CSE6250: Big Data Analytics in Healthcare
Homework 3

Jimeng Sun

Deadline: Feb 25, 2018, 11:55 PM AoE

- Discussion is encouraged, but each student must write his/her own answers and explicitly mention any collaborators.
- Each student is expected to respect and follow the GT Honor Code.
- Please type the submission with LaTeX or Microsoft Word. We **will not** accept hand written submissions.
- Please **do not** change the filenames and function definitions in the skeleton code provided, as this will cause the test scripts to fail and subsequently no points will be awarded.

Overview

Accurate knowledge of a patient's disease state is crucial to proper treatment, and we must understand a patient's phenotypes (based on their health records) to predict their disease state. There are several strategies for phenotyping including supervised rule-based methods and unsupervised methods. In this homework, you will implement both type of phenotyping algorithms using Spark.

Prerequisites [0 points]

This homework is primarily about using Spark with Scala. We strongly recommend using our bootcamp virtual environment setup to prevent compatibility issues. However, since we use the Scala Build Tool (SBT), you should be fine running it on your local machine. Note this homework requires Spark 1.3.1 and is **not compatible** with Spark 2.0 and later. Please see the build.sbt file for the full list of dependencies and versions.

Begin the homework by downloading the hw3.tar.gz from Canvas, which includes the skeleton code and test cases.

You should be able to immediately begin compiling and running the code with the following command (from the *code/* folder):

```
sbt/sbt compile run
```

And you can run the test cases with this command:

```
sbt/sbt compile test
```

1 Programming: Rule based phenotyping [30 points]

Phenotyping can be done using a rule-based method. The Phenotype Knowledge Base (PheKB) provides a set of rule-based methods (typically in the form of decision trees) for determining whether or not a patient fits a particular phenotype.

In this assignment, you will implement a phenotyping algorithm for type-2 diabetes based on the flowcharts below. The algorithm should:

- Take as input event data for diagnoses, medications, and lab results.
- Return an RDD of patients with labels (*label*=1 if the patient is case, *label*=2 if the patient is control, *label*=3 otherwise).

You will implement the *Diabetes Mellitus Type 2* algorithms from PheKB. We have reduced the rules for simplicity, which you can find in the images below. However, you can refer to the full description for more details if desired.

The following files in *code/data/* folder will be used as inputs:

- **encounter_INPUT.csv**: Each line represents an encounter and contains a unique encounter ID, the patient ID (Member_ID), and many other details about the counter. *Hint: sql join*
- **encounter_dx_INPUT.csv**: Each line represents an encounter and contains any resulting diagnoses including a description and ICD9 code.
- **medication_orders_INPUT.csv**: Each line represents a medication order including the name of the medication.
- **lab_results_INPUT.csv**: Each line represents a lab result including the name of the lab (Result_Name), the units of the lab output, and lab output value.

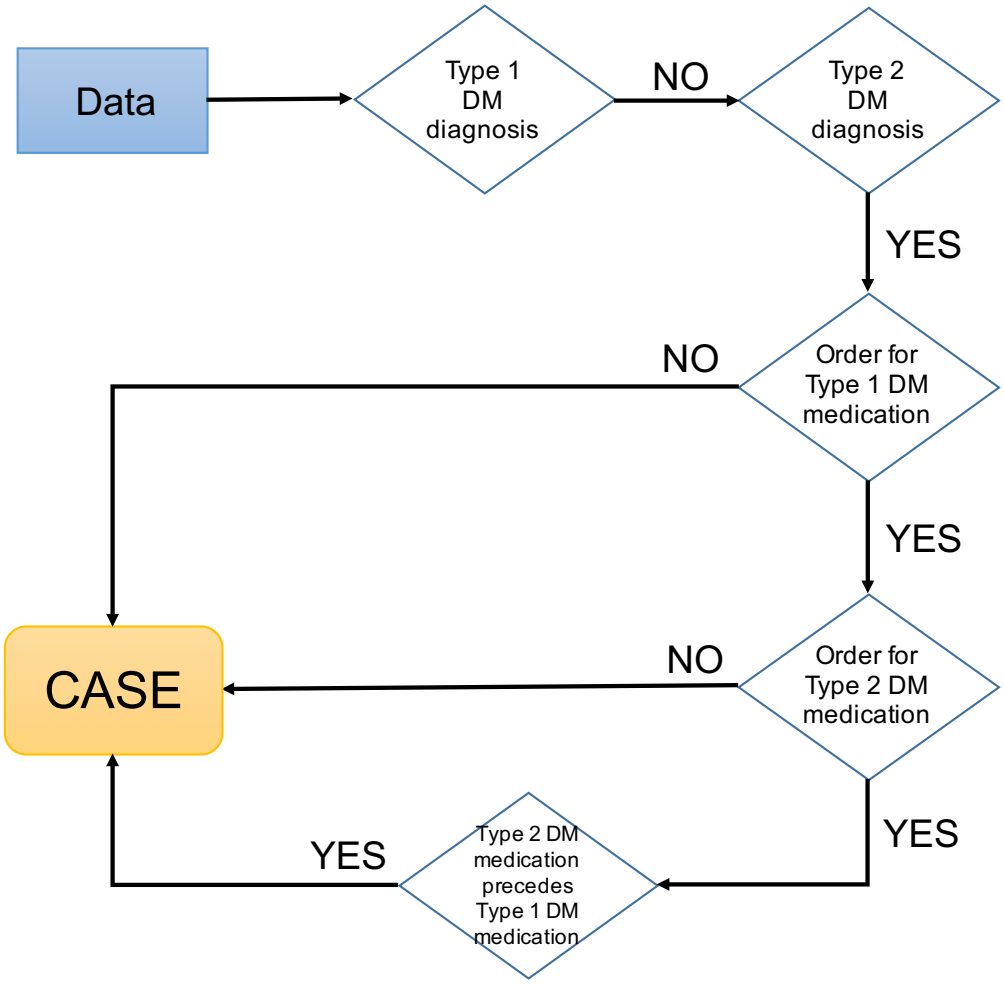

Figure 1: Determination of cases

For your project, you will load input CSV files from the *code/data/* folder. You are responsible for transforming the .csv's from this folder into RDDs.

The simplified rules which you should follow for phenotyping of Diabetes Mellitus Type 2 are shown below. These rules are based off of the criteria from the PheKB phenotypes, which have been placed in the *phenotyping_resources/* folder.

- **Requirements for Case patients**: Figure 1 details the rules for determining whether a patient is case. Certain parts of the flowchart involve criteria that you will find in the *phekb_criteria/* folder as outlined below:

  - **T1DM_DX.csv**: Any ICD9 codes present in this file will be sufficient to result in `YES` for the *Type 1 DM diagnosis* criteria.
  - **T1DM_MED.csv**: Any medications present in this file will be sufficient to result in `YES` for the *Order for Type 1 DM medication* criteria. Please also use this list for the *Type 2 DM medication preceeds Type 1 DM medication* criteria.

- **T2DM_DX.csv**: Any of the ICD9 codes present in this file will be sufficient to result in `YES` for the *Type 2 DM diagnosis* criteria.
- **T2DM_MED.csv**: Any of the medications present in this file will be sufficient to result in `YES` for the *Order for Type 2 DM medication* criteria. Please also use this list for the *Type 2 DM medication preceeds Type 1 DM medication* criteria.

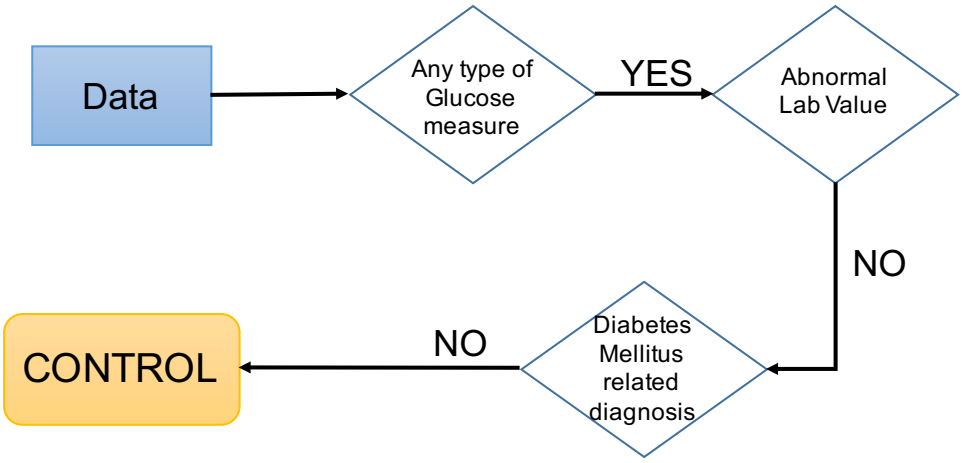

Figure 2: Determination of controls

- **Requirements for Control patients**: Figure 2 details the rules for determining whether a patient is control. Certain parts of the flowchart involve criteria that you will find in the *phekb_criteria/* folder as outlined below:
  - **ABNORMAL_LAB_VALUES_CONTROL.csv**: Any values described in this file should be considered abnormal for the *Abnormal Lab Value* criteria.
  - **DM_RELATED_DX.csv**: Any ICD9 codes present in this file will be sufficient to result in `YES` for the *Diabetes Mellitus related diagnosis* criteria.

In order to help you verify your steps, expected counts along the different steps have been provided in:

- `phenotyping_resources/expected_count_case.png`
- `phenotyping_resources/expected_count_control.png`

Any patients not found to be in the **control** or **case** category should be placed in the **unknown** category. Additional hints and notes are provided directly in the code comments, so please read these carefully.

a. Implement *edu.gatech.cse8803.main.Main.loadRddRawData* to load the input .csv files in the data folder as structured RDDs. [5 points]

b. Implement *edu.gatech.cse8803.phenotyping.T2dmPhenotype* to:
- Correctly identify case patients [10 points]
- Correctly identify control patients [10 points]
- Correctly identify unknown patients [5 points]

2 Programming: Unsupervised Phenotyping via Clustering [40 points]

At this point you have implemented a supervised, rule-based phenotyping algorithm. This type of method is great for picking out specific diseases, in our case diabetes, but they are not good for discovering new, complex phenotypes. Such phenotypes can be disease subtypes (i.e. severe hypertension, moderate hypertension, mild hypertension) or they can reflect combinations of diseases that patients may present with (e.g. a patient with hypertension and renal failure). This is where unsupervised learning comes in.

2.1 Feature Construction [16 points]

You will need to start by constructing features out of the raw data to feed into the clustering algorithms. You will need to implement ETL using Spark with similar functionality as what you did in last homework using Pig. Since you know the diagnoses (in the form of ICD9 codes) each patient exhibits and the medications they took, you can aggregate this information to create features. Using the RDDs that you created in *edu.gatech.cse8803.main.Main.loadRddRawData*, you will construct features for the COUNT of medications, COUNT of diagnoses, and AVERAGE lab test value.

a. Implement the feature construction code in *edu.gatech.cse8803.features.FeatureConstruction* to create two types of features: one using all the available ICD9 codes, labs, and medications, and another using only features related to the phenotype. See the comments of the source code for details.

2.2 Evaluation Metric [8 points]

Purity is a metrics to measure the quality of clustering, it's defined as

$$purity(\Omega, C) = \frac{1}{N} \sum_{k} \max_{j} |w_k \cap c_j|$$

where N is the number of samples, k is index of clusters and j is index of class. w_k denotes the set of samples in k-th cluster and c_j denotes set of samples of class j.

a. Implement the *purity* function in *edu.gatech.cse8803.clustering.Metrics*

2.3 K-Means Clustering [5 points]

Now you will perform clustering using Spark's MLLib, which contains an implementation of the k-means clustering algorithm as well as the Gaussian Mixture Model algorithm.

From the clustering, we can discover groups of patients with similar characteristics. You will cluster the patients based upon diagnoses, labs, and medications. If there are d distinct diagnoses, l distinct medications and m medications, then there should be $d + l + m$ distinct features.

a. Implement k-means clustering for $k = 3$. Follow the hints provided in the skeleton code in *edu.gatech.cse8803.main.Main.scala:testClustering*.

b. Compare clustering for the $k = 3$ case with the ground truth phenotypes that you computed for the rule-based PheKB algorithms. Specifically, for each of *case*, *control* and *unknown*, report the percentage distribution in the three clusters for the two feature construction strategies. Report the numbers in the format shown in Table 1 and Table 2.

Percentage Cluster	Case	Control	Unknown
Cluster 1	x%	y%	z%
Cluster 2	xx%	yy%	zz%
Cluster 3	xxx%	yyy%	zzz%
	100%	**100%**	**100%**

Table 1: Clustering with 3 centers using all features

Percentage Cluster	Case	Control	Unknown
Cluster 1	x%	y%	z%
Cluster 2	xx%	yy%	zz%
Cluster 3	xxx%	yyy%	zzz%
	100%	**100%**	**100%**

Table 2: Clustering with 3 centers using filtered features

2.4 Clustering with Gaussian Mixture Model (GMM) [5 points]

a. Implement GaussianMixture for $k = 3$. Follow the hints provided in the skeleton code in *edu.gatech.cse8803.main.Main.scala:testClustering*.

b. Compare clustering for the $k = 3$ case with the ground truth phenotypes that you computed for the rule-based PheKB algorithms. Specifically, for each of *case*, *control* and *unknown*, report the percentage distribution in the three clusters for the two feature construction strategies. Report the numbers in the format shown in Table 1 and Table 2.

2.5 Clustering with Streaming K-Means [5 points]

When data arrive in a stream, we may want to estimate clusters dynamically and update them as new data arrives. Spark's MLLib provides support for the streaming k-means clustering algorithm that uses a generalization of the mini-batch k-means algorithm with **forgetfulness**.

a. Show why we can use streaming K-Means by deriving its update rule and then describe how it works, the pros and cons of the algorithm, and how the forgetfulness value balances the relative importance of new data versus past history.

b. Implement StreamingKMeans algorithm for $k = 3$. Follow the hints provided in the skeleton code in *edu.gatech.cse8803.main.Main.scala:testClustering*.

c. Compare clustering for the $k = 3$ case with the ground truth phenotypes that you computed for the rule-based PheKB algorithms. Specifically, for each of *case*, *control* and *unknown*, report the percentage distribution in the three clusters for the two feature construction strategies. Report the numbers in the format shown in Table 1 and Table 2.

2.6 Discussion on K-means and GMM [6 points]

We'll now summarize what we've observed in the preceeding sections:

a. Briefly discuss and compare what you observed in 2.3b using the k-means algorithm and 2.4b using the GMM algorithm.

b. Re-run k-means and GMM from the previous two sections for different k (you may run it each time with different k). Report the purity values for all features and the filtered features for each k by filling in Table 3. Discuss any patterns you observed, if any.

NOTE: Please change k back to 3 in your final code deliverable!

k	K-Means All features	K-Means Filtered features	GMM All Features	GMM Filtered features
2				
5				
10				
15				

Table 3: Purity values for different number of clusters

3 Advanced phenotyping with NMF [20 points]

Given a feature matrix $\boldsymbol{V}$, the objective of NMF is to minimize the Euclidean distance between the original non-negative matrix $\boldsymbol{V}$ and its non-negative decomposition $\boldsymbol{W} \times \boldsymbol{H}$ which can be formulated as

$$\underset{\boldsymbol{W} \geq 0, \boldsymbol{H} \geq 0}{\operatorname{argmin}} \quad \frac{1}{2}||\boldsymbol{V} - \boldsymbol{W}\boldsymbol{H}||_2^2 \tag{1}$$

where $\boldsymbol{V} \in \mathbb{R}_{\geq 0}^{n \times m}$, $\boldsymbol{W} \in \mathbb{R}_{\geq 0}^{n \times r}$ and $\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{r \times m}$. $\boldsymbol{V}$ can be considered as a dataset comprised of n number of m-dimensional data vectors, and r is generally smaller than n.

To obtain a $\boldsymbol{W}$ and $\boldsymbol{H}$ which will minimize the Euclidean distance between the original non-negative matrix $\boldsymbol{B}$, we use the Multiplicative Update (MU). It defines the update rule for $\boldsymbol{W}_{ij}$ and $\boldsymbol{H}_{ij}$ as

$$\boldsymbol{W}_{ij}^{t+1} = \boldsymbol{W}_{ij}^{t} \frac{(\boldsymbol{V}\boldsymbol{H}^\top)_{ij}}{(\boldsymbol{W}^t\boldsymbol{H}\boldsymbol{H}^\top)_{ij}}$$

$$\boldsymbol{H}_{ij}^{t+1} = \boldsymbol{H}_{ij}^{t} \frac{(\boldsymbol{W}^\top\boldsymbol{V})_{ij}}{(\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{H}^t)_{ij}}$$

Pseudo-code for the rule is listed below.

```
Initialize W, H randomly;
repeat
    /* Updating W[i,:]                                   */
    Save HH^T as a global variable H_s;
    W^{t+1}[i,:] = W^t[i,:] ⊙ V[i,:]H^T ⊙ (W^t[i,:]H_s)^{-1}
    /* Updating H[:,i]                                   */
    Save W^T W as a global variable W_s;
    H^{t+1}[:,i] = H^t[:,i] ⊙ W^T V[:,i] ⊙ (W_s H^t[:,i])^{-1}
until (1/2)||V - WH||_2^2 < ε;
```

You will decompose your feature matrix $\boldsymbol{V}$, from **2.1**, into $\boldsymbol{W}$ and $\boldsymbol{H}$. In this equation, each row of $\boldsymbol{V}$ represents one patient's features and a corresponding row in $\boldsymbol{W}$ is the patient's cluster assignment, similar to a Gaussian mixture. For example, let $r = 3$ to find three phenotype(cluster), if row 1 of $\boldsymbol{W}$ is $(0.23, 0.45, 0.12)$, you can say this patient should be group to second phenotype as 0.45 is the largest element.

$\boldsymbol{W}$ can be very large, i.e. a billion patients, which must be worked on in a distributed fashion while $\boldsymbol{H}$ is relatively small and can fit into a single machine's memory. You will define these two types of matricies as distributed RowMatrix and local dense Matrix respectively in the skeleton code.

a. Implement the algorithm, as previously described, in *edu.gatech.cse8803.clustering.NMF*. [15 points]

b. Run NMF clustering for $k = 2, 3, 4, 5$ and report the purity for two kinds of feature construction. [5 points]

c. Compare clustering for the $k = 3$ case with the ground truth phenotypes that you computed for the rule-based PheKB algorithms. Specifically, for each of *case*, *control* and *unknown*, report the percentage distribution in the three clusters for the two feature construction strategies. Report the numbers in the format shown in Table 1 and Table 2. [5 points]

d. Show why we can use MU update rule by deriving the equation for it. [10 points bonus]

4 Submission [5 points]

The folder structure of your submission should be as below or your code will not be graded. You can display fold structure using *tree* command. All other unrelated files will be discarded during testing. You may add additional methods, additional dependencies, but make sure existing methods signature doesn't change. It's your duty to make sure your code can be compiled with the provided SBT. Be aware that writeup is within code root.

```
<your gtid>-<your gt account>-hw3
|-- homework3answer.pdf
|-- build.sbt
|-- project
|   |-- build.properties
|   \-- plugins.sbt
|-- sbt
|   \-- sbt
\-- src
    \-- main
        |-- java
        |-- resources
        \-- scala
            \-- edu
                \-- gatech
                    \-- cse8803
                        |-- clustering
                        |   |-- NMF.scala
                        |   |-- Metrics.scala
                        |   \-- package.scala
                        |-- features
                        |   \-- FeatureConstruction.scala
```

```
                        |-- ioutils
                        |   \-- CSVUtils.scala
                        |-- main
                        |   \-- Main.scala
                        |-- model
                        |   \-- models.scala
                        \-- phenotyping
                            \-- PheKBPhenotype.scala
```

Create a tar archive of the folder above with the following command and submit the tar file.

```
tar -czvf <your gtid>-<your gt account>-hw3.tar.gz \
  <your gtid>-<your gt account>-hw3
```