SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Intro to Big Data Science: Final Project

Due Date: Jan 15

✏ **Problem**(Investigating breast cancer data, **totally 30 points**)

This data set contains published iTRAQ proteome profiling of 77 breast cancer samples generated by the Clinical Proteomic Tumor Analysis Consortium (NCI/NIH). It contains expression values for more than 12.000 proteins for each sample, with missing values present when a given protein could not be quantified in a given sample. The expression values may affect the breast cancer, e.g., the property and geometry of tumor, and cancer classifications in the clinical examination.

There are three CSV files whose descriptions are given below:

1. “77_cancer_proteomes_CPTAC_itraq.csv”:

   This file includes all the protein information. Each row represents one certain type of protein. Its attributes include:

   - RefSeq_accession_number: RefSeq protein ID (each protein has a unique ID in a RefSeq database)
   - gene_symbol: a symbol unique to each gene (every protein is encoded by some gene)
   - gene_name: a full name of that gene
   - Remaining columns: log2 iTRAQ ratios (real-valued) of each protein for each healthy or unhealthy sample (These are protein expression data, most important for our analysis). There are totally 83 columns while three last columns are from healthy individuals. And the other 80 columns only have 77 useful data from breast cancer samples, with 3 redundant columns replicated from the 77 useful columns. Please remove the 3 replicated columns before doing the data analysis.

2. “clinical_data_breast_cancer.csv”:

   This file contains the clinical examination results of 105 breast cancer samples. Each row represents a certain unhealthy sample. The attributes of each row are the ID and examination results of this sample. First column "Complete TCGA ID" is used to match the sample IDs (see column names) in the main file “77_cancer_proteomes_CPTAC_itraq.csv”. All other columns have self-explanatory names, contain data about the cancer classification of a given sample using different methods.

   Why there are 77 samples in the protein gene expression data while there are 105 samples in the clinical examination data? Explanation: In the original study the proteomics data were acquired for 111 samples (105 tumors, 3 replicated analyses that you removed and 3 healthy samples). However, as it’s usually the case with mass spectrometry, not all sample acquisitions work properly. To assess the quality people use dozens of different parameters and sometimes one has to simply discard a chunk of data that didn’t pass your quality requirements. This was the case here and the researchers removed 28 samples from the final data.

   Therefore, before doing analysis you have to match the sample IDs of clinical data to those of the protein expression data.

3. “PAM50_proteins.csv”:

   Contains the list of genes and proteins used by the PAM50 classification system. Totally there are 100 patients participating the PAM50 classification system. If you want to use the ’PAM50 mRNA’ classification result in the clinical data, to make your analysis reliable, you have to filter the sample IDs such that they appear in the file “PAM50_proteins.csv”. The column RefSeqProteinID contains the protein IDs that can be matched with the IDs in the main protein expression data set.

The original study in the past research can be found in the website:
http://www.nature.com/nature/journal/v534/n7605/full/nature18003.html

In brief: the data were used to assess how the mutations in the DNA are affecting the protein expression landscape in breast cancer. Genes in our DNA are first transcribed into RNA molecules which then are translated into proteins. Changing the information content of DNA has impact on the behavior of the proteome, which is the main functional unit of cells, taking care of cell division, DNA repair, enzymatic reactions and signaling etc. They performed K-means clustering on the protein data to divide the breast cancer patients into sub-types, each having unique protein expression signature. They found that the best clustering was achieved using 3 clusters (original PAM50 gene set yields four different subtypes using RNA data).

An example python script using k-means, named “breast_kmeans.py”, is also attached with this file.

In this final project, you will have to do the following tasks and answer questions step by step:

1. **(5 points) Data preprocessing:** including detecting missing values (if any) and outlier samples (if any), data conversion and normalization (if necessary), splitting the data into training set and test set (if needed in case of supervised learning), and so on. It is helpful to follow the example script "breast_kmeans.py";
2. **(5 points)** Take the protein gene expression (numerical) values as the input, do **clustering analysis** and compare your clustering results with "PAM50 mRNA" examination results. You may try the example script "breast_kmeans.py", and examine whether it is possible to improve the clustering performance.
3. **(10 points) Using classification analysis instead of clustering.** Also take the protein gene expression (numerical) values as the input, but take "PAM50 mRNA" results as labels. You can divide the 77 unhealthy samples into training and test sets: among the 77 samples, there are 43 patients participating PAM50 examination; select these 43 samples as training set and the other 34 samples as test set. Do classification analysis. Can you explain which gene expression attributes are important for "PAM50 mRNA" label?
4. **(10 points)** You can also take other cancer examination results, e.g., "Tumor", "ER Status", "miRNA Clusters", and even "Gender", etc., as your concerns. **Design your own problem, supervised (classification, regression), or unsupervised (clustering) learning, and analyze the problem using these data.** Please state your problem clearly, and write down your analysis step by step.
5. Conclusion.

The following skills shall be integrated in this project:

1. Data exploration: including data statistics and data visualization;
2. Data preprocessing;
3. Model construction: you could use any model you prefer, even the model we did not cover in class;
4. Feature selection and model selection;
5. Model evaluation;
6. Report writing.

These should be included in your report. You should also have a complete set of codes. You report (typically in pdf format) and codes should be compressed in a zip file. Please use your student ID and name to rename your zip file. Then the zip file shall be uploaded to the classroom of hackdata.

Your project will be graded based on several factors, including the accuracy, comparison of different methods, whether your methods are innovative, the quality of your report, the analysis on your methods and you results (e.g., computational efficiency, model interpretability, etc.), and the quality of your codes.

The final project must be done individually.