Assignment 2

Psych 253 (Stanford University, Spring 2020)

Released: 2020/04/22; Due: 2020/05/13

1 Assignment Questions

1.1 Regression in neural data

In this assignment, we will first train regression models from neural responses to predict object sizes in the stimulus or neural responses. **In the following problems, please report the variance explained on the validation set as regression performances.**

1.1.1 Problem 1: Size regression from neural responses

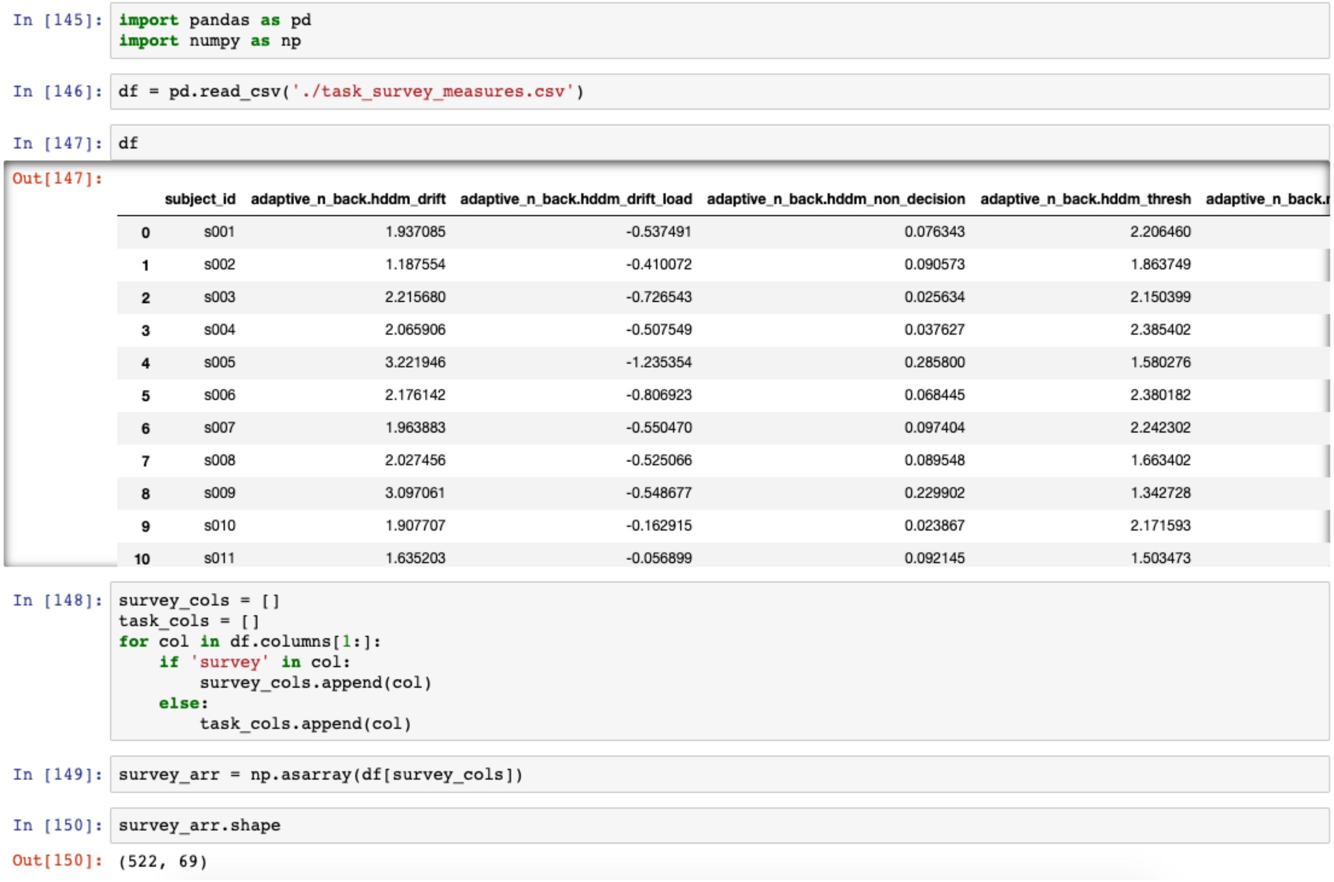
As described before, the stimuli presented to the animals are generated by rendering one of 64 3-dimensional objects at some chosen position, pose, and size, on a randomly chosen background photograph. In this problem, we are interested in predicting the sizes of the objects in the stimuli ("size" stored in "image_meta") by regressing on the neural responses. Besides, we are also curious about the differences between V4 and IT neurons on their abilities to predict object sizes in stimuli. **Therefore, in the following questions of this problem, please train regression models from V4 and IT neurons separately.** Also, as discovered in Assignment 1, the last 40 neurons of V4 need to be ignored. If comparisons between results from V4 and IT neurons are needed, please subsample neurons in IT to equalize the neuron numbers.

- **OLS regression:** Please train OLS regression models from V4 and IT neurons to predict object sizes. You can run multiple cross-validation image splits with different neuron subsampling in order to compare the performances between V4 and IT neurons. Meanwhile, please also train the regression models from all IT neurons. Finally, draw a barchart for regression performances of V4 neurons, subsampled IT neurons, and all IT neurons with error bars. Use proper statistic test to find the better area.
- **Ridge regression:** Ridge regression model can avoid overfitting by adding the L2 regularization on weights. In this question, we will find the best parameter for Ridge regression and compare the performances between Ridge and OLS regression. Specifically:
  - Please train Ridge regression from V4 and IT neurons to predict object sizes using alpha values from {1e-7, 1e-5, 1e-3, 1e-2, 1e-1, 1, 10, 100}. Please use the same cross-validation image splits and the same neuron subsampling as OLS regression. After training regression models for all alpha values, please find the best value for each neuron group (V4 neurons, subsampled IT neurons, all IT neurons). Finally, draw a barchart for regression performances from the best alpha parameters for three groups of neurons with error bars. In the same barchart, please also include the performance bars from previous question beside the bar of the same neuron group. Use proper statistic tests to find the better area and also determine whether Ridge regression is better than OLS regression for each group.
  - RidgeCV class in sklearn can be used to explore the best alpha value automatically using cross-validation. In this question, please use the same alpha value set to train RidgeCV regression model for V4 neurons and all IT neurons (training on subsampled IT neurons is NOT required). Please choose your cross-validation splitting strategy. Is your best alpha value the same as previous question?

1.1.2 Problem 2: Predicting neural responses from neural responses

In this problem, we are interested in predicting neural responses of one animal from another animal (let's say, from Chabo to Tito, but please remember to exclude the last bad V4 neurons in each animal). Please use Ridge, Lasso, and PLS regression with the best parameter for each type of regression model (alpha for Lasso and Ridge, "n_components" for PLS). In order to find the optimal parameter, choose your own parameter sets properly for each regression model and use the same cross-validation image splits for fair comparisons. Finally, find the best parameter separately and draw a barchart containing three bars for each model with error bars. Use proper statistic test to find the best method.

t



```
In [145]: import pandas as pd
          import numpy as np

In [146]: df = pd.read_csv('./task_survey_measures.csv')

In [147]: df

Out[147]:
```

	subject_id	adaptive_n_back.hddm_drift	adaptive_n_back.hddm_drift_load	adaptive_n_back.hddm_non_decision	adaptive_n_back.hddm_thresh	adaptive_n_back.r
0	s001	1.937085	-0.537491	0.076343	2.206460	
1	s002	1.187554	-0.410072	0.090573	1.863749	
2	s003	2.215680	-0.726543	0.025634	2.150399	
3	s004	2.065906	-0.507549	0.037627	2.385402	
4	s005	3.221946	-1.235354	0.285800	1.580276	
5	s006	2.176142	-0.806923	0.068445	2.380182	
6	s007	1.963883	-0.550470	0.097404	2.242302	
7	s008	2.027456	-0.525066	0.089548	1.663402	
8	s009	3.097061	-0.548677	0.229902	1.342728	
9	s010	1.907707	-0.162915	0.023867	2.171593	
10	s011	1.635203	-0.056899	0.092145	1.503473	

```
In [148]: survey_cols = []
          task_cols = []
          for col in df.columns[1:]:
              if 'survey' in col:
                  survey_cols.append(col)
              else:
                  task_cols.append(col)

In [149]: survey_arr = np.asarray(df[survey_cols])

In [150]: survey_arr.shape

Out[150]: (522, 69)
```

Figure 1: Illustration for how to load the self-regulation ontology data and how to get task and survey related columns.

1.2 Regression in self-regulation ontology data

Next, we will use regression to examine the prediction power of survey and task measures to self-regulation measures. The data used in the following problems has been introduced in the class during Lecture 2. A detailed description of this data can be found in [1]. **Similarly to previous problems, please report the variance explained on the validation set as regression performances.**

1.2.1 Data explanation

The data is stored in three csv files:

- "task_survey_measures.csv": a table containing both the survey and task measures for the subjects, where each row is one subject, each column with "survey" in its name is one survey measure, and other columns are task measures.
- "retest_task_survey_measures.csv": a table containing the retested survey and task measures for some of the subjects, organized similarly to "task_survey_measures.csv".
- "self_regulation_measures.csv": a table containing the self regulation measures for the subjects, where each row is one subject and each column is one self regulation measure.

The measures provided are already abstract and processed measures computed from the raw survey responses and task performance. Please check [1] to know the details about the processing procedure. We provide an example about how to load the data in Figure 1.

1.2.2 Problem 3: Regression to self regulation measures

In this problem, we will train regression models from task and survey measures to self-regulation measures and then compare the performance of regression models from these two measures. Please use Ridge regression model and search for optimal alpha values. Please use the same cross-validation subject splits in both two measures and all candidate alpha values, so that their regression performances can be compared fairly. Finally, draw a barchart for regression performances from the best alpha parameters for two measures with error bars. What is your conclusion?

2 Further reading

Some additional reading you might find useful:

- Andrew Ng's "Feature Selection, L1 and L2 Regularization, and Rotational Invariancee" is a useful example of how machine learning experts have analyzed regularized regression techniques. (See `https://icml.cc/imls/conferences/2004/proceedings/papers/354.pdf`.)
- WN van Wieringen has a nice set of lecture notes on ridge regression that dive into the details of linear algebra of regressions as well as given some nice non-neural-related examples. `https://arxiv.org/pdf/1509.09169.pdf`.
- The classic work by Robert Tibshirani (here at Stanford) on the Lasso is worth reading: [2]. The same can be said for Trevor Hastie's (also at Stanford) work on the ElasticNet [3].
- My work with Ha Hong and Jim DiCarlo on regressing category-orthogonal properties in macaque IT, which came up in class, might be of interest as an application [4].
- As a reference to using neural responses as both targets and predictors of regressions, you can have a look at Charle Cadieu's 2007 work [5], or some of the modeling work also from Jim DiCarlo's group (and in which I was involved) e.g. [6].

References

[1] Eisenberg, I. W. *et al.* Uncovering mental structure through data-driven ontology discovery. *PsyArXiv* (2018).

[2] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288 (1996).

[3] Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320 (2005).

[4] *Hong, H., *Yamins, D. L., Majaj, N. J. & DiCarlo, J. J. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience* **19**, 613–622 (2016).

[5] Cadieu, C. *et al.* A model of v4 shape selectivity and invariance. *J Neurophysiol* **98**, 1733–50 (2007).

[6] Yamins*, D. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* (2014).