COMP90089: Machine Learning Applications for Health

Semester 2, 2022
Douglas Pires, PhD; Daniel Capurro, MD, PhD

Assignment 2: Cohort characterisation and clustering

For assignment 1, you have identified a group of patients of interest (a patient cohort) and built a digital phenotype for a relevant clinical condition in critical care using MIMIC (*profound hypotension*).

Before embarking in a potentially laborious and challenging predictive modelling task (for example, employing a machine learning flavours such as deep learning or an interpretable learning approach), it might be valuable discovering simple patterns in the data that might segregate patients into subphenotypes (that might be related to different patient outcomes or increased risk). This can be achieved via Exploratory Data Analysis as well as Unsupervised Learning (in this case, clustering).

For this assignment, you will investigate the main characteristics of the cohort developed for assignment 1, identify clusters of patients and assess whether they have clinical meaning. As input, you will work on a profound hypotension patient cohort that we provide here.

The main patient characteristics that you will use include: Demographics, Co-morbidity index (Charlson), Severity of illness scores (APSIII) as well as patient outcomes, including length of stay (LoS) and in-ICU mortality.

1. **Describe via summary statistics the main properties of the *profound hypotension* cohort**. (Up to 200 words. Also provide plots for visualising the distributions).

2. **Perform *k-means* clustering to identify subgroups of patients**. Use the elbow criteria to identify an adequate number of clusters. Justify your choice. Since selecting clusters based on elbow criteria is a heuristic, you may select two different values of k for further analysis. (Up to 300 words. Provide a plot to justify the number of clusters chosen).

3. **What are the main property differences between the clusters you identified? Are they related to patient outcomes?** Provide visualisation plots that demonstrate these differences. (Up to 300 words. Also provide plots for visualising the distributions between clusters).

You are allowed to use Python visualisation libraries or your choice. We recommend using the *scikit-learn* implementation of K-means.

The provided cohort has the following **columns**:
"ID": Patient de-identified number;
"anchor_age": Age of the patient (years);
"gender": F for female and M for male;
"dod": Date of Death (if empty means survived - patient outcome variable);
"apsiii": Severity of illness score;
"LoS": Length of stay in ICU (days - patient outcome variable);
"charlson_comorbidity_index": Co-morbidity index.

Submit your analysis as a single *Jupyter notebook* with the outputs of each block of code.