Problem set 2
Due: April 25th by midnight
Through: Gradescope

Instructions:

Late policy: You are allowed to turn in your problem set late, but your grade will be deducted 5% with every late day.

Collaboration policy: You are expected to try solving the problem set on your own. However, if stuck on a problem, you are encouraged you to collaborate with other students in the class, subject to the following rules:

(1) You may discuss a problem with any student in this class, and work together on solving it. This can involve brainstorming and verbally discussing the problem, but should not involve any sharing of code, code fragments, or solutions.
(2) In your solution for each problem, you must write down the names of any person with whom you discussed it. This will not affect your grade.

Output: You are required to submit a single pdf on gradescope by April 25th at midnight. Please include all the relevant code as well as your answers for each question in the pdf.

Data:

Log on to your physionet account. You will find a new dataset called LLDA_data.gz in the MLHC17PS1 you already have access to. This is the only dataset you will need for this problem set. It has information, charts, and ICD9 codes for patients in the Cardiac Surgery Recovery Unit (CSRU) who were in the ICU for longer than 48 hours. Specifically LLDA_data.gz has one line for every CSRU stay and the following variables:

- Icustay_id: unique ids of CSRU stays.
- Hadm_id: hospitalization id
- Subject_id: patient id
- 2 'label' variables (explained in detail later)
    - Multi_label: a combination of patient ICD9 diagnosis codes to be used in question 1
    - Single_label: a combination of patient ICD9 diagnosis codes to be used in question 2
- 4 variables which contain charts from the patients' ICU stay. *Please make sure you only include the relevant chart information in each question.* [1]

[1] The chart variables were created by concatenating all the relevant raw chart data extracted from MIMIC, then we removed digits, special symbols and carriage returns. We also removed

    - Dc_chart: this column has only the patient's chart upon discharge (for question 1)
    - chart0_24 → these are all the patients charts written in the first 24 hours (for question 2)
    - Chart_inbetween: these are all the patients charts written in the first 24 hours (for question 2)
    - Chart24_out: these are all the patients charts written in the last 24 hours, including their discharge summaries (for question 2)
- Additional demographic information about the patient and the stay such as age, ethnicity, etc.

Software:

Github repo:

You will find useful resources and information in the pset2materials github repository, found here https://github.com/mlhc17mit/pset2materials

Mallet:

You will be using Mallet for this assignment. Mallet is a Java-based package for statistical natural language processing, and of particular relevance for this problem set is its MCMC-based learning algorithm for latent Dirichlet allocation (LDA), a probabilistic topic model. Visit http://mallet.cs.umass.edu/ for instructions on how to download and setup the software and useful tutorials.

Hints:

The easiest way to import your wrangle your data into a Mallet friendly dataset is to extract the relevant information into a tab delimited text file, with three columns: ID, label, chart text. Then you can use the Mallet import-file command changing the line-parsing regular expression to separate columns with tabs, for example: --line-regex '([^\t]+)\t([^\t]+)\t(.*)'

Even more help: (It's your lucky day!)

You will find a simple python script (topic_labels.py) in the pset2materials repo that can take in a file with the topic word distribution from your LLDA Mallet runs (explained in the next section), and a icd9_label_ds.csv (a file also found in the repo) and label the ICD9 diagnoses using their description. It will return the same topic-word distribution file and add a column with the descriptions right after the topic name. This will help you when looking at and analyzing the topic-word distributions. Refer to the script to understand the parameters you need to pass to get the cleaned up output.

---

some of the words that appear in the headers and footers of the chart such as "admission date" or "namepattern" which is used to mask the patient or MD names.

Background:

In this problem set, you will train *Labeled* Latent Dirichlet Allocation (LLDA) models. Please consult the LLDA-background.pdf in the pset2materials repo for background about LLDA and additional readings.

Setup:

Dr. Thorin was invited to give a talk at the annual International Conference for Medical Liaisons (ICML) about the recent improvements in patient outcomes at Mirkwood General Hospital (MGH) which were driven by the analyses you did during your visit to MGH. Among the attendees was the world renowned cardiothoracic surgeon, Dr Aragorn. Motivated by Dr. Thorin's talk, Dr Aragorn schedules a meeting with you during which he explains that *he is interested in understanding and exploring diverse and complex subpopulations of patients who are admitted to the Cardiac Surgery Recovery Unit (CSRU) and how their status changes over time*. After hearing about how you turned unstructured chart data from MGH into profound medical insights, Dr Aragon is convinced that unstructured data will help with his exploratory task. You immediately suggest using Latent Dirichlet Allocation (LDA; also known as topic models) as an unsupervised learning technique to discover medical topics from patient charts.

Much to your dismay, Dr Aragorn looks unsatisfied. "The concept of unsupervised learning is certainly exciting but it seems like we are throwing out a lot of precious information. You see, we already know something about the patient topics. In fact, the ICD9 diagnoses that are associated with the patient stay encode what the attending physician believes to be the "patient's topic" as you call it and hence reflects useful medical insight. Can we somehow incorporate these ICD9s when we learn about new subtopics in the population?", Dr Aragorn asks.

"Finally", you think to yourself. "My moment to shine. I know just the right thing to do here: LLDA". You start explaining to Dr Aragorn what you've learned about LLDA from the background section. "LLDA is similar to LDA in that it allows us to learn topics associated with patient charts. Unlike LDA, we can use labels to guide the learning process allowing us to discover multiple subtopics associated with a given label.

You decided to breakdown Dr Aragorn's requests into two subtasks: Identifying diverse subpopulations and learning the patient state evolution.

Question 1: Subpopulation discovery

In this task, you will tackle Dr Aragorn's request to uncover and explore patient subpopulations. Here, we want the model to discover different subtopics within the same ICD9 label. To do so, we create the multi_label variable which has duplicate labels, meaning we repeat each label three times. In order to capture general topics that might exists in the population as a whole, we

create several (10) general topics for the entire population. As an example, if patient A has ICD codes = [42821, 7907], his labels become [42821 _1, 42821_2, 42821_3, 7907_1, 7907_2, 7907_3, general_1, general_2, ...general_10]. This allows the model to learn different sub-topics for different subpopulations of ICD9's. For example, you might learn different topic distributions for populations with Acute systolic heart failure (ICD9 = 428.21), based on co-morbidities [2].

Since you are concerned with the overall patient state rather than its evolution over time, you will only use the discharge charts (dc_chart) to train your model. These charts have the most concise and complete description of the patients' entire ICU stay. You are strongly encouraged to read some of the charts and familiarize yourself with their contents.

(a) List the top 20 words associated with all three topics learned for the following disease categories. Top 20 here refers to the 20 words with the highest \beta values for each of the topics.

(i) Aortic valve disorders (ICD9 4241)
(ii) Tobacco use disorder (ICD9 3051)
(iii) Urinary tract infection, site unspecified (ICD9 5990)
(iv) Bacteremia (ICD9 7907)

(b) Discuss how each of these words could be related to its corresponding disease category. You are expected to do some brief research on the meaning of words, acronyms or abbreviations which you do not understand. Give a name to each of the subtopics learned that describes the concepts captured by the subtopics. Did you expect to see these words? Why? Are there any "odd" words that you are surprised to see? If so, how would you explain to Dr Aragorn why these odd words are showing up?

(c) Explore other topics and report one that you find interesting or perplexing. Explain why you chose that topic and why it is interesting.

(d) Once the learning phase is over, we can interpret and analyze what the model has discovered by computing the posterior over the topic distribution (\theta) for each patient. This gives us an idea of the "mixture" of topics that a patient belongs to. You will find that a patient typically has a number of relevant topics with a non-negligible theta value. This number varies from patient to patient.

Dr. Aragorn is interested to know the subpopulations that two specific patients belong to: icustay id = 238491 and icustay id = 268664. List the relevant learned topics for these two ICU stays and report the corresponding expected values of theta. Note, you should report at least 5 topics for each of these ICU stays. What are the main differences

[2] The term co-morbiditiy refers to diseases that often co-occur.

between the topic distributions for these patients? Why do you think these differences exist? Look at their chart data and their additional demographic information to support your answers.

(e) Identify another ICU stay with interesting topics. Explain why you think this stay is particularly interesting and comment on the patient's topics. You may use the additional demographic data to locate patients of interest.

Question 2: Patient evolution

Next, you decided to address Dr. Aragorn second request, understanding how patients evolve over time during their ICU stay. In order to do so, you separated patient chart data by time: chart0_24 has the charts from the first 24 hours of the patient's ICU stay, while chart_inbetween has all the patient's chart from day 2 of their stay till the 24 hours before their discharge from the ICU (if there are any). Finally chart24_out has the patient's chart in the last 24 hours of their visit (including their discharge charts).

Your plan is to train a separate LLDA using each of the chart data. You explain to Dr Aragorn your reasoning for doing so: the LLDA trained on the patient charts collected during the first 24 hours (task 1), should help you discover the topics associated with the first stages of the patient state. The charts collected in between admission and discharge should help you discover topics that reflect important procedures or complications that the patient experiences during his/her stay (task 2). Lastly, the final set of charts will give you the culmination of the patient evolution (task 3). By comparing the most relevant words of the same topic across time, one can learn how subpopulations change over time. Since you're not particularly interested in discovering subgroupings of each label in this task, you choose to represent each label only once, and include one general label to capture overall population trends. This is encoded in the variable 'single_label'.

(a) List the top 20 words for each of the following diagnosis categories for each of the three time periods:
   (i) Cardiac arrest (ICD9 4275)
   (ii) Diseases of tricuspid valve (ICD 3970)
   (iii) Infection and inflammatory reaction due to cardiac device, implant, and graft (ICD9 99661)

(b) Comment on topic word composition that you see with an emphasis on the temporal patterns you observe. Explain the medical relevance of your findings. How does the topic change over time and what does that change reflect? Are the words you discovered describing treatments or patient conditions or other concepts? Again, you are expected to explain any odd words and express a general understanding of the acronyms or abbreviations you see.

(c) Dr Aragorn is interested in understanding the evolution of the ICU stay = 253003. Look at all this patient's information to piece together a coherent story of her ICU stay. How do her learned topics align with the events of her ICU stay? Are there important events during her stay that are not captured by the learned topics?

(d) Identify a patient whose trajectory over time you find interesting. Report his/her topics and the expected value of their corresponding theta parameter. Why do you find that patient interesting? You may use the additional information to identify interesting patients.

At the end of your long meeting, Dr Aragorn thanks you and quickly walks out of the conference room. He seems to join seven other people, one of them holding something resembling a ring, maybe it's a stethoscope? Eh, you have not time to ask about it. You need to leave to finish your class project.