UDS Midterm Study Guide

Yabei Zeng

Table of contents

1 UDS Midterm Study Guide

1.1 Proscriptive vs. Descriptive Questions

- Descriptive Questions: Questions about the **state of the world** and about objective reality (have right or wrong answers) include the following examples:
  - "What kinds of users are clicking our ads?
  - "Do high-income and low-income countries emit similar amounts of carbon dioxide?"
- Proscriptive Questions: Questions about **how the world should be** , don't have right or wrong answers because answers to proscriptive questions require evaluating the desirability of possible outcomes, include the following examples:
  - "Should higher income and low income countries be expected to meet the same carbon emission reduction standards?"
  - "Do high-income countries have a moral obligation to provide tuberculosis drugs to developing countries for free (or at cost)?"

1.2 Exploratory Questions

- Using Exploratory Questions: questions about elicit information, questions about broader patterns in the world, could be answered by simple summary statistics or plots, answering the exploratory questions makes you understand the contours of the problem you seek to solve, include the following examples:
  - What type of buildings (industrial, residential, commercial) consume the most power in the US?
  - In what region of the US are buildings consuming the most power?
  - Is there a region of the US where buildings are generating the most CO2?
  - Does the average energy use per building vary by region or building type?
  - In what season is most building energy consumed? Is more energy consumed by heating or AC needs, or do the two use similar amounts of power?

1.2.1 Internal versus External Validty

- Internal validity: how well you have analyzed the data you have
- External validity: how well you expect the answer you generated from that data to generalize to your stakeholder's context
- Internal validity concerns
  - Relate to your ability to properly characterize the causes of severe accidents in this data, accidents causes you generate from this data are meaning and faithful representations of the patterns in the data

  - Concerns over the accuracy with which things are measured

- External Validty Concerns
  - New features rolls out, whether there was anything exceptional about the data generating process

- The more control one has over a study context, the more likely one is to have good internal validity, but that control can often create an artificially to limits external validity, Thus internal and external validity should not necessarily be thought of as things to try and simultaneously maximize at all costs, rather, they are best thought of as distinct features of analysis that should always be considered

1.3 Passive Prediction Questions

- Passive prediction questions: about **the future or potential** outcomes of individuals entities, including examples:
  - "How likely is patient X to experience a heart attack in the next two years?"
  - "How likely is it that Mortgage Holder Y will fail to make their mortgage payment next month?"

- Passive Prediction don't usually have "an answer", need by considering the feasibility of developing a model to give individual-level answers to a Passive Prediction Question
- Differentiating between exploratory and passive prediction questions
  - Passive prediction questions focus on the values that get spit out of a model for each entity in the data, the only thing we care about in the passive prediction is **the quality of these predictions**
  - With exploratory question, our interest is in improving our understanding of the problem space

- Passive prediction and causal questions
  - Both questions are trying to predict future outcome
  - The difference is that passive prediction are about make accurate prediction, not causal relationships that we can directly manipulate to shape outcomes (causal questions)
  - Correlation does not necessarily imply causation

- **Internal validity and external validity** in answering passive prediction questions
  - Internal validity: measure of how well a model captures the meaningful variation in the data we already have
  - External validity: measure of how well we think that our model is likely to perform when faced with new data

  - Some of the factors that influence the External Validity to Passive Prediction Questions are the same as those that shape the External Validity of Exploratory Question
    * Population represented in the data
    * The time period in the questions
  - Two external validity concern (different from exploratory questions):
    * Overfitting
    * Adversarial users: attempt to subvert a statistical or machine learning model

- Not Using Black Box:
  - Problem with black box machine learning models: these models are used in high-stakes areas like `healthcare` and `criminal justice` but lack transparency, leading to severe consequences due to their inaccessibility and complexity
  - Explanation towards the black box could be problematic, lacking fidelity and can be misleading
  - Well-structured data with meaningful features can be equally well-predicted by simpler, interpretable models

- pyGAM
  - Generalized Additive Models(GAMs)
    * $g(\mathbb{E}[y|X]) = \beta_0 + f_1(X_1) + f_2(X_2, X_3) + ... + f_M(X_N)$
    * $X_1, X_2, ..., X_N$ are independent variables, and y is the dependent variable
    * $g()$ is the link function

1.4 Causal Questions

- Causal questions: designed to help us predict the consequences of our actions
- Stakeholder will ask a causal question when they want to know whether an action may be beneficial
- An answer to causal questions will involve designing a study that not only measures the effect of treatment but also knows any measured effect will generalize to the stakeholder's context.
- Internal and External Validity
  - Internal: the study itself in the context of studying
  - External: The context in which the stakeholder wishes to generalize the result

- Answering causal questions:
  - An action X on an outcome Y
  - `Fundamental Problem of Causal Inference`: directly measure the causal effect of X on Y for a given entity
  - `Randomized experiment`: most familiar tool for answering causal questions
    * Randomized Control Trials (RCTs) or A/B Tests

1.4.1 Potential Outcomes Framework

- T is the treatment, a binary treatment, meaning it can only take on two values : {0,1}
- Y_i is the potential outcome variable for i individual, ${Y_i}^1$ is the treated potential outcome variable for i individual
- D is the observed receipt of either treatment or control, takes two values {0, 1}
- Given value of D, we will only be able to observe
  - Y_i^0 if $D_i = 0$ or Y_i^1 if $D_i = 1$
- Defining the causal effect:
  - $\delta_i = Y_i^1 - Y_i^0$
  - The difference between the individual potential outcome
  - This quantity is not able to observe
  - For all individuals, the value is call ATE
    - $ATE = \frac{1}{N} \sum_{i \in 1,2,3...N} \delta_i$
    - $ATE = \frac{1}{N} \sum_{i \in 1,2,3...N} Y_i^1 - Y_i^0$
    - If we assume our data is a random sample from the population, then we can write the ATE in the following form
      - $ATE = E(\delta_i)$
      - $ATE = E(Y_i^1 - Y_i^0)$
      - $ATE = E(Y_i^1) - E(Y_i^0)$
    - The binary treatment
      - $E(Y_i^1|D_i = 1) - E(Y_i^0|D_i = 0)$
      - This is a exact quantity you get from a linear regression of Y on D
- Correlation and Causation
  - $\hat{ATE} = E(Y_i^1|D_i = 1) - E(Y_i^0|D_i = 0) + E(Y_i^0|D_i = 1) - E(Y_i^0|D_i = 1)$
  - because $E(Y_i^0|D_i = 1) - E(Y_i^0|D_i = 1) = 0$, therefore $\hat{ATE} = E(Y_i^1|D_i = 1) - E(Y_i^0|D_i = 0)$
  - We can also write $\hat{ATE}$ as
    - $\hat{ATE} = E(Y_i^1|D_i = 1) - E(Y_i^0|D_i = 1) + E(Y_i^0|D_i = 1) - E(Y_i^0|D_i = 0)$
    - where $E(Y_i^1|D_i = 1) - E(Y_i^0|D_i = 1)$ is ATT
    - and $E(Y_i^0|D_i = 1) - E(Y_i^0|D_i = 0)$ is the baseline difference
- Average Treatment Effect on the Treated (ATT)
  - $E(Y_i^1|D_i = 1) - E(Y_i^0|D_i = 1)$
  - $ATE = E(Y_i^1) - E(Y_i^0)$
  - $= \lambda \left(E(Y_i^1|D_i = 1) - E(Y_i^0|D_i = 1)\right) + (1 - \lambda) \left(E(Y_i^1|D_i = 0) - E(Y_i^0|D_i = 0)\right)$
  - where $E(Y_i^1|D_i = 1) - E(Y_i^0|D_i = 1)$ is the ATT (Average treatment effect on the treated)
  - where $E(Y_i^1|D_i = 0) - E(Y_i^0|D_i = 0)$ is the average treatment effect on the untrated

  - ATT = ATE when $E(Y_i^1|D_i = 1) - E(Y_i^0|D_i = 1) = E(Y_i^1|D_i = 0) - E(Y_i^0|D_i = 0)$, meaning no differential treatment effects
  - neither $E(Y_i^0|D_i = 1)$ nor $E(Y_i^0|D_i = 1)$ are observable

WHAT is `SUTVA`?

- The Stable Unit Treatment Value Assumption (SUTVA) is a fundamental principle in causal inference that assumes `the treatment given to one unit does not affect the outcomes of other units.` This assumption comprises three main ideas:
  - `homogeneity of treatment effects`: each unit receives the same treatment effect
  - `No spillovers or externalities`: The treatment of one unit does not influence the outcomes of another unit.
  - `No general equilibrium effects`: The assumption that the treatment does not lead to broader systemic changes that could affect the outcomes.

1.4.2 Indicator Variables

`Indicator variables` :sometimes referred to as dummy variables, takes values of 0 and 1, used to indicate whether a given observation belongs to a discrete category in a way that can be used in statistical models.

- the coefficient on an indicator variable is an estimate of the average **DIFFERENCE** in the dependent variable for the group identified by the indicator variable
- the **REFERRENCE GROUP**, is the set of observations for which the indicator variable is always zero.
- the coefficient on an indicator variable is an estimate of a **DIFFERENCE** with respect to a **REFERENCE GROUP**
- For categorical variables with more than 2 categories, use `one-hot encoding`

1.4.3 Experiment

Internal Validity