n-Sample Test Classifier on Binary Outcomes of Stratified Randomized Experiments

MA 590 Special Topics: Causal Inference

Aukkawut Ammartayakun

17 February, 2023

Two-Sample Case

Two-Sample Case

Let say we have two group of data D: D_t and D_c. We want to test whether there is a difference between the two groups.

Classifier Two-Sample Test (Lopez-Paz and Oquab 2017)

- Combined two dataset into one dataset D.
- Split the dataset into training and testing set.
- Fit the classifier (like logistic regression) to the training set and predict the testing set.
- Calculate the empirical loss l_e of the classifier. If $|l_e - 0.5| < \epsilon$, then $\bar{\tau} = 0$.

Testing the Algorithm

Let's test the algorithm with the random homogenity data.

```
set.seed(590)
# generate random multivariate gaussian data
n <- 1000
d <- 5
X <- matrix(rnorm(n*d), n, d)
# assign target variable, first half is 0, second half is 1
y <- c(rep(0, n/2), rep(1, n/2))
c(c2st(X, y)$emp_loss)
```

```
## [1] 0.52
```

Testing the Algorithm

Let's test the algorithm with the random heterogenity data.

```
set.seed(590)
# generate two dataset: two gaussians
n <- 500
d <- 5
X0 <- matrix(rnorm(n*d, 0,1), n, d)
X1 <- matrix(rnorm(n*d, 2,2), n, d)
# combine two dataset
X <- rbind(X0, X1)
y <- c(rep(0, n), rep(1, n))
c(c2st(X, y)$emp_loss)
```

```
## [1] 0.06
```

Generalization to n-Sample Case

Possible Solution

- Combine both treatment and control group within each stratum into one dataset D_i^s.
- For each group, fit the classifier (like logistic regression) to the training set and predict the testing set.
- Calculate the empirical loss l_e of the classifier. If $|l_e - 0.5| < \epsilon$, then $\bar{\tau}_{\text{within}} = 0$
- Find the way to infers $\bar{\tau}_{\text{between}}$

References

References

Lopez-Paz, David, and Maxime Oquab. 2017. "Revisiting Classifier Two-Sample Tests." In *International Conference on Learning Representations*. https://openreview.net/forum?id=SJkXfE5xx.