C3M1_Assignment

August 10, 2020

1 Estimating Treatment Effect Using Machine Learning

Welcome to the first assignment of **AI for Medical Treatment**!

You will be using different methods to evaluate the results of a randomized control trial (RCT).

You will learn: - How to analyze data from a randomized control trial using both: - traditional statistical methods - and the more recent machine learning techniques - Interpreting Multivariate Models - Quantifying treatment effect - Calculating baseline risk - Calculating predicted risk reduction - Evaluating Treatment Effect Models - Comparing predicted and empirical risk reductions - Computing C-statistic-for-benefit - Interpreting ML models for Treatment Effect Estimation - Implement T-learner

1.0.1 This assignment covers the folowing topics:

* Section **??**
* Section **??**
* Section **??**

1.1 Packages

We'll first import all the packages that we need for this assignment.

- `pandas` is what we'll use to manipulate our data
- `numpy` is a library for mathematical and scientific operations
- `matplotlib` is a plotting library
- `sklearn` contains a lot of efficient tools for machine learning and statistical modeling
- `random` allows us to generate random numbers in python
- `lifelines` is an open-source library that implements c-statistic
- `itertools` will help us with hyperparameters searching

1.2 Import Packages

Run the next cell to import all the necessary packages, dependencies and custom util functions.

```
In [25]: import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import sklearn
         import random
         import lifelines
         import itertools

         plt.rcParams['figure.figsize'] = [10, 7]
```

1 Dataset ### 1.1 Why RCT?

In this assignment, we'll be examining data from an RCT, measuring the effect of a particular drug combination on colon cancer. Specifically, we'll be looking the effect of Levamisole and Fluorouracil on patients who have had surgery to remove their colon cancer. After surgery, the curability of the patient depends on the remaining residual cancer. In this study, it was found that this particular drug combination had a clear beneficial effect, when compared with Chemotherapy. ### 1.2 Data Processing In this first section, we will load in the dataset and calculate basic statistics. Run the next cell to load the dataset. We also do some preprocessing to convert categorical features to one-hot representations.

```
In [26]: data = pd.read_csv("levamisole_data.csv", index_col=0)
```

Let's look at our data to familiarize ourselves with the various fields.

```
In [27]: print(f"Data Dimensions: {data.shape}")
         data.head()
```

```
Data Dimensions: (607, 14)
```

Out[27]:

	sex	age	obstruct	perfor	adhere	nodes	node4	outcome	TRTMT
1	1	43	0	0	0	5.0	1	1	True
2	1	63	0	0	0	1.0	0	0	True
3	0	71	0	0	1	7.0	1	1	False
4	0	66	1	0	0	6.0	1	1	True
5	1	69	0	0	0	22.0	1	1	False

	differ_2.0	differ_3.0	extent_2	extent_3	extent_4
1	1	0	0	1	0
2	1	0	0	1	0
3	1	0	1	0	0
4	1	0	0	1	0
5	1	0	0	1	0

Below is a description of all the fields (one-hot means a different field for each level): `- sex (binary): 1 if Male, 0 otherwise - age (int): age of patient at start of the study - obstruct (binary): obstruction of colon by tumor - perfor (binary): perforation of colon - adhere (binary): adherence to nearby organs - nodes (int): number of lymphnodes with detectable cancer - node4 (binary): more than 4 positive lymph nodes - outcome (binary): 1 if died within 5 years - TRTMT (binary): treated with levamisole + fluoroucil - differ (one-hot): differentiation of tumor - extent (one-hot): extent of local spread`

In particular pay attention to the `TRTMT` and `outcome` columns. Our primary endpoint for our analysis will be the 5-year survival rate, which is captured in the `outcome` variable.

Exercise 01

Since this is an RCT, the treatment column is randomized. Let's warm up by finding what the treatment probability is.

$$p_{treatment} = \frac{n_{treatment}}{n}$$

- $n_{treatment}$ is the number of patients where `TRTMT = True`
- n is the total number of patients.

```
In [28]: # UNQ_C1 (UNIQUE CELL IDENTIFIER, DO NOT EDIT)
         def proportion_treated(df):
             """
             Compute proportion of trial participants who have been treated

             Args:
                 df (dataframe): dataframe containing trial results. Column
                                 'TRTMT' is 1 if patient was treated, 0 otherwise.

             Returns:
                 result (float): proportion of patients who were treated
             """

             ### START CODE HERE (REPLACE INSTANCES OF 'None' with your code) ###
```

```
    proportion = sum(df.TRTMT==1)/len(df.TRTMT)

    ### END CODE HERE ###

    return proportion
```

Test Case

```
In [29]: print("dataframe:\n")
         example_df = pd.DataFrame(data =[[0, 0],
                                          [1, 1],
                                          [1, 1],
                                          [1, 1]], columns = ['outcome', 'TRTMT'])
         print(example_df)
         print("\n")
         treated_proportion = proportion_treated(example_df)
         print(f"Proportion of patient treated: computed {treated_proportion}, expected: 0.75")
```

```
dataframe:

   outcome  TRTMT
0        0      0
1        1      1
2        1      1
3        1      1


Proportion of patient treated: computed 0.75, expected: 0.75
```

Next let's run it on our trial data.

```
In [30]: p = proportion_treated(data)
         print(f"Proportion Treated: {p} ~ {int(p*100)}%")
```

```
Proportion Treated: 0.49093904448105435 ~ 49%
```

Exercise 02

Next, we can get a preliminary sense of the results by computing the empirical 5-year death probability for the treated arm versus the control arm.

The probability of dying for patients who received the treatment is:

$$p_{\text{treatment, death}} = \frac{n_{\text{treatment,death}}}{n_{\text{treatment}}}$$

- $n_{\text{treatment,death}}$ is the number of patients who received the treatment and died.
- $n_{\text{treatment}}$ is the number of patients who received treatment.

The probability of dying for patients in the control group (who did not received treatment) is:

$$p_{\text{control, death}} = \frac{n_{\text{control,death}}}{n_{\text{control}}}$$

- $n_{\text{control,death}}$ is the number of patients in the control group (did not receive the treatment) who died. - n_{control} is the number of patients in the control group (did not receive treatment).

```
In [31]: # UNQ_C2 (UNIQUE CELL IDENTIFIER, DO NOT EDIT)
         def event_rate(df):
             '''
             Compute empirical rate of death within 5 years
             for treated and untreated groups.

             Args:
                 df (dataframe): dataframe containing trial results.
                                 'TRTMT' column is 1 if patient was treated, 0 otherwise.
                                   'outcome' column is 1 if patient died within 5 years, 0 o

             Returns:
                 treated_prob (float): empirical probability of death given treatment
                 untreated_prob (float): empirical probability of death given control
             '''

             treated_prob = 0.0
             control_prob = 0.0

             ### START CODE HERE (REPLACE INSTANCES OF 'None' with your code) ###

             treated_prob = sum((df.TRTMT == 1) & (df.outcome == 1)) / sum((df.TRTMT == 1))
             control_prob = sum((df.TRTMT == 0) & (df.outcome == 1)) / sum((df.TRTMT == 0))

             ### END CODE HERE ###

             return treated_prob, control_prob
```

Test Case

```
In [32]: print("TEST CASE\ndataframe:\n")
         example_df = pd.DataFrame(data =[[0, 1],
                                          [1, 1],
                                          [1, 1],
                                          [0, 1],
                                          [1, 0],
                                          [1, 0],
                                          [1, 0],
                                          [0, 0]], columns = ['outcome', 'TRTMT'])
         #print("dataframe:\n")
         print(example_df)
         print("\n")
```

```
        treated_prob, control_prob = event_rate(example_df)
        print(f"Treated 5-year death rate, expected: 0.5, got: {treated_prob:.4f}")
        print(f"Control 5-year death rate, expected: 0.75, got: {control_prob:.4f}")
```

```
TEST CASE
dataframe:

   outcome  TRTMT
0        0      1
1        1      1
2        1      1
3        0      1
4        1      0
5        1      0
6        1      0
7        0      0


Treated 5-year death rate, expected: 0.5, got: 0.5000
Control 5-year death rate, expected: 0.75, got: 0.7500
```

Now let's try the function on the real data.

```
In [33]: treated_prob, control_prob = event_rate(data)

        print(f"Death rate for treated patients: {treated_prob:.4f} ~ {int(treated_prob*100)}%
        print(f"Death rate for untreated patients: {control_prob:.4f} ~ {int(control_prob*100
```

```
Death rate for treated patients: 0.3725 ~ 37%
Death rate for untreated patients: 0.4822 ~ 48%
```

On average, it seemed like treatment had a positive effect.

Sanity checks It's important to compute these basic summary statistics as a sanity check for more complex models later on. If they strongly disagree with these robust summaries and there isn't a good reason, then there might be a bug.

1.2.1 Train test split

We'll now try to quantify the impact more precisely using statistical models. Before we get started fitting models to analyze the data, let's split it using the `train_test_split` function from `sklearn`. While a hold-out test set isn't required for logistic regression, it will be useful for comparing its performance to the ML models later on.

```
In [34]: # As usual, split into dev and test set
        from sklearn.model_selection import train_test_split
        np.random.seed(18)
```