Causal Inference I

Mixtape Session

MIXTAPE
SESSIONS

Roadmap

- Causal Forests
  - Introduction to causal forests
  - Unconfoundedness and common support assumptions
  - A Brief History of Trees
  - Fundamental Machine Learning Concepts

Causal Forests: Bridging Machine Learning & Causal Inference

Causal forests are . . .

- A powerful tool for estimating heterogeneous treatment effects using tree-based methods.
- Combining the strengths of random forests with the principles of causal inference to uncover nuanced relationships.
- Addressing challenges in observational data: leveraging the unconfoundedness assumption and advanced modeling techniques.

Origins of Causal Forests

- Birthed from the intersection of causal inference and machine learning.
- Preceded by methods like propensity score matching, regression discontinuity, and instrumental variables.
- Motivated by the need for more flexible, non-parametric methods to estimate heterogeneous treatment effects.

Athey and Wager: Pioneering Work on Causal Forests

- Susan Athey & Stefan Wager’s foundational work: ”Estimation and Inference of Heterogeneous Treatment Effects using Random Forests” in 2019 JASA
- Addressed challenges in traditional methods: Bias-variance trade-off and overfitting.
- Introduced causal trees as building blocks for causal forests, leveraging bootstrapping and honest splitting.

Overfitting in Propensity Score Estimation

- Estimating propensity scores often involves logistic regression with many covariates.
- Overfitting can arise with a large set of covariates relative to sample size or with high-degree interactions.
- Overfitted models yield scores too close to 0 or 1, complicating matching.

Nearest Neighbor Matching & Common Support

- Common support ensures overlap in propensity score distributions.
- Yet, nearest neighbor matching might pair units not very close in propensity scores.
- Especially problematic when untreated units greatly outnumber treated ones.

Quality of Matches

- Common support as measured by propensity scores doesn’t guarantee high-quality matches.
- This is because common support should be a concept we are thinking of as holding, not in the propensity score, but in the actual stratification of the data using the dimensions of the covariates
- Units with similar propensity scores might differ on key covariates given the curse of dimensionality kicks in very quickly as we increase covariates with high dimensions.
- Matching without replacement can lead to lower-quality subsequent matches.

Sensitivity to Specification

- Recall that we are *estimating* the propensity score; we don’t know the truth even if we know the covariates to use for estimation
- Match quality and causal estimates can be sensitive to propensity score model specification.
- Minor model changes can lead to different matched samples and different matched samples means we have variation in treatment effect estimation that is due to these matching irregularities
- Highlights the importance of robustness checks.

Challenges of High-Dimensional Covariate Spaces

- Modern datasets often have a vast number of covariates, making the ”curse of dimensionality” a prominent concern.
- Even with a few covariates, the curse can arise, but it’s especially pronounced in high-dimensional settings.
- While the assumption of unconfoundedness requires controlling for all confounders, in practice, ensuring this in high dimensions is challenging.
- The ”kitchen sink” approach, adding numerous covariates to control for confounding, can introduce its own problems.
- Causal forests offer a flexible way to address these challenges, capturing complex relationships without the need for overly restrictive linear specifications.

Heterogeneous Treatment Effects

- Causal forests estimate heterogeneous treatment effects, capturing variations across subgroups.
- In high-dimensional settings, treatment effects can vary across many dimensions.
- Enables insights into how treatment effects manifest in different segments of the data.

Adaptive Partitioning

- Trees partition data based on covariate values.
- Causal forests adaptively find splits relevant for estimating treatment effects.
- ”Zooming in” on areas with pronounced or variable treatment effects.

Regularization

- Random forests introduce randomness via bootstrapping and random subsets of predictors.
- This randomness acts as a form of regularization.
- Helps prevent overfitting in high-dimensional settings.

Key Challenges in Causal Inference

- Confounding: Hidden biases that can skew results.
- Selection bias: Non-random assignment to treatment.
- Measurement error: Inaccuracies in data collection.

If we can address the problem using covariates, then we may be in a situation to use causal forests (but not all problems fit this scenario)

Assumption of Unconfoundedness

- Unconfoundedness: Treatment is independent of potential outcomes (i.e., as good as random) for units with identical covariate values or “strata”.

$$(Y^1, Y^0) \perp\!\!\!\perp D \mid X$$

- Causal forests are in the “branch” of causal inference that assumes unconfoundedness (like DML, propensity scores, regression and matching)
- But their strength is in their adaptability to large dimensions which makes them powerful tools under this assumption.

Common Support in High Dimensions

- After unconfoundedness, we need “common support” – non-empty cells for the entire stratification of the data based on the covariates in treatment and control
- Unconfoundedness says we are allowed to use covariates for causal inference, but common supports says it’s actually possible to do it because we have units in treatment and control for all dimensions of X
- In high-dimensional settings, ensuring common support becomes challenging though (slide after next for more)

Target Estimand: The Conditional Average Treatment Effect (CATE)

- Definition: $\tau(x) = \mathbb{E}[Y^1 - Y^0 | X = x]$
- Represents the difference in potential outcomes for treated vs. untreated units, conditioned on covariate values x.
- Allows us to capture heterogeneous treatment effects across different subgroups.
- Causal forests excel at estimating CATE by leveraging the adaptiveness of trees to capture interactions and non-linearities in the relationship between covariates and treatment effects.

Example: Breakdown of Common Support

- Consider two binary covariates, sex (male/female) and age (adult/child).
- Even if sex and age individually have overlap, the joint distribution (e.g., sex=0, age=1) might lack overlap.
- With more covariates, gaps in joint distribution become more probable: the ”curse of dimensionality.”

Table 1: Stratified sample with common support

Table: Counts and Titanic survival rates by strata and first class status.

	First class		**All other classes**		
Strata	Obs	Mean	Obs	Mean	Total
Male adult	175	0.326	1,492	0.188	1,667
Female adult	144	0.972	281	0.626	425
Male child	5	1	59	0.407	64
Female child	1	1	44	0.613	45
Total observations	325		1,876		2,201

Table 2: Stratified sample without common support

Table: Counts and Titanic survival rates by strata and first class status.

	First class		All other classes		
Strata	Obs	Mean	Obs	Mean	Total
Male adult	175	0.326	1,492	0.188	1,667
Female adult	144	0.972	281	0.626	425
Male child	5	1	59	0.407	64
Female child	0	n/a	44	0.613	44
Total observations	324		1,876		2,200

Strata vs. Partition

- Historically, in many statistical contexts, we use ”strata” and ”stratification” to refer to dividing data into subsets based on certain criteria.
- In the literature on causal forests and decision trees, the term ”partition” is more commonly used.
- This terminology traces back to the computer science and machine learning origins of decision trees, where data is ”partitioned” into subsets during the tree-building process.

Why the Difference?

- ”Stratification” often implies a deliberate, predefined division of data based on known, important criteria.
- ”Partitioning” in trees is more dynamic and adaptive, with divisions made based on data-driven decisions to optimize a specific objective.
- While they conceptually overlap, the terms have different historical and contextual connotations.

A Word of Caution

- As we delve into causal forests, be mindful of the term ”partition” and its usage.
- It aligns closely with ”stratification,” but remember the adaptive, data-driven nature of partitioning in this context.
- Terminology can sometimes be a barrier, but understanding the underlying concepts bridges the gap.

Causal Forests: Addressing the Challenge

- Causal forests partition the covariate space, estimating effects within partitions.
- Allows for local treatment effect estimation where there’s overlap.
- Highlights regions where common support might be violated.

Origins of Decision Trees: The 1960s

- The foundational concepts of decision trees emerged in the 1960s.
- Initially explored in cognitive psychology to model human decision processes.
- Used in medical decision-making to aid doctors in diagnosing diseases based on a hierarchical structure of symptoms and outcomes.
- These early trees were simplistic and manually constructed, but they paved the way for algorithmic tree-building methods in the subsequent decades.
- The concept gained traction as it offered a visual and interpretable way to make decisions based on multiple criteria.

Early Beginnings

- Ross Quinlan’s ID3 in the 1980s: pioneering tree algorithm.
- ID3 uses an information-theoretic approach: It selects the attribute that provides the best split (maximizing information gain) at each node, building the tree iteratively.
- Real-world example: Diagnosing medical conditions. ID3 could be applied to a dataset where symptoms are attributes and diseases are classes. The tree would guide medical professionals by asking about the most informative symptoms first, helping narrow down potential diagnoses.

Classification and Regression Trees (CART)

- Introduced by Breiman et al. in 1986.
- CART allows for the creation of binary trees for both classification (categorizing data into classes) and regression (predicting numerical values).
- Uses a ”greedy” approach: At each step, it selects the best split based on a specific criterion (like Gini impurity for classification or mean squared error for regression) without concern for future decisions.
- Real-world example: Predicting housing prices. Using a dataset with features like house size, location, and age, CART can be used in regression mode to predict the price of a house based on these attributes.

Ensemble Methods and Random Forests

- Leo Breiman introduced Random Forests in 2001.
- Random Forests are an ensemble method: They combine predictions from multiple decision trees to produce a more robust and accurate result.
- The method introduces randomness in two ways: By bootstrapping samples for training each tree and by selecting a random subset of features at each split.

Ensemble Methods and Random Forests Example

- Real-world example: Credit scoring.
- Banks use Random Forests to predict the likelihood of a loan applicant defaulting.
- The model takes into account various factors like income, employment history, and credit score, aggregating insights from multiple trees to assess risk.

Boosting and Gradient Boosted Trees

- Boosting introduced by Schapire in 1990; later refined by Freund; is an ensemble technique that focuses on reducing bias by giving more weight to misclassified instances in subsequent models.
- Gradient Boosted Trees introduced by Friedman in the late 1990s and early 2000s builds trees sequentially, where each tree tries to correct the errors made by the previous ones and uses gradient descent to minimize the loss function.
- Both methods aim to improve prediction accuracy by combining weak learners (models slightly better than random guessing) to form a strong learner (a model with high predictive accuracy).

Boosting and Gradient Boosted Trees Example

- Real-world example: Customer churn prediction.
- Telecom companies use Gradient Boosted Trees to predict which customers are likely to terminate their services.
- By analyzing data like call patterns, customer complaints, and billing information, the model identifies high-risk customers, helping companies take proactive retention measures.

Modern Use and Software Development

- Trees are staples in machine learning toolkits.
- Libraries: scikit-learn (Python), rpart and randomForest (R).
- Popular for interpretability and visualization.

Decision Trees

- Flowchart-like structure used for making decisions.
- Decisions made by asking a series of questions.
- Comprises nodes (questions), branches (answers), and leaves (decisions/outcomes).

Random Forests

- An ensemble of decision trees.
- Uses bootstrapped samples to build individual trees.
- Aggregates predictions: majority vote or averaging.

Leaves in Trees

- Leaves are the terminal nodes of trees.
- In causal forests, leaves estimate treatment effects.
- Allows capturing heterogeneous effects across data segments.

Regularization

- A technique to prevent overfitting.
- In trees: limit depth, minimum samples in a leaf, randomness in feature selection.
- Ensures the model generalizes well to new data.

Benefits of Ensemble Methods

- Strength in numbers: multiple trees reduce variance.
- Can capture complex, non-linear relationships.
- Improved accuracy and robustness.

Decision Trees in Causal Inference

- Which brings us to now – causal inference
- Transition to causal inference frameworks in the 2000s.
- Scholars like Susan Athey, Stefan Wager and Guido Imbens developed Causal Trees and Forests.