Some interesting real-world problems viewed through the data science lens

Nandan Sudarsanam – IIT-Madras

Introduction

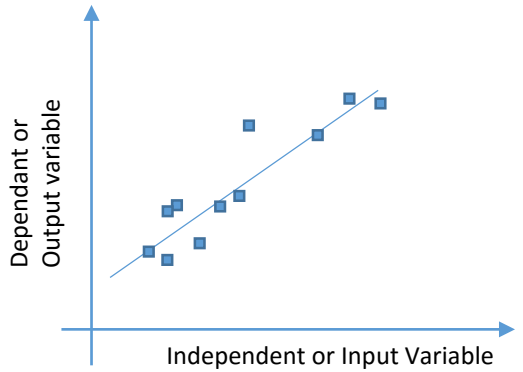
- Is this one talk or four talks mashed together?
- What techniques?
  - Machine Learning, Statistics and Operations Research: All of which have large applied math components
- Four specific examples and their applications
  - PU (Positive and Unlabelled) Learning: Application to Urban Mobility
  - Graph Colouring: Application to Railways
  - Reinforcement Learning: Bandit Problems
    - Ticket checkers
    - E-commerce recommendations
    - Software Testing
  - Temporal Difference Learners: Application to Microfinance

Machine Learning - 101

- Machine Learning
  - Related to Data Mining, Pattern Recognition, statistical learning, etc.
  - Supervised Vs Unsupervised Learning
    - Supervised Learning: Task of creating a function/relationship from training data which has one or more explicit output (dependant) variables. Also, indicated as data that is labelled. This can then be used to map new instances of the inputs.
      - Regression (the word means something different here) – The output is continuous
      - Classification – The output is categorical
    - Unsupervised Learning: Task of creating patterns from data which have no explicit measure or signal guiding us. The data is unlabelled
      - Clustering
      - Association Rule Mining
  - Semi-supervised Learning, Reinforcement Learning

Supervised Learning and Regression Analysis

- Regression Analysis: The first supervised learner
  - Simple/Multiple Linear regression: Fitting a line to data
    - Relationship between Dependent (Output) variables and Independent (Input) variables
    - Multiple Regression: More variables, or transformations/high order extensions of the same input
    - Examples: Sales: Population density and number of customers



The four examples

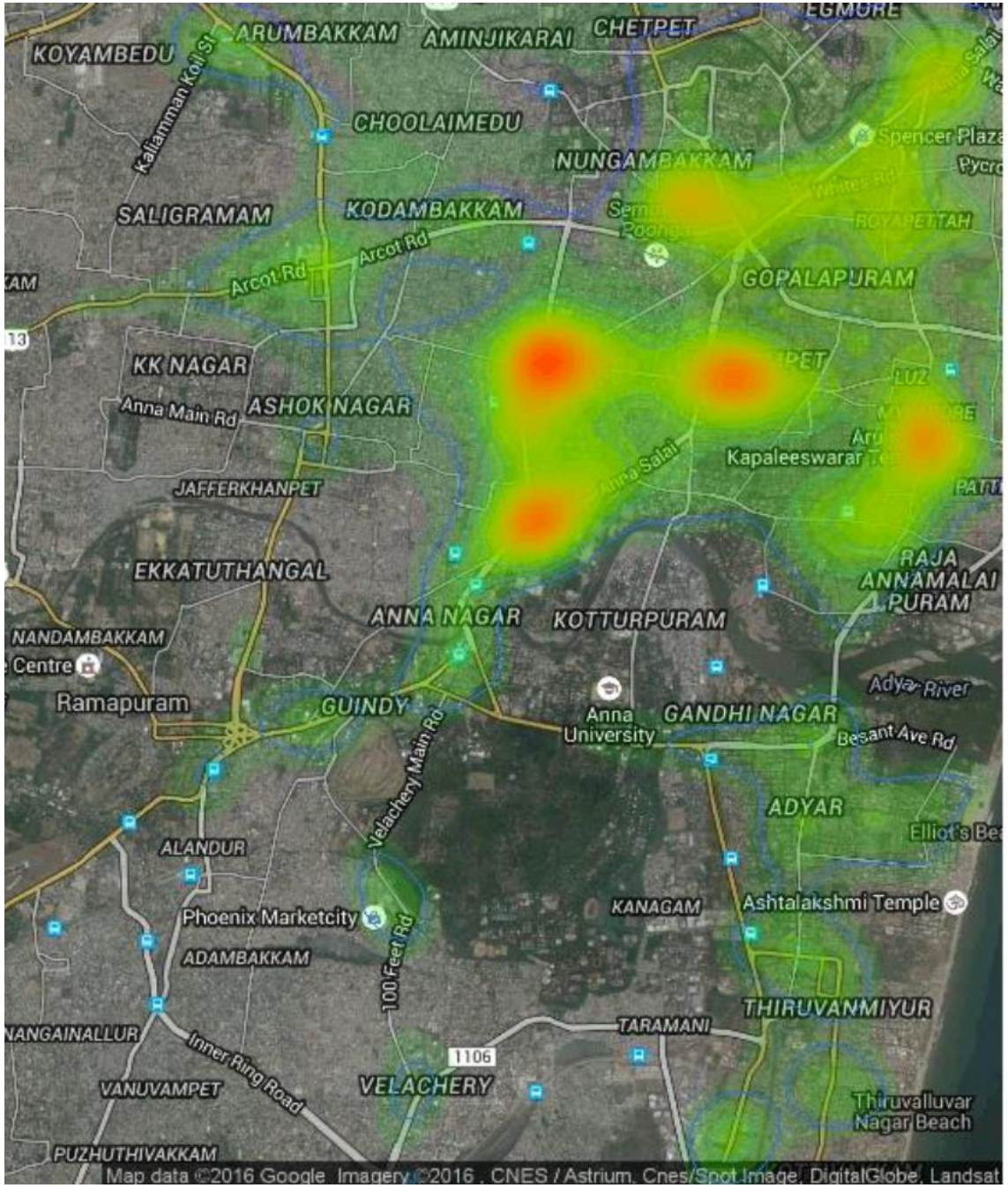
- PU (Positive and Unlabelled) Learning: Application to Urban Mobility
- Graph Colouring: Application to Apps and Railways
- Reinforcement Learning: Bandit Problems
  - Ticket checkers
  - E-commerce recommendations
  - Software Testing
- Temporal Difference Learners: Application to Microfinance

The first problem: Urban mobility and the For-hire vehicle market

- For-hire vehicles are a major mode of transportation in the urban environment.

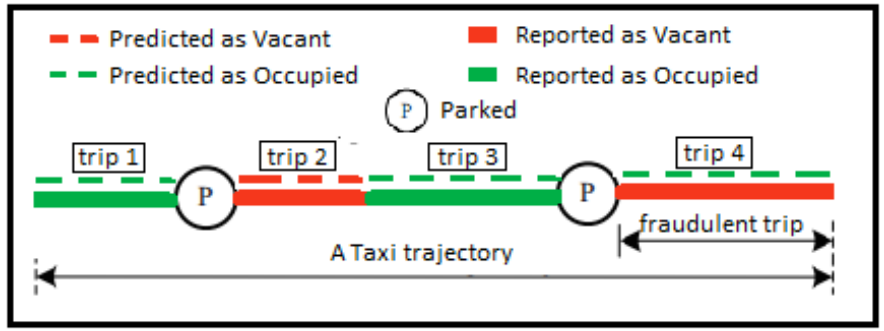
- These vehicles fitted with GPSs and auto meters can provide us with data, which in turn can provide interesting insights and answer questions:
  - Patterns in urban mobility and migration, with implications for policy
  - From the organizations perspective:
    - What route should a vehicle take?
    - What should a for-hire vehicle do when it has dropped off a passenger?

The first problem: Urban mobility and the For-hire vehicle business



The first problem: Urban mobility and the For-hire vehicle market

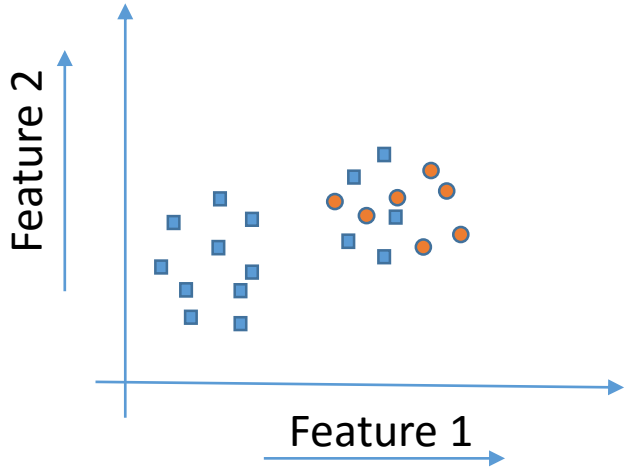
- Is the data reliable?
- What does cleaning the data buy us?
  - The data becomes usable for data analytics
  - It can be used to catch fraudulent drivers (inductive and transductive)
- Can we use classification to clean the data/catch fraudulent drivers?



- But the training data also has fraudulent trips!

The first problem: Urban mobility and the For-hire vehicle market

- Is this a case of just some noise (or mislabelled classes)?
- The asymmetric nature of the mislabelled classes
- Why is this a PU(positive and unlabelled) learning problem?
- The approach to tackling PU learning problems
  - Using unsupervised with supervised



The four examples

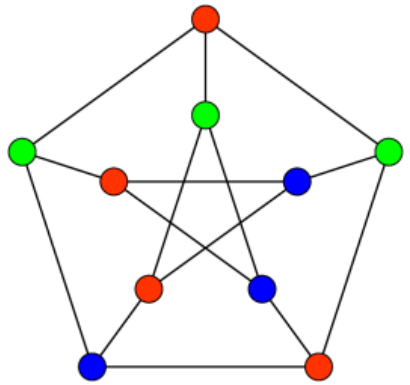
- PU (Positive and Unlabelled) Learning: Application to Urban Mobility
- Graph Colouring: Application to Autism Apps and Railways
- Reinforcement Learning: Bandit Problems
  - Ticket checkers
  - E-commerce recommendations
  - Software Testing
- Temporal Difference Learners: Application to Microfinance

The second problem: Allocation through graph colouring

- App for autistic children which helps them communicate
- Data presentation problem
- Dataset – corpus of sentences through crowdsourcing
- Assignment of words to slots
- Objective: Minimize the number of instances where a word needs to be moved to a new location due to a conflict in the slot

The second problem: Allocation through graph colouring

- Applications to other problems.
  - What if we wanted to assign trains to platforms. (flights to gates)
  - Dataset: Actual train arrival and departure information
- What is graph colouring and how can it help us:
  - Label (colour) the vertices of a graph with minimum number of labels such that no two vertices share the same label
    - In the app: Nodes are words, colours are slots
    - In the train problem: nodes are trains, colours are platforms
  - But this conception still does not help us solve our problems. Why?
- The need for the MCE (Minimizing Conflicting Edges) variant



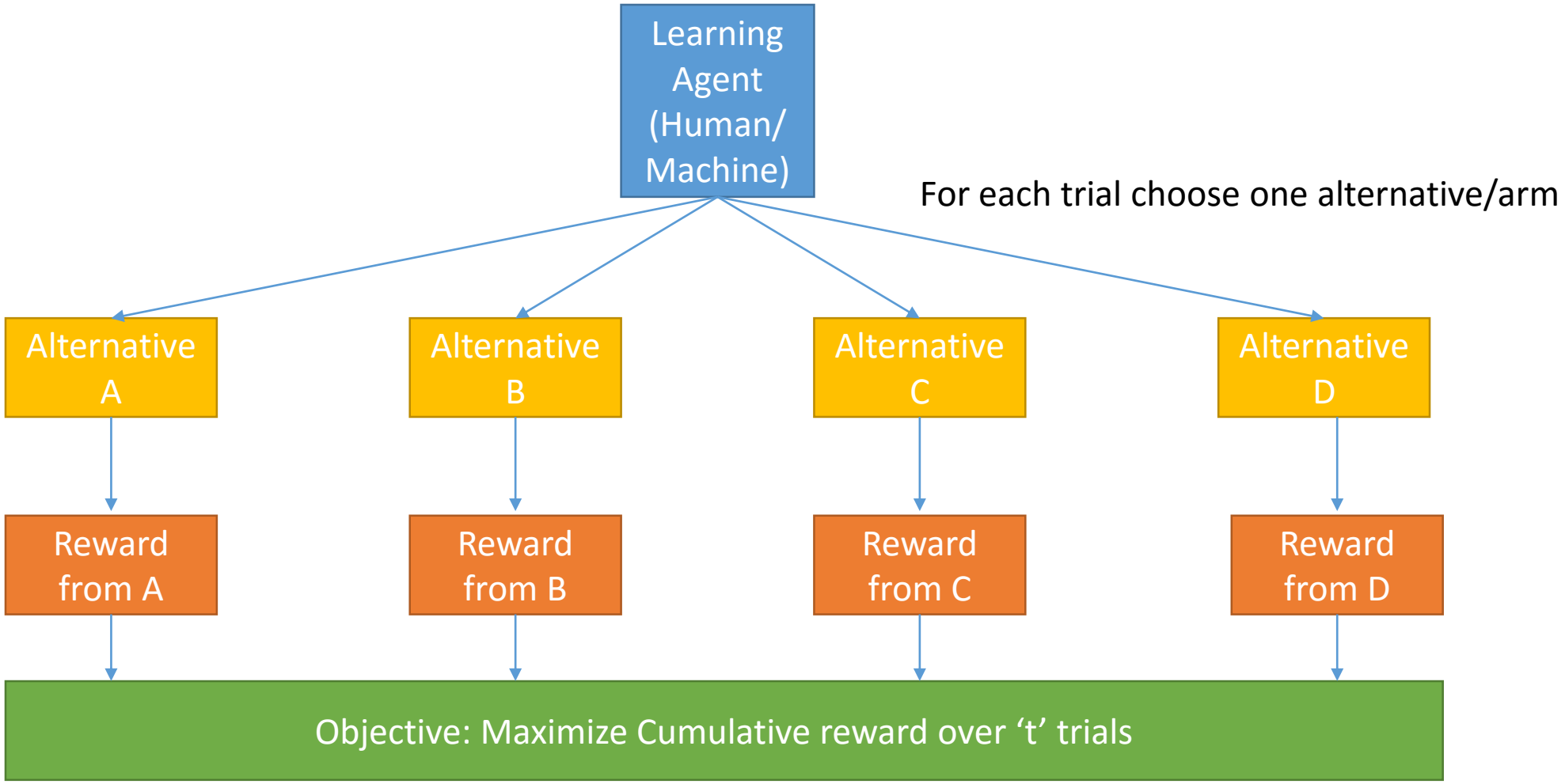
The four examples

- PU (Positive and Unlabelled) Learning: Application to Urban Mobility
- Graph Colouring: Application to Autism Apps and Railways
- Reinforcement Learning: Bandit Problems
  - Ticket checkers
  - E-commerce recommendations
  - Software Testing
- Temporal Difference Learners: Application to Microfinance

The third problem: Bandit Problems

- Data Science and analytics need data (not to mention Big-Data)
- What if you don’t have data? Creating Data and then analysing it
- Online vs Offline context of creating data
  - Offline: Design of Experiments (DOE), A/B testing
  - Online: Reinforcement Learning: Bandit Problems
- Some Bandit examples: online Ads/News articles, clinical trials, financial strategies
- The exploration-exploitation dilemma

The third problem: Bandit Problems



The third problem: Bandit Problems

- Why not just choose the best?
- What if you have too many options? Linear Bandits
- Some novel applications:
  - Ticket Checkers: The context matters
  - E-Commerce Recommendations: need to go beyond linear and multi-armed bandits
  - Software testing: The arms are related to each other

The four examples

- PU (Positive and Unlabelled) Learning: Application to Urban Mobility
- Graph Colouring: Application to Autism Apps and Railways
- Reinforcement Learning: Bandit Problems
  - Ticket checkers
  - E-commerce recommendations
  - Software Testing
- Temporal Difference Learners: Application to Microfinance

The fourth problem: TD Learners in MFIs

- Understanding and Predicting the customers' transactional behaviour is critical for financial institutions that provide debt
- The standard framework for making such predictions, typically, involves the use of demographic information along with historic repayment patterns.
- Unlike traditional retail lending, microfinance institutes (MFIs) operate in a different environments. They may look at previously unexplored :
  - Customer demographics
  - Geographical locations
  - Completely new products

The fourth problem: TD Learners in MFIs

- These make the relevant historic data both sparse and different from the traditional institutions
- This difference in data requires predictive modelling approaches that are uniquely suited to operate in the microfinance context
- Specifically, we focus on the idea that there would be ***limited cases where we have data over the complete life cycle of the debt. Instead a significant portion of the data could be from customers who are mid-cycle***. We focus our efforts on utilizing this partial repayment information to aid the cold-start problem of a data driven decision-making initiative.

The fourth problem: TD Learners in MFIs

- Using Temporal Difference Learners to make predictions
  - No need to wait till the end of the sequence of repayments for a debt to utilize the data
  - Useful when a major portion of the clients are at some point mid cycle in repaying the debt
  - Agnostic to the method used to predict prepayments.

Visual Representation

- Visual Representation of data Utilization of a traditional Supervised learner versus Temporal Difference learner:
- The diagram shows a four month repayment cycle for a loan
- Earlier months have more data (represented by more grids) and later months have little or no data
- The shaded portions indicate that the data is being utilized as input or output in the exercise to predict month four from month one