COMP 565 Assignment 5: topic modeling in EHR

This assignment is worth 8% of your total grade and due at **23:59 on November 29, 2023**.

Download the electronic health record (EHR) data containing ICD-9 codes for a subset of the MIMIC-III ICU patients' records and the meta information about the ICD codes from here:

`https://drive.google.com/drive/folders/1L4yWqJcL93qJHfO7xCuv4vbs20hlaAcV?usp=sharing`

The data file named `MIMIC3_DIAGNOSES_ICD_subset.csv.gz` contains two columns with the first column for the patient ID (SUBJECT_ID) and the second column for the ICD-9 code (ICD9_CODE). For the purpose of this assignment, the patients were chosen to have at least one of 3 ICD-9 categories: 331, 332, 340, which correspond to Alzheimer's disease, Parkinson disease, and Multiple Sclerosis, respectively.

To make this assignment even more manageable, the codes for external injuries, supplementary classification code, and hypertension NOS (ICD-9 401.9) were removed. We will work with only 1000 ICD codes that were randomly sampled from the ICD codes of those pre-filtered patients. As a result, among the 1000 ICD codes from the data file, there are $D = 689$ unique patients based on the SUBJECT_ID and $M = 389$ unique ICD-9 codes.

The file named D_ICD_DIAGNOSES.csv.gz contain description of the ICD-9 codes in the `SHORT_TITLE` and `LONG_TITLE` columns.

File to submit

Submit your code that implements the LDA Gibbs sampling algorithm and generates the subsequent required heatmap plots in ONE file with name `COMP565_A5_LDA.ipynb` or `COMP565_A5_LDA.Rmd` via MyCourses. Write your name and student ID on the top of your script to indicate that **the work is solely yours**.

Useful plotting libraries

The heatmap plots shown below were plotted using `ComplexHeatmap` (`https://jokergoo.github.io/ComplexHeatmap-reference/book/`) in R. If you are programming in Python, you can also plot similar heatmaps using `seaborn` `https://seaborn.pydata.org/`

Review code of conduct

Review **Avoid plagiarism and cheating** policy under Content tab and make sure that you strictly follow them in completing this assignment.

1 Implementing collapsed Gibbs sampling LDA (5%)

Implement the *collapsed Gibbs sampling algorithm* discussed in class. Set the number of topics K to 5. Set the hyperparameters $\alpha = 1$ and $\beta = 0.001$ for the document topic Dirichlet prior and the ICD-9 topic Dirichlet prior, respectively.

As shown in the equations below, your implementation involves 3 key updates: (1) update the $K \times 1$ topic distribution of z_{id} for each ICD code i from each patient d, (3) update $n_{.dk}$ for the patients-by-topics count matrix (i.e., the $D \times K$ count matrix), (4) update $n_{w.k}$ for the ICDs-by-topics counts (i.e., the $M \times K$ count matrix):

$$\gamma_{idk} = (\alpha_{z_{id}} + n_{.dk}^{-(i,d)}) \left(\frac{\beta + n_{x_{id}.k}^{-(i,d)}}{\sum_w \beta + n_{w.k}^{-(i,d)}} \right), \quad p(z_{id} = k | z^{-(i,d)}, x_{id}) = \frac{\gamma_{idk}}{\sum_k \gamma_{idk}} \tag{1}$$

$$z_{id} \sim p(z_{id} = k | z^{-(i,d)}, x_{id}) \tag{2}$$

$$n_{.dk} = \sum_{i=1}^{n_d} [z_{id} = k] \tag{3}$$

$$n_{w.k} = \sum_{d=1}^{D} \sum_{i=1}^{M_d} [z_{id} = k][x_{id} = w] \tag{4}$$

Run 100 iterations of your collapsed Gibbs sampling algorithm, which will take under one minute.

Normalize the final ICDs-by-topics and the patients-by-topics matrix, respectively:

$$\phi_{wk} = \frac{\beta + n_{w.k}}{W\beta + \sum_w n_{w.k}} \tag{5}$$

$$\theta_{dk} = \frac{\alpha + n_{.dk}}{K\alpha + \sum_k n_{.dk}} \tag{6}$$

2 Visualizing the top ICD codes under each topic (1%)

For each topic k, choose the top 10 ICD-9 codes defined under the distribution $\mathbf{\Phi}_k$. Concatenate the top 10 ICD codes per topic together, resulting in a 50×5 ICDs-by-topics matrix as illustrated in Figure 1.

Because of the stochastic nature of the sampling algorithm and implementation differences, your inferred topics may differ from Figure 1 and may not clearly separate the above 3 ICD-9 categories due to disease complications and comorbidities. However, your inferred topics should be meaningful and represent clear distinction from one another. Otherwise, there are bugs in your code.

Topic 1 and topic 5 have ICD codes prefix 331 as the top code, implying their connections with Alzheimer's disease. Topic 2 has the top second code beginning with 332 (i.e., Paralysis agitans), which codes Parkinson's disease although we see comorbidity code 4280 CHF for Congestive heart failure, unspecified (`https://icdlist.com/icd-9/428.0`), appearing as the top 1 code. Interestingly, topic 4 involves code 294 w/o behavioral disturbance and the target code 340 for multiple sclerosis (among others). Topic 3 does not have any of the target codes and contain codes for infection such as urinary tract infection (5990), pneumonia (486), etc.

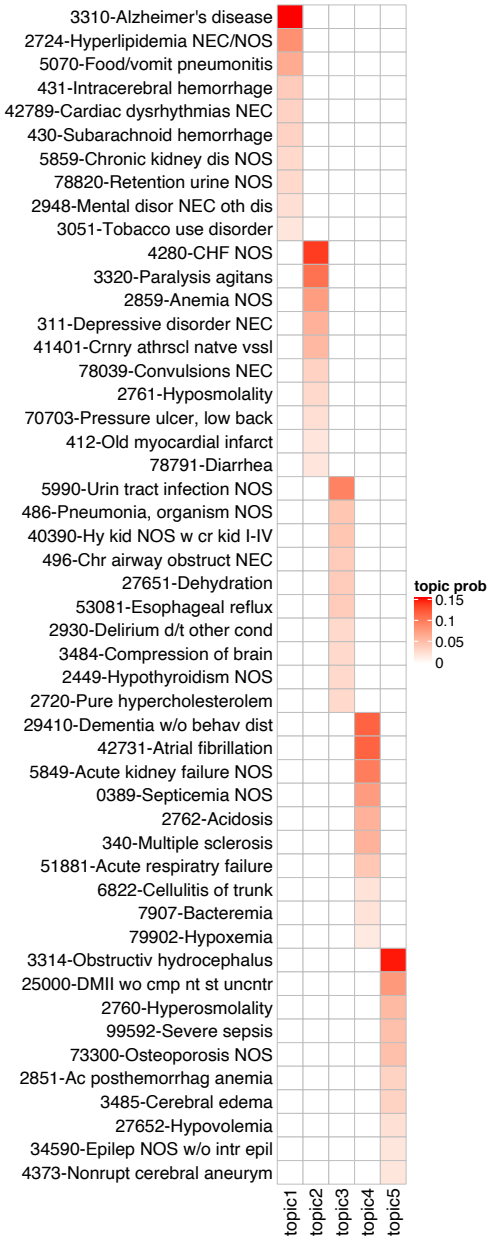

Figure 1: Latent topics inferred from the collapsed Gibbs sampling. The intensities ranges from 0 (white) to 1 (red). The rows are the top ICD-9 code under each of the 5 topics. The row names are concatenation of ICD-9 codes and the corresponding `SHORT_TITLE` from D_ICD_DIAGNOSES.csv.gz.

3 Correlating topics with the target ICD codes (1%)

To further make sense of the 5 topics, compute the normalized patient-by-topic mixture θ_{dk} for each patient d and topic k using Eq (6). Then, correlate each topic from the $N \times 5$ patient topic mixtures θ with each binary target ICD code (331,332,340) over the N patients.

Indeed, we see that topic 1 and 5 are positively correlated with ICD 331, topic 2 correlates with ICD 332, topic 4 correlates with code 340, and topic 3 does not correlate with any of the target

codes (Figure 2).

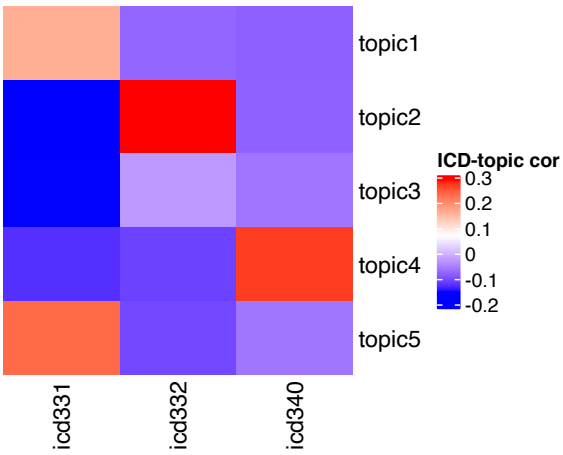

Figure 2: Topics by target ICD correlation.

4 Visualizing patient topic mixtures (1%)

Now choose top 100 patients under each topic and display them in a heatmap as shown in Figure 3. Reassuringly, we observe that the top patients with high probabilities under topic 1 and 5 are enriched for ICD codes 331, the top patients under topic 2 and 4 are enriched for ICD codes 332 and 340, respectively. In contrast, top patients under topic 3 do not have any of the 3 target ICD codes. Once again, your heatmap may differ from the one shown below but some of your topics should prioritize patients in one of the 3 disease groups.

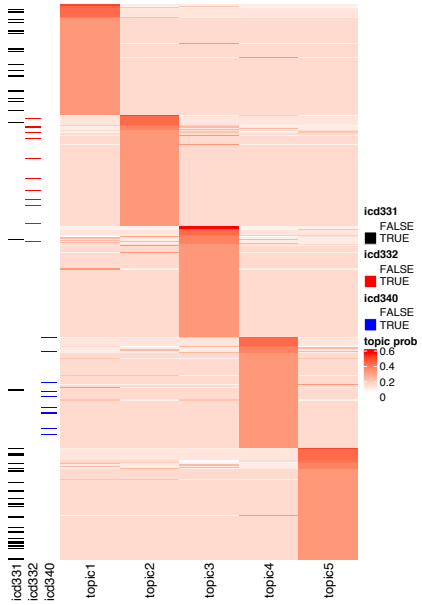

Figure 3: Top 100 patients per topic. The row annotations indicate whether each patient has the ICD code 331, 332, or 340.