3rd assignment – Clustering MIMIC data

Introduction to Clinical Data Science, 2022b

1 Preface

- We will cluster admissions based on clinical data, and we will compare the clinical data of the clusters.

2 Model Learning

1. Based on lab results (values, not presence) from the first 4 hours of ICU admission, as well as demographic data including gender, age, admission type, admission location, insurance, marital status, and ethnicity.
2. Limit to 40 most common labs, based on your selection from last assignment.
3. Use K-means algorithm to cluster admission. Next snippet can be used:

```
CREATE OR REPLACE MODEL `my_model`
OPTIONS(model_type='kmeans', kmeans_init_method =
'KMEANS++',
  num_clusters=k, standardize_features = true)
```

4. K should be set to 4.
5. **Bonus:** (10 points): Optimize K.

3 Describe the results

- Describe the distributions of demographic features (that were used as features for clustering).
- Describe what is the death probability of each cluster, and compare that to the overall probability of death.
- Description can be given graphically, or in a table.

4 CDSS

- You want to use the clusters for a clinical decision support. Suggest a support system which can be developed for each cluster. Do not stay on the theoretical leve, but suggest a data-driven diagnosis/procedure/condition for each cluster. Notice that some cluster may be hard to have such, so it's OK to skip those.

5 Notes

- Next tutorial can be a useful starting point: https://cloud.google.com/bigquery-ml/docs/kmeans-tutorial
- You don't have to use Bigquery ML for learning, you can learn using your favorit package (e.g., scikitlearn).
- Make sure to run all the notebook before submitting.
- Submission should be the notebook itself (if you used Colab, download it fist), not a URL.

- Use Moodle's forum for discussion, questions and answers. In this way everyone can benefit from the questions, answers and discussions.