IN-STK-5000, Medical Project

Christos Dimitrakakis `chridim@ifi.no`

October 12, 2018

1 Introduction

Before you start, make sure you have

- Joined one of the project groups in Piazza.
- Forked the code in `https://github.com/olethrosdc/ml-society-science`
- All questions should go through the QA platform in Piazza.

2 Historical data

First, we shall take a look at historical data present in Matlab/Octave format in `data/medical/historical.dat` For each patient, we observe the attributes $\boldsymbol{x}$, with

- $x_1 \in \{0, 1\}$, sex.
- $x_2 \in \{0, 1\}$, smoker.
- $x_{3:128} \in \{0, 1\}^4$, gene expression data. These variables can have missing values, but here they are all included in the historical data.
- $x_{129:130} \in \{0, 1\}^2$, symptoms. These can be taken to be akin to labels in supervised learning. There may be missing data here too, but for now we can assume they are included.

We also observe a therapeutic intervention $a \in \mathcal{A}$, which is followed by an outcome $y_t \in \{0, 1\}$. Consequently, historical data can be described by $(\boldsymbol{x}_t, a_t, y_t)$

Discovering structure in the data. It is uncertain if the symptoms present are all due to the same disease, or if they are different conditions with similar symptoms. (a) looking at only the attributes, estimate whether a single-cluster model is more likely than a multiple-cluster model. You can use anything, starting from a simple clustering algorithm like k-means to a hierarchical Bayesian model. (b) Try and determine whether some particular factors are important for disease epidemiology and may require further investigations.

You need to be able to validate your findings either through a holdout-set methodology, appropriately used statistical tests, or Bayesian model comparison.

Measuring the effect of actions. We also observe the effects of two different therapeutic interventions, one of which is placebo, and the other is an experimental drug. Try and measure the effectiveness of the placebo versus the active treatment. Are there perhaps cases where the active treatment is never effective, or should it always be recommended?