On permutation testing of classification accuracy for signal detection

Jonathan Rosenblatt *

*Ben-Gurion University of the Negev, Beer-Sheva, Israel

[TODO]

XXX | YYY | ZZZ

Introduction

A common workflow in genetics or neuroimaging consists of fitting a classifier, and estimating its predictive accuracy using cross validation. Given that the cross validated accuracy is a random quantity, it is then common to test if the cross validated accuracy is significantly better than chance using a permutation test. Genetic examples include [TODO: elaborate][1, 2]. Neuroscience examples include the recently popularized *multivariate pattern analysis* (MVPA) [3, 4, 5].

To fix ideas, we will adhere to a Neuroscientific example: In [10], the authors seek to detect auditory brain regions which distinguish between vocal and non-vocal stimuli. According to the MVPA analysis workflow, the localization problem is cast as a supervised learning problem: if the type of the stimulus can be predicted from the spatial activation pattern, significantly better than chance, then a region is declared to encode vocal/non-vocal information.

This same signal detection task can be also approached as a two-group multivariate test: Inferring that a region encodes vocal/non-vocal information, is essentially inferring that the spatial distribution of brain activations is different given a vocal/non-vocal stimulus. A practitioner may then call upon a two-group Hotelling test [6]. If the size of the brain region is too large compared to the number of observations, so that the spatial covariance cannot be fully estimated, then a high dimensional version of the Hotelling test can be called upon, such as [7].

At this point, it becomes unclear which is the preferred avenue. This question was indeed the subject of [8], which compared the Hotelling test to the classification accuracy of *Fisher's linear discriminant analysis* (LDA) [9]. Their asymptotic analysis concluded that methods are equivalent with respect to the order of convergence to a consistent test. The same authors did find that the methods differ in the constants governing the convergence.

Those constants, are the focus of this contribution. In particular, which signal detection method is to be preferred in finite samples? Our conclusion is quite simple: signal detection using multivariate tests almost always has more power than using the accuracy test statistic.

The main argument for our statement rests upon the observation that with typical sample sizes, the accuracy test statistic is highly discrete. Discrete test statistics are known to be conservative, since they cannot exhaust the permissible type I error rates. For the accuracy test statistic, the degree of discretization is governed by the number of samples. in our running example [10], the classification is performed based on 40 trials, so that the test statistic may assume only 40 possible values. This number of examples is not unusual if considering this is the number of subject in a genetic study, or the number of trial repeats in an fMRI brain scan.

The discretization effect is further augmented if the test statistic is highly concentrated. For some intuition consider the usage of the train-accuracy test statistic (i.e., not cross validated). Because the testing problem is high dimensional, the observed train accuracy will be high. But the same will occur in every permutation. The permutation accuracies will tend to concentrate around 1 with a granularity depending on the sample size. In a very high dimensional setup, the accuracy will be 1 for the observed data, but also for any permuted data set. The permutation p-value will thus be 1 for almost all samples, and the null will never be rejected.

In the following, we look into the conservativeness of signal detection using an the classification accuracy. We start by establishing a best practice for permutation testing using the accuracy test statistic,

We discuss the problem characteristics that govern the magnitude of the conservativeness, and try to offer an intuition to the scope of the observation that a multivariate test should always be preferred over a classification approach.

Problem setup

Adhering to our neuroscientific example, we now formalize terminology and notation. Let $y \in \mathcal{Y}$ be a class encoding. In our vocal/non-vocal example, using effect coding, we have $\mathcal{Y} = \{-1, 1\}$. Let $x \in \mathcal{X}$ be a p dimensional feature vector. In our vocal/non-vocal example p is governed by the number of voxels in a regions, which is the number of voxels in each brain region tested. We thus have $\mathcal{X} = \mathbb{R}^{27}$.

Given n pairs of (x_i, y_i), typically assumed i.i.d., the *testing* approach to localization amounts to testing whether $x|y = 1$ has the the same distribution as $x|y = -1$. I.e., the multivariate voxel activation pattern has the same distribution when given a vocal stimulus, as when given a non-vocal stimulus. The *classification* approach to the localization problem amounts to learning a predictive model $\hat{f}(x)$ from some assumed model class $\hat{f} \in \mathcal{F}$. The prediction accuracy, denoted $T_{\hat{f}}^{acc}$, is defined as the probability of a given classifier $\hat{f}$ of making a correct prediction $T_{\hat{f}}^{acc} := \mathbf{E}\left[I_{\{\hat{f}(x)=y\}}\right]$ when given a new, randomly drawn data point, (x, y).

Reserved for Publication Footnotes

How to perform the permutation test?

To discuss the power of a permutation test using the accuracy test statistic, we need to establish some best practices for such a test. The design of the test consists of the following choices:

What test statistic?

Cross validated or not? Is the statistic cross validated or not?

Refolding? For a K-fold cross validated test statistic: is the data refolded in each permutation?

Permute labels of features? Should the y be permuted or should the x?

Balanced folding? For a K-fold cross validated test statistic: is the data folding balanced? (a.k.a. stratified).

How many folds? We will now address these questions while bearing in mind that unlike the typical supervised learning setup, we are not interested in an unbiased estimate of the prediction error, but rather in the mere detection of a difference between two groups, leading to a better-than-chance accuracy.

What test statistic? Given a predictor $\hat{f}$, a natural test statistic is some estimate of its accuracy $T_{\hat{f}}^{acc}$. Then again, very low accuracies, even 0, is evidence that the classes are separated, and we only need to invert the predictions. We can thus consider some estimate of $|T_{\hat{f}}^{acc} - 0.5|$ as the test statistic. This, however, implies that if the classes are identical, random guessing has a 0.5 accuracy. This is not true if the classes are not balanced. The chance level in which case is the prevalence of the dominant class, we denote by $\hat{p}_{max}$. This suggests the following test statistic $|T_{\hat{f}}^{acc} - \hat{p}_{max}|$. Since we will latter be aggregating these statistic over random data foldings, where the dominant class may have varying frequencies, it seems appropriate to standardize the scale of this statistic. We thus also consider a z-scored accuracy: $|T_{\hat{f}}^{acc} - \hat{p}_{max}|/\sqrt{\hat{p}_{max}(1 - \hat{p}_{max})}$.

Cross validated or not? Were we interested in an unbiased estimator of the prediction error, there is no question that some validation is in order. Since we are merely interested in detecting a difference between groups, a biased error estimate is not an issue provided that it is consistent over all permutations. The underlying intuition is that if the exact same computation is performed over all permutations, then a permutation test will be "fair", i.e., will not inflate the false positive rate. We will thus be considering both cross validated accuracies, and train accuracies as our test statistics.

Refolding? The standard practice in neuroimaging is to refold the data after each permutation [TODO:ref]. This is imperative if permuting labels while aiming at balanced data folds. This is not, however, imperative in general. In this work, we will adhere however to the standard practice of refolding the data within each permutation.

Permute labels of features? While seemingly identical, the compounding of permutations with data foldings renders these two approaches distinct. As an example, consider balanced (stratified) K-fold cross validation where the initial data folding is balanced. After a label permutation, the folds will probably not be balanced, and will thus have to be refolded. If the features are permuted, then the labels conserve their original fold assignments, and the data need not be refolded.

Balanced folding? The standard practice when cross validating is to constrain the data folds to be balanced (i.e. stratified)[TODO: ref]. This is well justified when aiming at unbiased accuracy estimation. This also simplifies matter when aiming at signal detection, as can be seen from the above discussion of the appropriate test statistic. Then again, it may complicate matters, as can be seen from the above discussion on label versus feature permutation. In general , it is not imperative in general, and we will indeed be comparing the effect of balanced foldings versus unbalanced.

How many folds? Different authors suggest different rules for the number of folds. We will be varying the number of folds, since it will affect the concentration of the estimated accuracy, which will have a crucial effect on the conservativeness of the permutation test. Our intuition suggests that since more folds imply a less concentrated estimate, then leave-one-out should be the less conservative, and 2-fold should be the most conservative.

Power

fff

Discussion

fff

1. Jiang W, Varma S, Simon R (2008) Calculating confidence intervals for prediction error in microarray classification using resampling. Statistical Applications in Genetics and Molecular Biology 7(1).
2. Radmacher MD, McShane LM, Simon R (2002) A Paradigm for Class Prediction Using Gene Expression Profiles. Journal of Computational Biology 9(3):505–511.
3. Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America 103(10):3863–3868.
4. Varoquaux G et al. (2016) Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. working paper or preprint.
5. Golland P, Fischl B (2003) Permutation tests for classification: towards statistical significance in image-based studies. (Springer), Vol. 3, pp. 330–341.
6. Fujikoshi Y, Ulyanov VV, Shimizu R (2011) Multivariate Statistics: High-Dimensional and Large-Sample Approximations. (John Wiley & Sons).
7. Srivastava MS (2013) On testing the equality of mean vectors in high dimension. Acta et Commentationes Universitatis Tartuensis de Mathematica 17(1):31–56.
8. Ramdas A, Singh A, Wasserman L (2016) Classification Accuracy as a Proxy for Two Sample Testing. arXiv:1602.02210 [cs, math, stat].
9. Hastie T, Tibshirani R, Friedman J (2003) The Elements of Statistical Learning. (Springer).
10. Gilron R, Rosenblatt J, Koyejo O, Poldrack RA, Mukamel R (2016) Quantifying spatial pattern similarity in multivariate analysis using functional anisotropy. arXiv:1605.03482 [q-bio].