Statistical Learning

Due Saturday, June 10 on Moodle

Homework-02

Exercise 1: Connect your brain

1. Background: MRI and fMRI

Since its invention in the early 70s by Lauterbur and Mansfield (2003 Nobel prize in Physiology and Medicine), **magnetic resonance imaging** (MRI) has evolved into a versatile tool for the in vivo examination of tissue. Unlike X-ray computed tomography (CT) and positron emission tomography (PET), it does not rely on high energetic radiation but on the nuclear magnetic resonance phenomenon. Consequently, it does in principle not harm the examined tissue and can be applied also in healthy subjects. Thus, MRI is a perfect tool for the examination of the **living brain** in neuroimaging.

Functional magnetic resonance imaging (fMRI) is a technique to examine the human (or animal) brain "at work". fMRI is used to analyze (neuro-)scientific questions, e.g., on the localization of neural capabilities, on the consequences of neuronal diseases or on brain function. For this, in fMRI, a **time series** of MRI volumes is acquired, while the subject in the scanner is typically performing some cognitive task.

What fMRI images visualize is the so called blood oxygenation level–dependent (BOLD) contrast: as active neurons rely on increased oxygen supply, the neural activity is related to a local change in support of blood oxygenation. Thus, fMRI can be used as a natural, yet indirect, contrast for detecting neural activity. In order to achieve a sufficient temporal resolution the spatial resolution of fMRI is typically limited. An fMRI dataset then consists of more than 100 image volumes with a spatial voxel dimension of about 2-4 mm.

⇝ IMPORTANT DISCLAIMER ⇜

Data from fMRI experiments suffer from several artifacts that require special preprocessing ahead of the statistical analysis, like *slice time correction*, *motion correction*, *registration*, *normalization*, *brain masking* and *brain tissue segmentation*.

For the sake of this exercise, I'll provide you with a clean, pre-processed dataset extracted from the *Autism Brain Imagine Data Exchange* (ABIDE) project, but be aware that these early data analytic stages are crucial and not at all trivial.

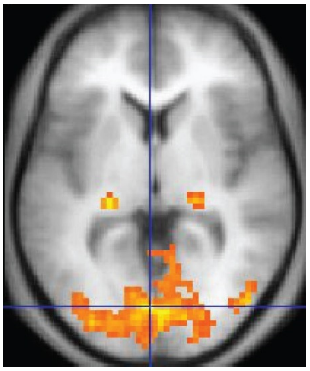


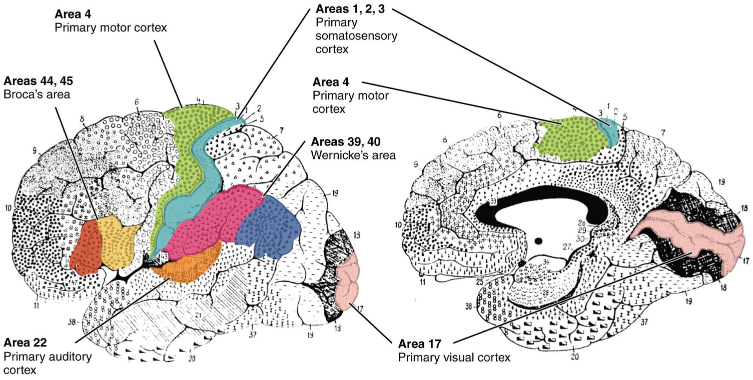
Figure 1: An fMRI image with yellow areas showing increased activity compared with a control condition

Functional Connectivity

The development of MRI and fMRI has paved the way to **connectomics**, i.e., modeling the brain as a network in order to tackle fundamental neuroscience research questions as a *graph-analysis* problem.

Generally speaking, a *connectome* is a map describing neural connection between brain **regions of interest** (`ROIs`), either by observing anatomic fiber density (*structural connectome*), or by computing a suitable statistical association measure (e.g. Pearson correlation) between time series of activity associated to `ROIs` (*functional connectome*). Of the two, the latter is the case of interest to us and from now on we will focus on it.

Nevertheless, before going any further, we need to clarify what these *regions of interest* actually are. Typically, `ROIs` are defined in terms of a suitable **functional brain atlas** which provides information about the spatial location of functional brain regions aggregating knowledge on brain functionality and anatomy accumulated over more that 100 year of brain research. In other words, we essentially use *these* atlases – yes, *these*, because there's more than one – to tag fMRI voxels with specific cortical brain regions. The oldest atlas system dates back to the German anatomist Brodmann who defined 52 cortical areas based on the cytoarchitectural organization of neurons.



Brodmann's cytotechtonic map (1909):
Lateral surface

Brodmann's cytotechtonic map (1909):
Medial surface

This is all nice and good, but to attach an observed fMRI voxels to a specific area of your functional atlas of choice we first need to *normalize* each individual brain or, in other words, we need to map it onto a "standard brain" in order to then be able to identify the corresponding brain regions. As an example, Talairach coordinates, also known as *Talairach space*, is one famous 3-dimensional coordinate system (atlas) that uses Brodmann areas as the labels for brain regions.

The Data: The `Autism Brain Image Data Exchange` Project

In this exercise we use a (*small part of a*) publicly available dataset released by the `Autism Brain Imagine Data Exhange` (ABIDE) project. The dataset contains neuroimaging data of patients suffering from *Autism Spectrum Disorder* (`ASD`) and *Typically Developed* (`TD`) subjects. Since fMRI data are strongly influenced by a variety of confounding factors, in an effort to mitigate this intrinsic variability we will also consider `age` and `sex` as additional covariate[1].

I will provide you with time series of activity associated to 116 distinct `ROIs` observed at 115 time instants in order to predict mental dysfunctions (…possibly building and then using a connectome structure for each subject in the study…). So, in other words, given data extracted from fMRI scans of patients affected by a mental disorder (`autism`) and scans of healthy individuals (`control`), the goal is to discover patterns that explain differences in the brain mechanism of the two groups.

More specifically, the **training** dataset has columns/features that can be broken down in the following way:

- `id`: simply the row index (ignore)
- `age`: the age of the subject
- `sex`: the sex of the subject
- `y`: the target variable to predict (only available in training, of course…)
- From column `5` on, you will find 116 vectorized **time series** (one for each `ROI`) of length 115 each.
  Of course you can handle/simplify these features as you wish, but some basic temporal statistical summaries may be a good starting point.

2. Variable Importance

The LOCO

Being able to quantify the importance of a covariate in predicting a response of interest is crucial in most real applications…even more so nowadays, when the use of very complex, nonlinear, overparametrized models is the rule rather than the exception.

Despite this, the very idea of "importance" is slippery and need to be precisely framed, defined and handled (…yes, even when talking about linear models!). Here we'll focus on a very simple and general technique.

[1] To extract the data, I have followed a preprocessing strategy called DPARSF, plus a band-pass filtering + global signal regression. To parcellate the brain we adopt the AAL atlas (116 `ROIs`). The final result for a single patient is a set of 116 time series of length 145 each.

At page 32 of their paper, Lei and coauthors proposed a simple, general and, essentially assumptions–free idea for measuring variable importance, called **leave-one-covariate-out** (LOCO) inference. The algorithm is extremely simple.

Let $\ell(y, \widehat{y})$ be a suitable error measure/metric/loss for the learning task at hand. Then,

1. Randomly split the training data into two, non overlapping, parts: $\mathsf{D}_n = \mathsf{D}_{n_1}^{(1)} \cup \mathsf{D}_{n_2}^{(2)}$ with $n_1 + n_2 = n$.
2. Run any algorithm you like to compute an estimate $\widehat{f}_{n_1}(\cdot)$ on first part $\mathsf{D}_{n_1}^{(1)}$.
3. Select some variable $\boldsymbol{X}[j]$ of interest to you, and recompute $\widehat{f}_{n_1}^{-j}(\cdot)$ on $\mathsf{D}_{n_1}^{(1)}$ again (rerun algorithm without access to variable $\boldsymbol{X}[j]$).
4. Use $\mathsf{D}_{n_2}^{(2)}$ to construct finite-sample, distribution-free confidence interval (e.g., use non-parametric bootstrap or sign-test or Wilcoxon-test) for the following new (population) parameter[2]:

$$\theta_j\big(\mathsf{D}_{n_1}^{(1)}\big) = \text{median}_{(Y,\boldsymbol{X})}\Big(\ell\big(Y, \widehat{f}_{n_1}^{-j}(\boldsymbol{X})\big) - \ell\big(Y, \widehat{f}_{n_1}(\boldsymbol{X})\big) \,\Big|\, \mathsf{D}_{n_1}^{(1)}\Big), \quad j \in \{1, \ldots, p\}. \tag{1}$$

   Since you're using the same dataset to build more than one confidence-interval, apply any reasonable correction to adjust for multiplicity (e.g. Bonferroni or Benjamini-Hochberg FDR).

θ_j has a very clear interpretation without resorting to linearity or any other *uncheckable* model assumption: it's just how much extra prediction error you would pay by not having access to variable $\boldsymbol{X}[j]$.

In addition, from a more technical point of view, this parameter is "smooth" enough (Hadamard differentiable) to guarantee the success of resampling techniques like the nonparametric bootstrap. In other words, confidence intervals, tests, etc for θ_j are easy to obtain no matter what is the underlying predictive model you picked (LASSO, LASSO + CV, Random Forest, Boosting, Deep nets, etc).

Of course there are also problems with the LOCO. More specifically:

1. It is **not** on an intuitive scale but we could fix this by rescaling.
2. Results are tied to our choice of the algorithm. In theory we could use a "meta–cross–validation" scheme that cycles over different candidate predictor classes (computational expensive but trivially parallelizable).
3. Results are also sensitive to: the ratio of training to test set sizes; the metric $\ell(\cdot, \cdot)$ selected; the correlation between features.
4. It is conditional on $\mathsf{D}_{n_1}^{(1)}$, meaning that it measures "*how important is variable X_j, to our algorithm's estimates on $\mathsf{D}_{n_1}^{(1)}$?*"...not obvious at all how to fix this...

⇝ Your job ⇜

1. Compete in the HW02 Hackathon to show me how good you are!
   Notice that this is a "lazy-competition" meaning that I'm not going for a strict check on code and submissions.
   In fact, no Kaggle notebook required.
   More traditionally, I expect you to upload on Moodle a well commented working code that covers your pipeline (from data loading/pre-processing and feature engineering/dim-reduction to model fit and prediction on `test`).
   REMARK: you can use any classifier/method you feel confortable with.
2. Starting from the model you adopted, evaluate variable importance via `LOCO` for a subset of 10 features.
   More in particular:
   - If the technique you picked comes with a *default* method to evaluate variable importance (e.g. Random Forest), compare the Top-10 in this ranking with their `LOCO` scores.
   - On the other hand, if the technique you picked does not come with a default method to evaluate variable importance, pick 10 features providing some reasoning behind your choice.
   - As usual, relevant & cute visualizations are always welcome...

Exercise 2: Teach-me-how-to-Teach

Do you remember Part II?

[2] Alternatively, you could also consider the median of the ratio or log-ratio between the two losses.