15.095: Machine Learning under a Modern Optimization Lens

Lecture 14: Optimal Prescription Trees

Motivation

- The Fundamental problem in Operations Research.
- Why it is important?
- Some of my core scientific beliefs and how they relate.
- Prescriptive Trees.

Outline

1. Some of my core scientific beliefs
2. The Fundamental Problem in OR
3. From Predictions to Prescriptions
4. Optimal Prescriptive Trees
5. Performance on Synthetic Data
6. Personalized Diabetes Management

Outline

Some of my core scientific beliefs

- Models exist in our imagination. All models are wrong some are useful.

Some of my core scientific beliefs

- Models exist in our imagination. All models are wrong some are useful.
- Simplicity and Interpretability are material properties of models especially in practice.

Some of my core scientific beliefs

- Models exist in our imagination. All models are wrong some are useful.
- Simplicity and Interpretability are material properties of models especially in practice.
- The only objective reality is data not probability distributions.

Some of my core scientific beliefs

- Models exist in our imagination. All models are wrong some are useful.
- Simplicity and Interpretability are material properties of models especially in practice.
- The only objective reality is data not probability distributions.
- The fundamental problem in our field is to make decisions from data over time.

Some of my core scientific beliefs

- Models exist in our imagination. All models are wrong some are useful.
- Simplicity and Interpretability are material properties of models especially in practice.
- The only objective reality is data not probability distributions.
- The fundamental problem in our field is to make decisions from data over time.
- George Dantzig:The final test of a theory is its capacity to solve the problems which originated it.

Outline

The Fundamental Problem in OR

- x: side data
- y: uncertain quantities
- z: controls
- $c(y, z)$: cost function.
- Given data (x^i, y^i), for $i = 1, \ldots, N$ and a cost function $c(y, z)$, solve

$$\min \mathbb{E}[c(Y, z) | x = x_0]$$

- Only data is known, not distributions.

Example: The News Vendor Problem

- Given x^i side data (weather in day i, S&P 500 in day $i-1$)
- y^i: demand for newspapers in day i.
- z: how many newspapers to order.
- $c(y,z) = p \times \min(y,z) - q \times z$, number of newspapers sold times price p minus cost of ordering.

Outline

A proposal

- B.+ Kallus, "From Predictions to Prescriptions", 2017
- Given data (x^i, y^i), for $i = 1, \ldots, N$, widely used machine learning (ML) methods estimate $\mathbb{E}[Y|X = x]$ for a new observation with $X = x$
- These predictions often take the form

$$\sum_{i=1}^{N} w_{N,i}(x) y^i$$

- k nearest neighbors

$$w_{N,i}(x) = \begin{cases} \frac{1}{k} & x^i \text{ is one of the } k \text{ nearest neighbors of } x \\ 0 & \text{o.w.} \end{cases}$$

- CART

$$w_{N,i}(x) = \begin{cases} \frac{1}{|R(x)|} & x^i \in R(x) \\ 0 & \text{o.w.} \end{cases}$$

where $R(x)$ is the set of training examples in the same partition of the feature space as x

- Random forest

$$w_{N,i}(x) = \frac{1}{N_{\text{tree}}} \sum_{t=1}^{N_{\text{tree}}} \frac{1}{|R^t(x)|} \mathbb{1}\{x^i \in R^t(x)\}$$

where $R^t(x)$ is the set of training examples in the same partition as x in tree t of the random forest

Prescriptions

- Full information problem

$$\min_{z \in Z} \mathbb{E}[c(z; Y) | X = x]$$

- Given data (x^i, y^i) for $i = 1, \ldots, N$, the approximate problem is given by

$$\min_{z \in Z} \sum_{i=1}^{N} w_{N,i}(x) c(z; y^i)$$

  where $w_{N,i}(x)$ is a weight function from an ML method

Performance

- Under certain regularity conditions, for certain weight functions, solution is asymptotically optimal and cost estimate is strongly consistent
- No additional computational cost compared to SAA problem (as long as weight functions are nonnegative).
- Should we do blood transfusion to certain patients before surgery to minimize probability of re-admission within 30 days? Reduction of re-admission rate by 8%
- Significant improvements in revenue in several real world problems.

Outline

Optimal Prescriptive Trees

- B+Dunn+Mundru, Optimal Prescriptive Trees, 2018.
- Consider a healthcare setting (personalized medicine, many other applications)
- Historical observational data $(X_i, z_i, Y_i),\ i = 1, \ldots, n$.
- $X_i \in \mathbb{R}^d$: Features of patient i.
- $z_i \in \{1, 2, \ldots, m\}$: Treatment assigned to patient i by doctor.
- $Y_i \in \mathbb{R}$: Outcome recorded of patient i (Lower the better).
- **Question:** When a new patient comes in with features x, what treatment $\tau(x) \in \{1, 2, \ldots, m\}$ is best for this person?

Can we use Machine Learning?

- For each patient x_i: **If** we knew the **best treatment** (treatment out of m options that leads to best outcome), then it is a *standard multiclass classification problem*.
- We could learn a classifier that predicts in $\{1, \ldots, m\}$ given $x \in \mathbb{R}^d$ using this historical data.
- **KEY CHALLENGE:** But, we only know the outcome for z_i (historically given treatment) and not the others.
- We do not know *what would have happened* ("counterfactuals") to patient i under the other $(m-1)$ treatments.

Why not predict these counterfactuals directly?

- For each treatment $t \in \{1, 2, \dots, m\}$, find the subset of subjects who received that treatment.
- Regress outcome y v/s features X for these subjects – Compute the regression function f_t for each t.
- **Test data**: For a new subject with features x, assign treatment that leads to *lowest predicted outcome*.
- We denote this method as **Regress-and-Compare** (**R&C**).

Can we do better?

- **Drawbacks of R&C**:

  1. Learning from subsamples - Less data to learn each f_t (Particularly when n/m is small)
  2. Splitting the data - can miss joint trends.
  3. Decision boundary is not explicitly characterized - not interpretable.

- **What is desirable?**
  1. To use all the data at once
  2. Tractability
  3. Interpretability
  4. Comparable or better performance to state of the art methods

Optimal Prescriptive Trees

- **Objective**: Determine $\tau(x)$ to minimize

$$\mu \texttt{ Mean outcome} + (1-\mu)\texttt{Prediction error}, \quad 0 < \mu < 1$$

$$\mu \left[\sum_{i=1}^{n} \left(y_i \mathbb{I}[\tau(\mathbf{x}_i) = z_i] + \sum_{t \neq z_i} \hat{y}_i(t) \mathbb{I}[\tau(\mathbf{x}_i) = t] \right) \right] +$$

$$(1-\mu) \left[\sum_{i=1}^{n} (y_i - \hat{y}_i(z_i))^2 \right],$$

- Need to predict **counterfactuals**.
  1. For each subject i : If he/she received treatment 1, we know $Y_i = Y_i(1)$.
  2. Estimate $Y_i(0)$ as average of patients in that leaf who received 0.
  3. Can also use linear regression.
- Use OCT or ORT algorithms.

Outline

Computations with Synthetic Data

- We generate n data points $x_i, i = 1, \ldots, n$ where each $x_i \in \mathbb{R}^d$ with $n = 1000, d = 20$ for the training set.
- Two treatments: 0 and 1.
- The outcome Y_t under each treatment t as a function of x is given by

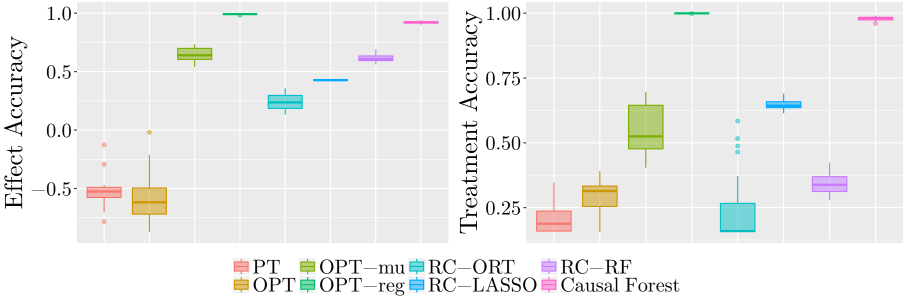
$$Y_0(x) = \texttt{baseline}(x) - \frac{1}{2}\texttt{effect}(x)$$
$$Y_1(x) = \texttt{baseline}(x) + \frac{1}{2}\texttt{effect}(x)$$

- To simulate an observational study, we assign treatments probabilistically depending on x.
- Metrics: (Reported on a test set)
  - **Treatment accuracy**: Fraction of units in the test set for which prescriptions match ground truth.
  - **Effect accuracy**: R^2 of predicted individualized treatment effect versus true value.

Methods we compare

- **Prescription Trees:** We include four prescriptive tree approaches:
  - Personalization trees, denoted PT;
  - Optimal Prescriptive Trees (OPT) with $\mu = 1$, i.e. only optimizing for personalization risk, denoted OPT;
  - OPT with $\mu = 0.5$ (jointly minimizing personalization risk and prediction error), denoted OPT-mu;
  - OPT with $\mu = 0.5$ and linear counterfactual estimation in each leaf, denoted OPT–reg.
- **Regress-and-compare:** We include three R&C approaches where the underlying regression uses either Optimal Regression Trees (ORT), LASSO regression or random forests, denoted RC–ORT, RC–LASSO and RC–RF, respectively.
- **Causal Forests:** While causal forests are intended to estimate the individual treatment effect, we use the sign of the effect to determine the choice of treatment.

Low noise, linear baseline and piecewise constant effect functions



Moderate noise, constant baseline and piecewise linear effect functions

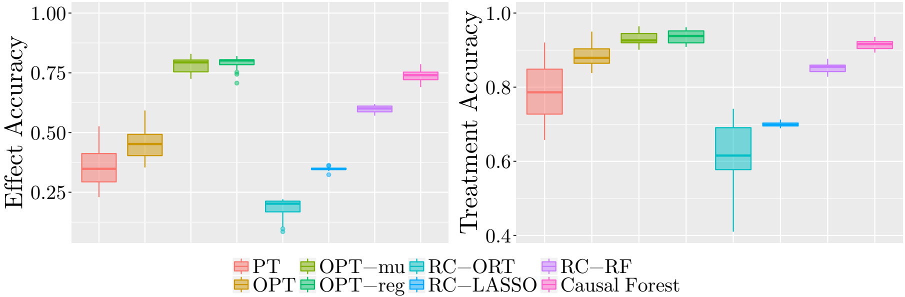
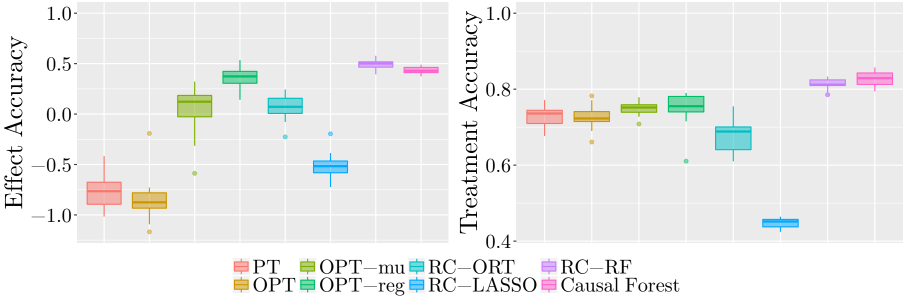


Figure: Comparisons for Combination 2

High noise, piecewise constant baseline and quadratic effect functions



Discussion

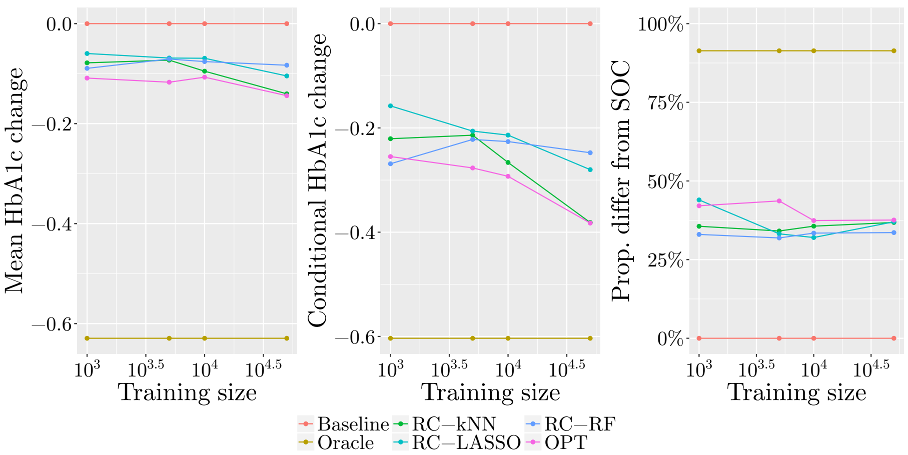
- Explicit representation of decision boundary leading to interpretability.
- R&C methods that fit separate functions for each treatment are generally outperformed by joint learning methods that learn from the entire dataset.
- Causal Forests and OPT are the strongest in terms of performance.

Outline

Personalized Diabetes Management

- Data from the Boston Medical Center, from 1999-2014.
- 100,000 patient visits for type 2 diabetes.
- 13 possible treatment options (regimens).
- Patient features include demographic information (sex, race, gender etc.), treatment history, and diabetes progression.
- Outcome of interest: $\mathrm{Hb}_{\mathrm{A1c}}$ level; smaller the better.
- Varied # training samples from 1,000–50,000 to examine the effect on out-of-sample performance. Averaged this process over ten different splits of the data.

OPT has a Performance and Interpretability Edge



Conclusions

- Optimization from data.

Conclusions

- Optimization from data.
- Optimal Prescriptive Trees have strong performance and are interpretable.

Conclusions

- Optimization from data.
- Optimal Prescriptive Trees have strong performance and are interpretable.
- The proposals today, we call them **Prescriptive Analytics** use data as primitives and combine ML and optimization, to make decisions over time.