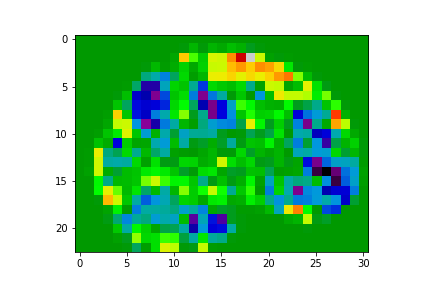
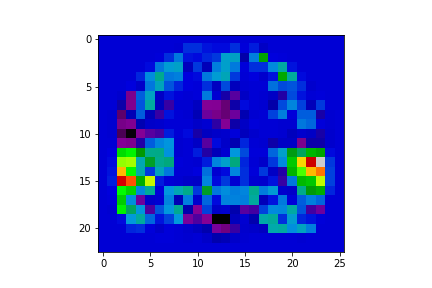
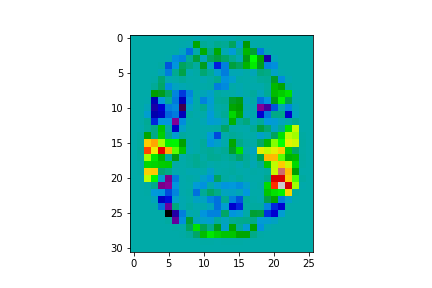
Project: Multilabel classification of brain imaging data

Overview

FMRI is a technology used to measure brain activity, typically while the subject is performing a task in the MRI scanner. It works by measuring how blood flow changes in response to neuronal activity. The resulting data are a time series of brain volumes. After pre-processing, we can compute fMRI brain maps which roughly reflect locations of brain activity corresponding to the task the subject was involved in i.e. a single map for each subject/activity. One important topic in neuroscience is to assess to what extent one can predict behavior from fMRI brain maps. This is known as “decoding” or “brain reading” [2, 3]. Each brain image may be associated with multiple cognitive processes (sub-tasks). Ideally, decoding can predict which sub-tasks are associated with each brain map [1] -- denoted as a set of labels.

The prediction problem which involves predicting a set of labels is known as multilabel classification. It is often thought of as the simplest kind of structured prediction [4, 5].

Your goal is to predict a set of labels associated with each fMRI brain image. You can use any technique you like. Your project will be judged based on performance on two metrics: Subset Accuracy and Micro-averaged AUC.

Reference

[1] Schwartz, Yannick, Bertrand Thirion, and Gaël Varoquaux. "Mapping paradigm ontologies to and from the brain." *Advances in neural information processing systems.* 2013.

[2] Varoquaux, Gael, and Bertrand Thirion. "How machine learning is shaping cognitive neuroimaging." *GigaScience* 3.1 (2014): 28.

[3] Pereira, Francisco, Tom Mitchell, and Matthew Botvinick. "Machine learning classifiers and fMRI: a tutorial overview." *Neuroimage* 45.1 (2009): S199-S209.

[4] Tsoumakas, Grigorios, and Ioannis Katakis. "Multi-label classification: An overview." *International Journal of Data Warehousing and Mining* 3.3 (2006).

[5] Koyejo, Oluwasanmi O., et al. "Consistent multilabel classification." *Advances in Neural Information Processing Systems.* 2015.

Learning Goals:

* Real-world experience on general ML principles, multilabel classification, blind test evaluation
* Investigate the importance of metrics on classification (the best answer for each of the metrics will differ)
* Investigate the effect of spatially correlated features on classification. Use of convolution and other image processing techniques may be helpful.
* Investigate learning with high dimensional input features
* Investigate data preprocessing/cleaning.

Kaggle Competition and grading

Your performance will be evaluated via two kaggle competitions -- as compared to a variety of given baselines. A grade is determined when your performance is better than the appropriate threshold. The final grade will be the average performance across the two metrics. Grading baselines will be added very soon.

[How to get started with Kaggle competitions](https://youtu.be/ez_eVGUjIP8)

Due date

Competition Ends **Dec 19, 11:59PM**

**IMPORTANT:** PLEASE START EARLY. THERE WILL BE NO EXTENSIONS ON THE DUE DATE.

Sign up!

Please sign-up for the following two competitions.:

-[Subset Accuracy Competition](https://www.kaggle.com/c/cs446-fa17) Use this [link](https://www.kaggle.com/t/8ad8d652fe8340f7a9a44dc117f364b5) to sign up.

-[Micro-averaged AUC Competition](https://www.kaggle.com/c/cs446-fa17b)

We will identify you using your NetID. An account created using your NetID is preferable for easy matching during the grading time; however, if you want to use your existing kaggle account, please email the instructor your Kaggle ID and NetID.

You will be restricted to a maximum of 4 submissions per day. The leaderboard shows scores on the validation set. Final grades are based on scores evaluated on a hidden test set with similar distributions as the training and the vaalidation set.

Team

You can work by yourself, or in teams of up to two people e.g. each person may focus on one metric.

Computation

You are free to use any compute resources you like. **Microsoft Azure** has graciously donated cloud computing credits for each student to use. Microsoft will be on campus on **Nov 1, 2017 at 6:30pm in SC 1109** to present a tutorial overview of how to use Azure Cloud for Computing.

Data Download data [here](https://uofi.box.com/v/cs446-project)(all numpy format)

* Label names: tag\_name.npy : identifies each label, not required for your model [Azure Link](https://cs446data.blob.core.windows.net/cs446-fall17-data/tag_name.npy)
* Training features: train\_X.npy (4-D tensor. first dimension is N = number of examples). Each sample is a low resolution brain image representing brain activity corresponding to the given labels. [Azure Link](https://cs446data.blob.core.windows.net/cs446-fall17-data/train_X.npy)
* Training labels: train\_binary\_Y.npy: labels represented as binary vector, Size = number of examples \* number of labels. [Azure Link](https://cs446data.blob.core.windows.net/cs446-fall17-data/train_binary_Y.npy)
* Test features: valid\_test\_X.npy [Azure Link](https://cs446data.blob.core.windows.net/cs446-fall17-data/valid_test_X.npy)

For loading data on python, see [here](https://docs.scipy.org/doc/numpy-1.13.0/reference/generated/numpy.load.html)

Prediction

Test labels are submitted as a text file to the kaggle site. Please use the provided python script to convert your binary matrix prediction into a format that Kaggle will accept. The order of your predictions should line up with the order of examples in valid\_test\_X.npy

Usage: python3 npy\_to\_csv.py [subset|auc] <prediction\_file> <output\_file>

e.g. python3 npy\_to\_csv.py subset result.npy subset\_submission.csv

e.g. python3 npy\_to\_csv.py auc result.npy auc\_submission.csv

Metrics

* Subset Accuracy Competition: The evaluation metric is identical to a. converting from multilabel to multiclass i.e. a different class for each label configuration b. Applying this metric to the resulting [multiclass labels](http://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html)
* Micro-averaged AUC Competition: The evaluation metric is identical to [#sklearn.metrics.roc\_auc\_score (with average = “micro”)](http://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html)

Useful links

[scikit-learn](http://scikit-learn.org/stable/modules/multiclass.html) reference for more on multiclass and multilabel classification:

[Essential Cheat Sheets for Machine Learning and Deep Learning](https://medium.com/@kailashahirwar/essential-cheat-sheets-for-machine-learning-and-deep-learning-researchers-efb6a8ebd2e5)