Mini Project 1

EN 601.751
Fall 2017
Prof. Alexis Battle

DUE DATE: 10/18/17

1. Writing introductory text.

- For each of two papers we have read in class (any paper we read up through 10/4), you will write alternative introductory text. This should be 4-8 sentences that would begin the introduction (**NOT the abstract**), introducing the topic and importance of the paper.
    - Select a paper from the readings that, as written, appears to target a broad rather than specialized audience. Write new introductory sentences as if you were targeting specialized experts in the topic of the paper.
    - Select a second paper that appears to target a specialized audience. Write new introductory sentences as if you were targeting a broad audience, such as Science/Nature/grant reviewers.

2. Data analysis:

Here, you will use gene expression data from the GTEx project to predict age of each individual. **Your goals are to:**

- Design an ML approach for predicting age from gene expression data.
- Evaluate how well age can be predicted is from various tissues of the body and compare between tissues.
    - Consider what decisions/properties of your approach make your results biologically meaningful or not.

You will be graded, not on the raw performance of your method, but on the correctness of your analysis, reproducibility, and thought you put into designing your approach. It's ok if you encounter challenges/confusing results – just do your best to describe them, and brainstorm why they might be occurring or what could improve them if you had more time.

To obtain the data:

1. Create an account on http://gtexportal.org/, and log in
2. From http://gtexportal.org/home/datasets, download:
    a. v6p, Single Tissue cis-eQTL Data (this will just be expression data per tissue):
        i. GTEx_Analysis_v6p_eQTL_expression_matrices.tar
    b. v6, Sample attributes:
        i. GTEx_Data_V6_Annotations_SubjectPhenotypesDS.txt
        ii. GTEx_Data_V6_Annotations_SubjectPhenotypes_DD.xlsx

Perform all steps from preprocessing onward to run a simple standard supervised machine learning approach to predict individual age from gene expression data, and determine which tissue is most informative regarding age.

- Build feature matrices **X**, separately for gene expression data from 4 tissues (Muscle_Skeletal, Whole_Blood, Adipose_Subcutaneous, Thyroid).
- Build response vectors **y** (age), one for each tissue over the corresponding individuals, from the sample attribute text file.
    - Note each tissue has a different set of individuals measured, so you will have to extract ages for the individuals in each tissue from the attribute file
- Preprocess the data as needed. This *may* include steps such as applying simple transformations, centering the data, etc. as needed by the assumptions of method you choose and what you think will work best:
    - One key question will be how to represent age in **y**
- For each tissue, separate data into training and test instances (70% train and 30% test)
- Select a subset of features to use (for efficiency you may choose a small set such as 1,000 total if you want to). This can be based on simple criteria such as correlation of each feature (gene) with the response y (**measured in the training set only**)
    - Optional: you are likely to get better performance if you consider feature selection carefully and try varying the number of features to find a good number using cross-validation.
    - Optional: consider removing batch effects with methods like PCA or SVA.
- Run an appropriate, standard supervised ML approach using an existing package in R or python (such as random forests, LASSO, ridge regression, support vector regression, or other existing method, your choice)
- Compute test set performance for each tissue as you vary the number of **training** examples used, for at least 5 settings. Appropriate metrics depend on your method but would include R^2, RSS, etc.
- If you chose to use a method with regularization, you also need to consider how to set any hyperparameters – pick a parameter by hand, cross-validation etc. Do NOT pick the hyperparameter setting by using your test data. It's ok to take a simple approach here, but in the report, discuss effects of this choice on performance.

2.1. Deliverables:

- Two figures:
    - **Figure 1**: For Muscle_Skeletal, plot test set performance across all settings of number of training samples that you tried.
    - **Figure 2**: Plot test set performance for all tissues using a) the total number of training samples available in each tissue and b) limiting all tissues to the same number of training samples based on the number available in the smallest tissue.
- A script in R or python (or other language with special permission) that runs every step of your analysis and produces the final result.
    - Must run on ugrad, grad, or hhpc clusters (Please tell me which).
    - The script can call other code you write, but you must provide all source code needed to run it.

    - You are free to use any standard libraries or easily installed packages in the language you are using (scikit-learn, anything from bioconductor etc)
    - **The script must include all steps of analysis. It should take me from raw downloaded data to final result.**

- A README file with instructions for running your analysis script (this should not be complicated), along with:
    - Brief comments on results/difficulties you had.
    - Which tissue appears to have the most signal for predicting age? Discuss any issues with interpreting your results in terms of biological meaning.
    - How could you improve your analysis in the future given a longer timeframe?
    - Does your analysis tell you anything about which genes are most relevant to age?