Home Assignment 2

Data Management Algorithm for Decision Making

December 10, 2025

General Instructions

Please submit a zip file named "Id1_Id2.zip" via ’webcourse’ course site with the following files:

- **Written Questions:** PDF file "HW2_Dry.pdf" (typed and in English).
- **Your Code ’.py’ files - there are 4**
- **Your Code notebook:** "`EX2.ipynb`" (Jupyter Notebook).
- **Execution Results (for control only, all the results should be reported in the PDF file):** "`EX2_Result.html`" HTML file with code output.

Instructions for Saving Jupyter Notebook as HTML

Follow these steps:

1. Download the notebook file (‘.ipynb‘).
2. Open a new Jupyter Notebook and upload the downloaded notebook file to your workspace.
3. Run the following code in the new notebook:

```
import nbformat
from nbconvert import HTMLExporter
from nbconvert.preprocessors import ExecutePreprocessor

with open("HW2_Code.ipynb") as f:
    notebook_content = nbformat.read(f, as_version=4)

ep = ExecutePreprocessor(timeout=600, kernel_name='python3')
ep.preprocess(notebook_content)

html_exporter = HTMLExporter()
html_data, _ = html_exporter.from_notebook_node(notebook_content)

with open("HW2_Result.html", "w") as f:
    f.write(html_data)
```

Note: You can adjust the ‘timeout‘ parameter in the code above if needed. There is no requirement for a specific runtime in this exercise. **Make sure that all the results match those you report in your PDF**

- Explain your solutions in your own words. Minimize using external *LLMs* tools (e.g., ChatGPT).
- If you use information not covered in the course slides, provide a proper reference.
- If you consulted with classmates, mention who you consulted with. Collaboration is fine as long as you can explain the solution independently.

- Be honest and fair in your work. Ensure you fully understand everything you submit.

Questions

Problem 1: Causal inference (40 points). Consider the causal DAG for the Stack Overflow developers' survey dataset in Figure 1 (the dataset is provided).

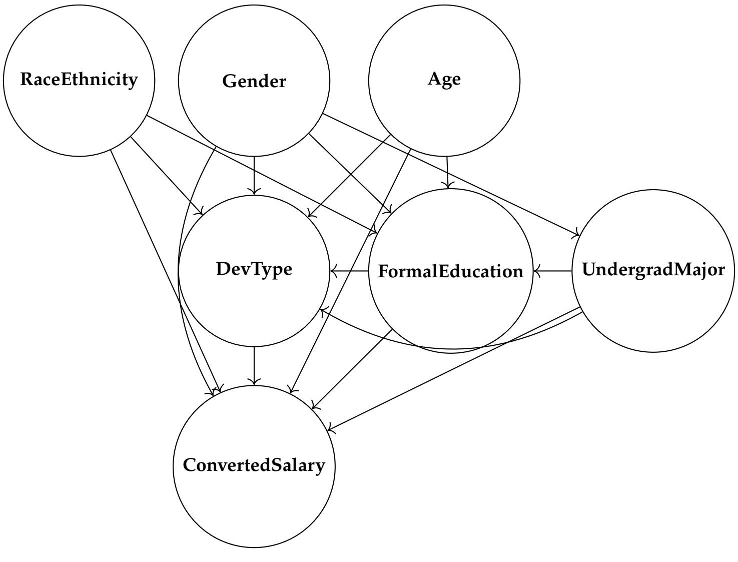


Figure 1: Partial causal DAG for the Stack Overflow dataset.

1. (5 points) For each variable pair (X,Y), find a set of variables (nodes) that blocks all backdoor paths between X and Y. Explain your answer.
   (a) X = DevType, Y = ConvertedSalary
   (b) X = Gender, Y = ConvertedSalary
   (c) X = UndergradMajor, Y = ConvertedSalary
2. (5 points) The set of parents of the treatment variable X always satisfies the backdoor criterion. Explain why this statement is correct. Does this always provide the minimal adjustment set? If yes, explain your answer. Otherwise, give an example.
3. Calculate the average treatment effects (ATE) using the provided code. Your task is to compute the average treatment effect for higher education on salary. Note that the code converted the education column into a binary column.
   (a) (5 points) Complete the code to find the confounding variable set given a treatment, outcome, and a causal DAG. Check your function on the given example DAG, treatment and outcome. **Report your results. For this example, mention one backdoor path that is blocked by one of your confounders.**
   (b) (5 points) Complete the code to compute the empirical ATE - namely, the ATE computed directly on the data without using any estimator, according to the ATE formula:

$$\text{ATE} = \mathbb{E}[Y \mid do(T = 1)] - \mathbb{E}[Y \mid do(T = 0)] = \sum_{c \in C} P(C = c) \left(\mathbb{E}[Y \mid T = 1, C = c] - \mathbb{E}[Y \mid T = 0, C = c]\right)$$

   Where:

   - T is the binary treatment
   - Y is the numeric outcome
   - C is the set of all confounder strata

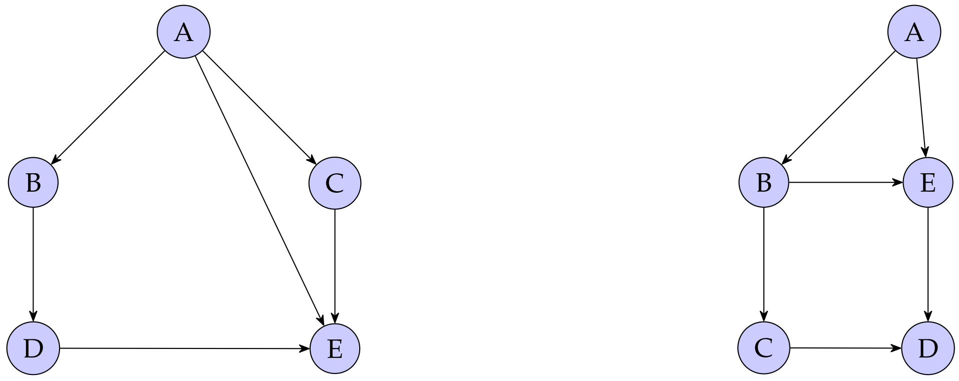
- $P(C = c)$ is the empirical probability of stratum c in the data

Report your results. Explain in your own words the meaning of the number you got. What is the conclusion we get from this number?

(c) (10 points) Run the code to estimate the ATE using linear regression and propensity score weighting estimators. **Report your results. Explain the differences between the values you get. Which estimate is likely the most accurate, and why?**

(d) (10 points) Repeat the process above with another chosen variable as the treatment. If this variable is not already binary, convert it to a binary variable as you see fit. **Report your results. Which variable did you choose? How did you convert it to binary (rules/threshold)? Explain the differences between the values you get from the different estimation methods. Which estimate is likely the most accurate, and why?**

Problem 2: *d*-separation (20 points). Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be three disjoint sets of variables. We say $\mathbf{X}$ and $\mathbf{Y}$ are d-separated given $\mathbf{Z}$, written as $\mathbf{X} \perp\!\!\!\perp_d \mathbf{Y}|\mathbf{Z}$ if every backdoor path between them is blocked. To determine whether X and Y are d-separated by Z: (1) Identify all backdoor paths between X and Y in the graph. (2) Check each path to see if it is blocked by Z.



(a) DAG 1.

(b) DAG 2.

Figure 2: Causal DAGs for problem 2.

1. (5 points) Choose two pairs of nodes (one from DAG 1 and one from DAG 2); what ~~is the adjustment set that blocks them~~ is a set that d-separates them, or, if there is no such set, explain why

2. (15 points)

   d-separation Axioms:

   (a) **Symmetry:**
   $$\mathbf{X} \perp\!\!\!\perp_d \mathbf{Y}|\mathbf{Z} \Leftrightarrow \mathbf{Y} \perp\!\!\!\perp_d \mathbf{X}|\mathbf{Z}$$

   (b) **Decomposition:**
   $$\mathbf{X} \perp\!\!\!\perp_d (\mathbf{Y} \cup \mathbf{W})|\mathbf{Z} \Rightarrow (\mathbf{X} \perp\!\!\!\perp_d \mathbf{Y}|\mathbf{Z}) \text{ and } (\mathbf{X} \perp\!\!\!\perp_d \mathbf{W}|\mathbf{Z})$$

   (c) **Weak Union:**
   $$\mathbf{X} \perp\!\!\!\perp_d (\mathbf{Y} \cup \mathbf{W})|\mathbf{Z} \Rightarrow (\mathbf{X} \perp\!\!\!\perp_d \mathbf{Y}|\mathbf{Z} \cup \mathbf{W}) \text{ and } (\mathbf{X} \perp\!\!\!\perp_d \mathbf{W}|\mathbf{Z} \cup \mathbf{Y})$$

   (d) **Contraction:**
   $$(\mathbf{X} \perp\!\!\!\perp_d \mathbf{Y}|\mathbf{Z}) \text{ and } (\mathbf{X} \perp\!\!\!\perp_d \mathbf{W}|\mathbf{Z} \cup \mathbf{Y}) \Rightarrow \mathbf{X} \perp\!\!\!\perp_d (\mathbf{Y} \cup \mathbf{W})|\mathbf{Z}$$

   (e) **Intersection:**
   $$(\mathbf{X} \perp\!\!\!\perp_d \mathbf{Y}|\mathbf{Z} \cup \mathbf{W}) \text{ and } (\mathbf{X} \perp\!\!\!\perp_d \mathbf{W}|\mathbf{Z} \cup \mathbf{Y}) \Rightarrow \mathbf{X} \perp\!\!\!\perp_d (\mathbf{Y} \cup \mathbf{W})|\mathbf{Z}$$

   Given the axioms above, prove or disprove (with a counterexample) the following statements:

   (a) If $\mathbf{X} \perp\!\!\!\perp_d (\mathbf{Y} \cup \mathbf{A})|\mathbf{Z}$ Does this imply $\mathbf{X} \perp\!\!\!\perp_d \mathbf{Y}|(\mathbf{Z} \cup \mathbf{A})$?

   (b) If $\mathbf{X} \perp\!\!\!\perp_d \mathbf{Y}|(\mathbf{Z} \cup \mathbf{A})$ and $\mathbf{X} \perp\!\!\!\perp_d \mathbf{A}|(\mathbf{Z} \cup \mathbf{Y})$ Does this imply $\mathbf{X} \perp\!\!\!\perp_d \mathbf{Y}|\mathbf{Z}$?

(c) If $\mathbf{X} \perp\!\!\!\perp_d \mathbf{Y}|\mathbf{Z}$ and $\mathbf{Y} \perp\!\!\!\perp_d \mathbf{Z}|\mathbf{X}$ Does this imply $\mathbf{X} \perp\!\!\!\perp_d \mathbf{Z}|\mathbf{Y}$?

Problem 3: Frequent Itemsets (20 points). **Apriori for data exploration**. We will use the Apriori algorithm to explore the Loan-Approval-Prediction data.

1. (5 points) To run the Apriori algorithm on a dataset, we first need to convert it into a list of transactions (e.g., a set of items a customer bought together). For example, if the columns are A and B and the values of the first row are (a, b), the corresponding transaction should be: (A:a, B:b). Complete the code to transform a given dataset into a list of transactions, which can then be used as input for the Apriori algorithm. **Report five examples from your output results for the example with 'loan_approval_dataset'.**

2. (5 points) Use the Apriori algorithm to find an example of an over-represented subpopulation: Use the Apriori algorithm to identify a subgroup with a high proportion compared to the complement subpopulation. For instance, assume there is an eye color column. You could find that more than 80% of the population have blue eyes (meaning that only 20% have green or brown eyes). **Report your results and explain the process you used to discover this subpopulation. You are allowed to discretize numerical columns (e.g., bin the Age column for example), what did you find?**

3. (10 points) Use the Apriori algorithm to find a property (pattern) associated with approved loan status: Apply the Apriori algorithm to discover a property (e.g., Eye color = brown) that frequently appears with approved loan status.**Report your results, explain your findings and the steps you took to identify this association. You are allowed to discretize numerical columns.**

Problem 4: Interesting Visualizations (20 points).

1. Choose a Column: Select a column from the "Loan-Approval-Prediction" dataset. If the column is numerical (e.g., Age), you may discretize it.

2. Plot Overall Data: Plot a graph showing the column you picked versus the average loan-status using the entire dataset. You are allowed to pre-process the data.

3. (10 points) Identify a Divergent Subpopulation: Find a subset of the data (defined by a pattern) where the trend of the column you pick versus average loan-status significantly deviates from the trend observed in the overall data. Measure this divergence using the Kullback-Leibler (KL) divergence code provided. For instance, if the overall data shows that the average loan status of people with blue eyes is higher than that of people with brown eyes, but among residents of Berlin, the trend is reversed, this subset represents a divergent subpopulation. A subpopulation should be defined by a conjunction of attribute-value assignments (e.g., `Eye color = Blue` or `City = Berlin AND Height >= 150 cm`). You should report the KL divergence score you got and a short explanation of your findings.

Important! Report all the results for problem 4 here in the PDF report. Attach any visualization you wish to discuss, describe your steps, and explain the result. You will not get points for any information that was written in the notebook but was not reported here.