CPH200B: CPH Cornerstone **Winter 2025**

Project 2: Machine Learning-based Causal Effect Estimation

Instructor: Ahmed Alaa *Total points: 30 pts*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Causality is a central concept in science and medicine. The fundamental goal of medical research is to understand whether an intervention has a causal effect on the outcomes of a specific patient population. Answering such causal questions require randomization of the intervention and comparison of outcomes of treated patients against a control group. However, conducting randomized trials is not always feasible, and we often need to estimate causal effects from observational data. The goal of this project is to test your understanding of some of the causal inference concepts studied in class, as well as apply and develop machine learning-based methods for causal effect estimation.

Instructions: This is not a group project and students will be graded individually. The project has two deliverables:

- **A report summarizing your results**. The report should include point-by-point answers to the questions below. Please submit your report (in pdf format) via the bcourses website for CPH 200B.
- **A zip file with your code**. Please submit both your code and report using the bcourses website for CPH 200B. You will get feedback on both your report and code via bcourses.

Please submit your report and code by Tuesday 2/25 11:59 PST.

2.1 Warm-up Exercise: Hypothesis Testing & Confounding [6 pts]

The most basic form of causal inference involves comparing survival curves in different groups stratified by an intervention of interest. In this task, we will implement hypothesis testing methods to examine whether differences between the outcomes of treated and untreated patients are statistically significant, and whether these difference reflect the causal effect of the intervention.

For all the tasks below, we will use the UNOS heart transplant [1] dataset from Project 1.

2.1.1 Tasks and Deliverables

Please complete the following tasks. Your report should include the results for each task (i.e., tables or plots) along with your answers to the questions associated with each task.

Task 2.1.1 [1 pts]. Implement the Log-Rank test from scratch in Python. Using the UNOS dataset, apply your implemented test to check whether the survival outcomes of patients on ventricular assist device (VAD) support differ from those of patients without VAD support.

Task 2.1.2 [1 pts]. Propose a method to determine if there are confounders in the UNOS dataset for the effect of VAD support on survival outcomes. List all detected confounders.

Task 2.1.3 [2 pts]. For the comparison of survival curves to have a causal interpretation, we need to adjust for confounding variables that may cause the patient groups being compared to have different clinical features.

Propose a **propensity-weighted** version of the Kaplan-Meier estimator you implemented in Project 1 that adjusts for confounding. Plot the propensity-weighted Kaplan-Meier curves in patients with and without VAD. Compare this plot with the survival curves of both groups using the standard Kaplan-Meier estimators.

Task 2.1.4 [2 pts]. Propose a **propensity-weighted** version of the Long-Rank test. Apply this test to check whether the survival outcomes of patients on VAD support differ from those of patients without VAD. Compare the result of this test with the unadjusted test you implemented in Task 2.1.1. Comment on the results.

2.2 ML-based Estimation of Average Treatment Effects [6 pts]

2.2.1 Clinical Background, Dataset, and Setup

In this task, we will use individual patient data from the International Stroke Trial (IST), one of the largest randomized trials ever conducted in acute stroke [2]. The trial investigated the impact of aspirin and subcutaneous heparin on patients with acute ischemic stroke, with treatment randomization within 48 hours of symptom onset. The trial findings indicated no effect of both aspirin and heparin on 14-day and 6-month mortality. The trial protocol and data dictionary have been provided to you.

The original IST data lacks confounding as it was generated through a randomized trial. The instructor introduced confounding artificially by filtering patients out of the trial using a random function that depends on the patient features. The resulting dataset mimics an observational dataset where treatment is assigned through a mechanism that depends on patient features. You will conduct the following tasks using the artificially confounded dataset with the goal of recovering the same treatment effects estimated in the randomized trial.

2.2.2 Tasks and Deliverables

Estimate the average effect of aspirin and heparin on 14-day mortality using the following estimators. Compare your estimates with those of the original trial and provide commentary on the results.

Task 2.2.1 [1 pts]. A standard difference-in-means estimator.

Task 2.2.2 [1 pts]. An inverse propensity weighting (IPW) estimator using a Gradient Boosting model for the propensity scores.

Task 2.2.3 [2 pts]. A covariate adjustment estimator using a Gradient Boosting model with T-learner, S-learner, and X-learner architectures.

Task 2.2.4 [2 pts]. An augmented IPW (doubly-robust) estimator that combines the propensity model from Task 2.2.2 and an outcomes model based on the S-learner in Task 2.2.3.

2.3 Counterfactual Inference and Domain Adaptation [8 pts]

In this task, we will explore the application of concepts from the machine learning literature to estimate heterogeneous treatment effects. The seminal work in [3] establishes a link between estimating treatment effects and the domain adaptation problem in machine learning. Using this insight, the authors repurpose ideas from domain adaptation literature to create a new deep learning model for estimating the conditional average treatment effects (CATE) function. The core idea of their algorithm is to eliminate confounding bias by learning a representation Φ of the features X that aligns the distribution of treated and control populations, $\Phi(X|T=1)$ and $\Phi(X|T=0)$, in the representation space, referred to by the authors as a "balancing" representation.

Please read the paper carefully and complete the following tasks.

Task 2.3.1 [3 pts]. Implement the $TARNet$ and CFR_{MMD} models proposed in [3] in PyTorch. Evaluate the performance of all models using the semi-synthetic benchmark dataset included in the Project 2 notebook.

Task 2.3.2 [1 pts]. Visualize the treated and control features before and after applying the balancing representation $\Phi(.)$ using t-SNE. Comment on the results.

Task 2.3.3 [1 pts]. Show the impact of the scaling parameter α (Eq. (3) in [3]) on the loss function on the test set for the Maximum Mean Discrepancy (MMD) regularizer.

Task 2.3.4 [3 pts]. Use the $TARNet$ and CFR_{MMD} models to estimate average treatment effects using the IST data in Task 2.2. Assess the alignment of your estimates with the trial results and compare them to the estimators in Tasks 2.2.3 and 2.2.4.

2.4 NeurIPS Reviewer for a Day: Reviewing & Reproducing Recent Research on ML-Based Causal Inference [10 pts]

In this task, we will focus on one paper that proposes new methods for estimating CATE inspired by ideas we studied in Lectures 7, 8 and 9. The paper is *"Adapting Neural Networks for the Estimation of Treatment Effects"* by Claudia Shi, David Blei and Victor Veitch, which was published in NeurIPS 2019. The objective of this task is to develop critical paper review skills and practice reproducing research results. Please read the paper carefully and complete the following tasks.

Task 2.4.1 [5 pts]. Please review the NeurIPS 2024 reviewing guidelines and write a comprehensive review of this paper in accordance with those guidelines.

Task 2.4.2 [5 pts]. Implement the `DragonNet` and `Targeted regularization` methods proposed in this paper in PyTorch and reproduce their performance results on the IHDP dataset (Table 1 in the paper).

References

[1] Weiss, Eric S., Lois U. Nwakanma, Stuart B. Russell, John V. Conte, and Ashish S. Shah. "Outcomes in bicaval versus biatrial techniques in heart transplantation: an analysis of the UNOS database." *The Journal of heart and lung transplantation*, vol. 27, no. 2 (2008): 178-183.

[2] International Stroke Trial Collaborative Group. "The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke." *The Lancet*, 349.9065 (1997): 1569-1581.

[3] Shalit, U., Johansson, F. D., and D. Sontag. "Estimating individual treatment effect: generalization bounds and algorithms." In International conference on machine learning (pp. 3076-3085). 2017.