Statistics 215A Final Project Fall 2019

Due: Friday December 13 at 11:59 PM

Note, there is no HW with this lab. Please push a **lab_final**/ folder with the following files by the deadline:

- **lab_final.Rmd** or lab_final.Rnw: the raw report + code.
- **lab_final.pdf**: the output of lab_final.Rnw/lab_final.Rmd. This output should be no more than 12 pages and should not contain any code.
- **R**/: a folder containing any .R scripts (e.g. load.R and clean.R) that will be sourced in lab_final.Rmd and any other pieces of code you use.
- **output**/: a folder containing the out-of-sample voxel predictions (`predv1_yourname.txt`).

You may create additional folders if you wish, but please keep your project directory organized. Push everything necessary to reproduce your report, but nothing else (e.g., do not push your **data**/ or **documents**/ folders). Note also that you do not have to create blinded files.

1 Introduction

For this project, we look at fMRI data measuring the brain's responses to visual images. This data is provided by the Gallant Lab here at UC Berkeley.

2 Background

Much like an image can be spatially discretized into units of 2D pixels, MRI can record measurements on discretized 3D volumes of the brain - cube-like units called voxels. For this project, we are focusing on the responses in 20 voxels located in the region of the brain responsible for visual functions.

The subject is shown pictures of everyday objects, such as a baby, stars, etc. Each picture is a 128 pixel by 128 pixel gray scale image, which can be represented by a vector of length $128^2 = 16384$. These image vectors can be reduced to length 10921 through a transformation. More details on the transformation below.

Although the actual fMRI response to a given stimulus/image is a function of time, the response of each voxel to each image has been reduced to a single number.

3 Data

The data is included in the data folder as `fMRIdata.RData`.

If you load this data into R and type `ls()` you will see four matrix objects listed:

`resp_dat`	1750 × 20	contains responses from the 20 voxels to 1750 images (responses of the training set).
`fit_feat`	1750 × 10921	contains the 1750 transformed images (features of the training set).
`val_feat`	120 × 10921	contains a separate validation set of 120 transformed images (features of the testing set).
`loc_dat`	20 × 3	spatial location of the voxels (20 voxels × 3).

There are two other datasets you may want to use. They are too large to be uploaded to GitHub (each is about 500 MB) and can be found instead on Google Drive (see the `data_location.txt` file in the data folder for the link).

`fit_stim.csv`	1750 × 16384	Each row contains the intensity values for each pixel in the 1750 untransformed images.
`real_wav.csv`	16384 × 10921	Each column contains the the real part of the 10921 Gabor wavelets.

3.1 The transformation

If `wav.pyr` contains the complex Gabor wavelet pyramid (you don't have this), then the transformation that has been used from the image vectors to the feature vectors is,

```
fit_feat = log ( abs( fit_stim %*% wav.pyr ) + 1 )
```

Note that this transformation has two non-linearities: The first is the use of absolute value, and the second is the $\log(x+1)$ transform. We are providing you with just the real part of the Gabor wavelet so you can see what the wavelet looks like, but you won't be able to perform this transform.

4 Tasks

There are two objectives for this project. The first is to predict the voxel response to new images. The second is to interpret the prediction models in relation to the scientific problem - how do voxels respond to images?

1. Partition your data into training, validation, and testing sets (see 7.2 in The Elements of Statistical Learning for advice). You should use the training set to select your model parameters and check performance with the validation set. The sizes of the subsets and how you want to partition the data are up to you.

2. Use some regression methods to come up with a model to predict the response of the all the voxels to new images. You should use LASSO and at least one other regression method in conjunction with model selection techniques. For LASSO, please compare the following model selection criteria: CV, ES-CV, AIC, AICc, and BIC, for selecting the smoothing parameter in Lasso. Describe the strengths and weaknesses of the different criteria.

3. The performance indicator used in the Gallant lab is the correlation between the fitted values and observed values on a separate validation set. Use the validation set you choose in Task 1. Report the correlation between fitted values and observed values based on your predictor for all 20 voxels.

4. Diagnostics: Choose a couple of your models to further investigate (you may restrict your further analysis to one or two voxels). Check the fit of your models. Are there any outliers? Discuss the stability of your prediction results and of the models.

5. Try to interpret your models. Do they share any features? Which predictors are important for which voxels? What features are stable across different bootstrap samples? Could you do hypothesis testing on the estimated parameters? How? What can be learned about how voxels respond to images? `fit_stim.csv` may be useful here.

6. Use your best model to predict the response of the first voxel on the 120 images in `val_feat`. Save your output one prediction per line in the same order as in the validation set in a plain text file (no header). Push a file called "`predv1_yourname.txt`" into a folder called "`output`" in your final project repo.

5 Notes

1. There are many parameters to tweak among the suggested methods. You may want to run R scripts on the SCF clusters.

2. The transformation from the pixels of the stimulus to Gabor wavelets is only one of many possible transformations. If you are feeling very ambitious you could try to predict the response from the stimulus values (`fit_stim.csv`) directly, or use some other transformation.

3. As this is your final project, please note that I will allot points towards visualizations, readability of code, thoughtfulness of your introduction/discussion/conclusion, overall formatting and presentation (e.g., cross-referencing of figures and tables, appropriate captions, no R output), and reproducibility in addition to the guided analysis above.