present 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 continuity of physical and chemical processes. Note that we are not excluding discontinuities: we only consider them less plausible at first.

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. This is summarized in the following (slightly informal) formula:

$$\mathrm{p}(Y \mid X) = \int f(Y \mid X)\, \mathrm{p}(f \mid \mathrm{data})\, \mathrm{d}f \tag{3a}$$

with

$$\mathrm{p}(f \mid \mathrm{data}) \propto f(\mathrm{data})\, \mathrm{p}(f \mid H) \tag{3b}$$

where $f(Y \mid X)$ is the frequency of predictand value Y among patients with features X; the probability of our data given such frequency is $f(\mathrm{data})$; and $\mathrm{p}(f \mid H)$ is the prior weight of such frequency.

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.

2.3 Brief overview of mathematical aspects

From a mathematical point of view the main problem is how to represent a generic frequency distribution f presented in the intuitive explanation above. There are an infinity of mathematically equivalent representations.

[1] Porta Mana 2019.

Here we use the representation as an *infinite mixture of product kernels*: a generic joint frequency distribution $f(Y, X, Z, \dots)$ is represented as

$$f(Y, X_1, X_2, \dots) = \sum_{k=1}^{\infty} w_k \, Q_{s_k}(Y) \, Q_{a_k}(X) \, Q_{b_k}(Z) \cdots \tag{4}$$

with an infinite set of coefficients $\{w_k, s_k, a_k, b_k, \dots\}$. These coefficients effectively are a coordinate system on the infinite-dimensional manifold of frequency distributions[2]. The various Q distributions are gaussian for continuous variates, binomial for integer variates, and Bernoulli for binary variates.

This representation allows for quick calculations of conditional and marginal frequencies. For more precise mathematical details see Dunson & Bhattacharya (2011), and Rasmussen (1999) who only considers gaussian kernels though.

This mathematical representation also allows, in principle, an interpretation of the frequency distribution in terms of clusters, and an interpretation of the population in terms of subpopulations. Such interpretations, however, are arbitrary and contentious if we do not first justify why the subpopulations should have distributions like the Qs and not some other distributions – say, why gaussian and not Cauchy or Gumbel. I therefore do not trust any such subpopulation interpretation of the formula above; its meaning is only in the final sum.

However, the number of clusters with high weights w_k does give us a very rough idea of the number of major and minor peaks of the frequency distribution, and therefore a vague idea of the number of possible subpopulations.

In the representation above, the prior probability over the frequencies is represented by a probability density over the coefficients $\{w_k, s_k, a_k, b_k, \dots\}$. This probability density determines our initial guess about the smoothness of the true frequency distribution.

The `narrow` and `broad` setup roughly correspond to two the following guesses:

`narrow` : with 97% probability, the widths of individual peaks can be two orders of magnitude broader or narrower than the average width

`broad` : with 61% probability, the widths of individual peaks can be two orders of magnitude broader or narrower than the average width

[2] Choquet-Bruhat et al. 1996 esp. Part VII.

The general probability distribution for such peaks is a compound gamma distribution. The `narrow` setup therefore favours a frequency distribution that does not have too narrow or too broad peaks; the `broad` setup gives much more freedom. The results show that the data overwhelm either choice, so our conclusions are robust to variation in this kind of preference.

The computation of the probability for the different frequencies is done by Markov-chain Monte Carlo sampling, with Gibbs sampling; see Neal (1993) for an exhaustive discussion, and again Dunson & Bhattacharya (2011), Rasmussen (1999) and references therein for details, and MacKay (2005 ch. 29) for an overview. The computation has converged when the Markov chain has become stationary; stationarity has been checked for various quantities derived from the frequency distribution, such as the joint probability of the data, the means and variances, joint covariances, and full-dimensional covariance. Convergence required roughly 48–60 hours of computation.

2.4 Entropy and mutual information

For a comprehensive discussion of the following concepts see the first chapters of MacKay (2005), especially ch. 4, and also his lectures [3], besides Shannon (1948) and the very accessible summary in § 14.7 of Press et al. (2007). Other references about its predictive meaning and practical use are for example Lindley (1956), Kullback (1978), Woodward (1964). These notions are their units are also international standards: see ISO (2008), items 13-23–13-42.

The *entropy* of the distribution for a variate Y, here denoted $H(Y)$, represents our uncertainty about future occurrences of that variate. For a binary variate it assumes a value between 0 bit (complete certainty and exact prediction) and 1 bit (complete uncertainty). A value of H (in bits) means that we are on average uncertain about $H\,N$ future outcomes.

[3] `https://www.youtube.com/watch?v=BCiZc0n6COY&list=PLruBu5BI5n4aFpG32iMbdWoRVAA-Vcso6`

This means that we will on average make

$$\begin{aligned}&\text{from}\quad (1-H/2)\,N \pm \sqrt{H\,N}\\ &\text{to}\quad C\,N \pm 2\sqrt{(1-C)\,C\,N}\quad \text{correct predictions,}\\ &\qquad\qquad\text{with}\quad C = \frac{1}{2}\left[1+\sqrt{1-H^{4/3}}\right]\,. \end{aligned} \tag{5}$$

The entropy is also related to the size of the typical set of different sequences of values we can see in N occurrences: the size is 2^{H}.

When we know the *specific* value of a feature (or features) X, we can usually make better predictions about Y, reflected by a sharper probability distribution for Y given the specific value x. The entropy of this distribution is denoted $H(Y \mid x)$. The meaning is as before.

In considering the *general* predictive power of the feature X, some important considerations are necessary. For example, it can happen that our predictions are almost perfect (entropy of 0 bit) when the value of X is in some range R_{good}; whereas their are very bad (entropy of 1 bit) when the value is in the complementary range R_{bad}. Is then the feature X a good or a bad predictor? The answer heavily depends *on which values of X we'll encounter in our future applications*. If all future applications typically have values of X in R_{good}, then X is a good predictor; if they have values in R_{bad}, then X is a bad predictor; for intermediate cases we have intermediate predictive power. Clearly this depends on the *probability distribution of future occurrences of X, $p(X)$*.

The *conditional entropy* of Y given features X is exactly the weighted combination of the entropies given each value of X, weighted by their frequency of occurrence:

$$H(Y|X) \coloneqq \sum_{x} H(Y \mid x)\, p(X) \tag{6}$$

Note that any kind of metric of predictive performance that neglects $p(X)$ is therefore missing a very important point; its results will be very hit-or-miss. Conditional entropy correctly keeps this factor into account, instead; this is the reason why its values can be given immediate operational meaning in terms of prediction.

Finally, the *mutual information* between Y and X is simply the difference between our uncertainty when we do not know X and when we do:

$$I(Y, X) \coloneqq H(Y) - H(Y \mid X). \tag{7}$$

From this definition and the operational meaning of the entropy we also obtain a concrete operational meaning for the mutual information. In particular, if we are completely uncertain about Y in absence of features, $H(Y) = 1$ bit, then a mutual information I means that the features X allow us to make on average

$$\text{from}\quad (1+I)\,N/2 \pm \sqrt{(1-I)\,N}$$
$$\text{to}\quad C\,N \pm 2\sqrt{(1-C)\,C\,N}\quad\text{correct predictions,}$$
$$\text{with}\quad C = \frac{1}{2}\Bigl[1+\sqrt{1-(1-I)^{4/3}}\Bigr]\,. \tag{8}$$

out of N cases, as discussed in § 1.2. The exact number depends on the distribution of features (which can be calculated from the probabilistic analysis).

Bibliography

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

Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., West, M., eds. (2011): *Bayesian Statistics 9*. (Oxford University Press, Oxford). DOI: 10.1093/acprof:oso/9780199694587.001.0001.

Choquet-Bruhat, Y., DeWitt-Morette, C., Dillard-Bleick, M. (1996): *Analysis, Manifolds and Physics. Part I: Basics*, rev. ed. (Elsevier, Amsterdam). First publ. 1977.

Dunson, D. B., Bhattacharya, A. (2011): *Nonparametric Bayes regression and classification through mixtures of product kernels*. In: Bernardo, Bayarri, Berger, Dawid, Heckerman, Smith, West (2011): 145–158. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.178.1521, DOI:10.1093/acprof:oso/9780199694587.003.0005, older version at https://www.researchgate.net/publication/228447342_Nonparametric_Bayes_Regression_and_Classification_Through_Mixtures_of_Product_Kernels.

ISO (International Organization for Standardization) (2008): *ISO 80000-13:2008: Quantities and units 13: Information science and technology*. International Organization for Standardization.

Kullback, S. (1978): *Information Theory and Statistics*. (Dover, New York). Republ. with a new preface and corrections and additions by the author. First publ. 1959.

Lindley, D. V. (1956): *On a measure of the information provided by an experiment*. Ann. Math. Stat. **27**^{4}, 986–1005. DOI:10.1214/aoms/1177728069.

MacKay, D. J. C. (2005): *Information Theory, Inference, and Learning Algorithms*, Version 7.2 (4th pr.) (Cambridge University Press, Cambridge). https://www.inference.org.uk/itila/book.html. First publ. 1995.

Neal, R. M. (1993): *Probabilistic inference using Markov chain Monte Carlo methods*. Tech. rep. CRG-TR-93-1. (University of Toronto, Toronto). http://www.cs.utoronto.ca/~radford/review.abstract.html, https://omega0.xyz/omega8008/neal.pdf.

Porta Mana, P. G. L. (2019): *A relation between log-likelihood and cross-validation log-scores*. Open Science Framework doi:10.31219/osf.io/k8mj3, hal:hal-02267943, arXiv:1908.08741.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. (2007): *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. (Cambridge University Press, Cambridge). First publ. 1986.

Rasmussen, C. E. (1999): *The infinite Gaussian mixture model*. Adv. Neural Information Processing Systems (NIPS) **12**, 554–560. https://www.seas.harvard.edu/courses/cs281/papers/rasmussen-1999a.pdf.

Shannon, C. E. (1948): *A mathematical theory of communication*. Bell Syst. Tech. J. **27**$^{3,\,4}$, 379–423, 623–656. https://archive.org/details/bstj27-3-379, https://archive.org/details/bstj27-4-623, http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf.

Woodward, P. M. (1964): *Probability and Information Theory, with Applications to Radar*, 2nd ed. (Pergamon, Oxford). doi:10.1016/C2013-0-05390-X. First publ. 1953.