IST 402 Week 13 Scribing Notes

Patrick Ryan
Ben Morgan
Jack DiPietro
Jack Woodburn

Week 13 Applications of Machine learning continued

Allocating interventions based on Predicted Outcomes (Homelessness Services Case Study)

Homeless people are increasing in number
-Homeless System à Federally funded system which assigns homeless people to 5 different kinds of housing
-4 housing programs
- Emergency shelter
- Transitional Housing
- Rapid Rehousing
- Homelessness Prevention
- Permanent Supportive Housing

Can we improve the assignment process using AI

Fairness needs to be kept in mind
Interpretability of ML models
- Your take?
- How important is interpretability
- Cost vs benefit

Causal Inference
- Need counterfactual estimates of different interventions before you can let the ML model decide which intervention/service you want to make an assignment to

Homeless Person assigned to Service A à Does not re-enter the system
Homeless person Assigned to Service B à Re-enters the system
But are there any confounding variables that can explain this change.

So how much can we potentially improve outcomes?

- What is each kind of data
  - Household Characteristics (Feature)
  - Which of the 5 services was assigned to them (Feature)
  - Predictor -> Whether they re-enter the Homeless system within 2 years of exit (Label) (Is this a good metric?) (Looking at 2 years)
  -

Their Approach:

- Clean this data to create a dataset
- Train an ML model on this that supports causal inference
    - BART (Bayesian Additive Regression Trees)
    - Provides counterfactual estimates for re-entry under different services
- See if the ML model is a good approximation of the real datasets
- (missed)

How did their Model do?

- Ground Truth (from their dataset) -> 43.04% households re-enter
- Prediction from BART model -> 43.72% households re-enter
- Fairly accurate - proved in the paper

How to Optimize

- Equation solved

After Optimization of Assignments -. 31.88% households re-enter
Reduction by 27.88%

Is the Optimized Solution Pareto Improving?: No

- What is Pareto-Improving?
    - Pareto Frontier??

- 33.17% assigned to services which drop their probability to re-enter
- 34.21% assigned to the same service
- 32.62% probability of re-entry increases

Thursday, Nov 21 Notes
Covering 2 Papers

1. Decision Trees
2. Bagged Decision Trees - Ensemble method
3. Boosted Decision Trees - Ensemble method

Looking at the standard decision tree

- Can be read as an if statement
- Each leaf node is associated with a decision
- Start from the top and follow the decisions to a leaf and with that leaf is your answer
- Oblique Splits: Decision boundaries which are not aligned with your axes

- Internal nodes test attributes
- Is determined by attribute value
- Branching is an output value

Decision tree algorithm

- Choose an attribute on which to descend at each level
- Condition on earlier (higher) choices
- Generally, declare an output value when you get to the bottom

- In the orange/lemon example…

Expressiveness

For Discrete Input, discrete-output case:

- Decision trees can express any function of the input attributes
- E.g., for boolean functions, truth table row -> path to a leaf

For Continuous Input, Continuous Output case

- Can approximate any function arbitrarily closely

2) Bagged Decision Tree

- Instead of training different models on the same data, train the same model multiple times on different data sets, and "combine" these "different" models
- Bagging stands for Bootstrap Aggregation
- Takes original data set D with N training examples
- …

3) Boosting

- Take a weak learning algorithm
- Only requirement: Should be slightly better than random
- Turn it into an awesome one by making it focus on difficult cases

Going over 2 Articles

- Water Mains break a lot, cutting water supply
- A problem in older cities, Syracuse is a prime example
- Looking to predict which water mains will break

Current Strategies

- A reactive system, fix the mains after they break
- Meaning residents won't have water
- Want an AI solution

Plan

- Frame it as a binary classification problem of whether a water main break will occur in a given city block within the next 3 years
- Result: Precision of 62% in the top 1% of our predictions
- Two ways in which system can be used
- 1) for preventative maintenance on the top 1% of the riskiest breaks
- 2) To use the risk scores to…

Paper 2

Animals are at a constant threat of being poached using traps and wires illegally

- How do you predict future poacher activity from past data

- "Missing" poaching data
- Limited patrol resources
- Imperfect observations

- Consequences

- Uncertainty in negative labels
- Class imbalance

Solving

- Booster Decision tree
- Built-up Ensemble
- A mixture of DT creates the most accurate data, used as an ensemble
-

Results

- Infrequent Hot Spots
- Predicted Hotspots
- Trespassing
- Poached Animals
- Snaring

Other notes on Decision Tree's

A classification tree has a descrete output

Regression tree has a continuous output

Algorithm

- An attribute at each level can be labeled true or false
- Output/solution at the bottom
- Boolean functions can be used to create a truth table

Creating a Decision Tree

- Pick the best attribute at each level and split the tree
- Repeat until a final solution is identified

Keep in mind…

- A good decision tree is…
  - not small enough to the point where it is inaccurate
  - Not too large so the solution is efficiently found
  - Regularization is necessary in order to ensure decision trees are compact