Taxi in New York City: Generating Lucrative Passenger Pick-up Strategy

Julian Gao, Qiujiang Jin
Graduate Department of Computer Science
Stanford University

Stanford | ENGINEERING
Computer Science

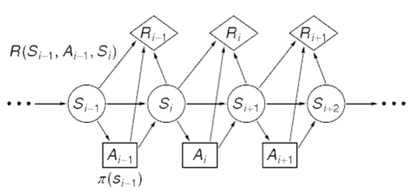
Project Description

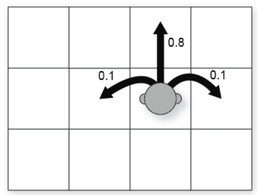
Data:
- 500 million records of NYC taxi trips from Jan 2015 to Jun 2016
- Gathered from NYC taxi & limousine commission official source
- Each record contains important details about the trip

Goal:
- To explore the most lucrative passenger pick-up location
- Produce useful data analytics as byproduct

Areas of Impact:
- Artificial Intelligence — state search, Markov Decision Process
- Machine Learning — pattern finding, modeling
- Data Mining — big data, statistical analysis

Markov Decision Process

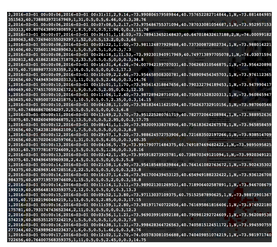


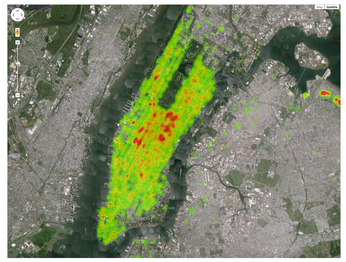


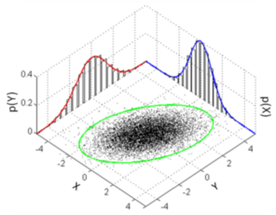
Notations
- t_{curr}, t_{ava}, t_{start}, t_{end} : current time, available time, start time, end time
- l_{curr}, l_{start}: current location (longitude, latitude grids), start location
- δ_t: Time elapse moving between locations
- f: Positive reward evaluation, sampled from payment
- c: Negative reward evaluation, sampled from time and distance

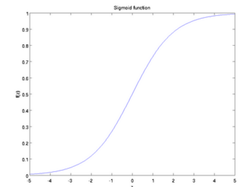
- State s := (t_{curr}, l_{curr}, t_{ava})
- S_{start} := (t_{start}, l_{start}, t_{end} - t_{start})
- Succ(s, a) := (t_{curr} + $\delta_t(a(s))$, $a(l_{curr})$, t_{ava} - $\delta_t(a(s))$, $p_{pickup}(l_{curr})$)
- Transition(s, a, s') := $1\{s' \in a(s)\}(1 - p_{pickup}(s)) + 1\{s' \notin a(s)\}p_{pickup}(s)p_{dest}(a(s), s')$
- Reward(s, a, s') := f(s, $a(s)$, s') - c(s, $a(s)$, s')
- IsEnd := $1\{t_{ava} = 0\}$, $\gamma = 1$

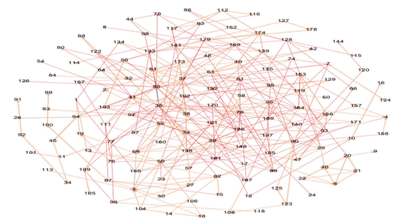
Strategy Generation Process







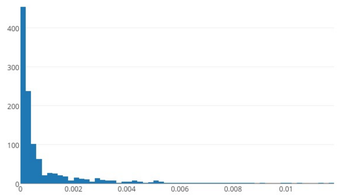




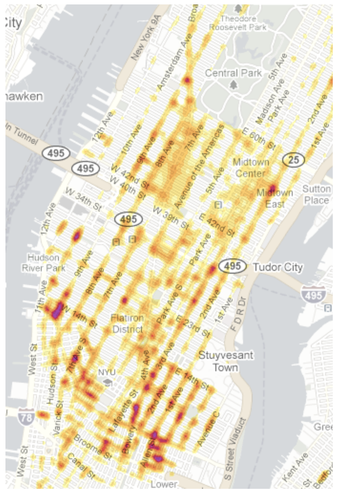
Input Data	Congestion Factor	Reward Function	Probability Map	Policy Generation
Data Pre-Processing	**Travel Time Estimation**	**Modeling Distribution**	**Pick-up Probability**	**MDP Simulation**
• Remove bad entries • Extract days • Discretize time and location intervals	• Sigmoid model of speed calibration • Congestion factor linear in number of activities	• Gather payment reward information • Gather time, distance cost information	• Calculate average waiting time • Estimate cruise speed • Sampling from Poisson process	• Defined by transition probability and reward • Maximize utility (money)

Data Exploration

Histogram of Hotspots
12am, Tuesday



Pick-up Probability Map



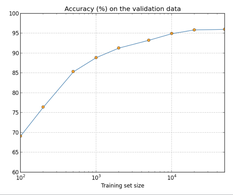
Theoretical Foundations

Travel Time Estimation
- Assume taxi driver drives through blocks without intentional stops
- Assume cruise speed linear with congestion factor and city average
- Congestion factor calculated by sigmoid

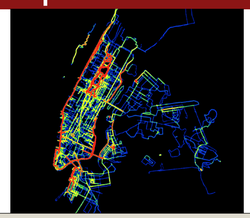
Pick-up Probability
- Average waiting time is representative when sample dataset is large
- Passenger arrival is a standard Poisson process
- Assume waiting time is equivalent to travel time

Reward Function
- Assume fixing starting position, the distributions of destination positions, trip time, trip distance, and payment amount are Gaussian models
- Assume taxi drivers maximize reward / cost.

Next Steps



Run on larger sets
More data means more representative parameters, which suggests more accurate prediction results



Visualize Path
Generate visualized optimal strategy path on city map real-time

References

NYC Taxi & Limousine Commission, http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
Harvard NYC Taxi Data Prediction, S. Daulton, S. Raman, T. Kindt, *http://sdaulton.github.io/TaxiPrediction/*
Tipsy Taxi project, http://lucianoiscool.github.io/Tipsy-taxi/
Nike's GPS NYC Visualization, Cooper Smith