Business Analytics

Session 9b. Optimization Basics and Linear Programming

Renyu (Philip) Zhang

New York University Shanghai

Spring 2019

What is Optimization

- We now move into the 3rd module of this course: Optimization.
- Most of the decision making problems in practice can be phrased into the following prototype:

> Maximze the reward (or minimize the cost) under some constraints.

- Examples:
  - Google maximizes advertising revenue without sacrificing user experiences.
  - VCs maximize return subject to risk tolerance.
  - Prediction models minimize loss on the training set without having an overly complex model.

Optimization Model

$$\max f(x), \text{ subject to } x \in \mathcal{X}$$

- x: Decision variables, which define a strategy or an action plan.
  - Control parameters to be decided.
  - Examples: price, advertisement alocation, investment portfolio, etc.
- $f(\cdot)$: Objective function, which defines the goal of the problem, the target to be optimized.
  - Expressed as a function of the decision variables.
  - Provides a criterion to compare alternate solutions/decisions.
  - Can be controlled by decision variables, but not directly.
  - Examples: profit, revenue, cost, risk, etc.
- $\mathcal{X}$: Constraints, defines when the decision variables are actually feasible.
  - Expressed in terms of decision variables.
  - Examples: Technical, financial, legal, and logical constraints, etc.
- Optimization triplet: decision variables, objective function, constraints.

Solution to Optimization Model

$$x^* = argmax\ f(x), \text{ subject to } x \in \mathcal{X}$$

- For a general optimization, i.e., $f(\cdot)$ is a general objective function, $\mathcal{X}$ is a general feasible set, it is impossible to obtain x^*
- When $f(\cdot)$ and $\mathcal{X}$ satisfy some conditions, solving x^* is (relatively) easy or at least plausible.
  - $f(\cdot)$ is linear or concave.
  - $\mathcal{X}$ is a convex set or a convex polytope.
- Two goals:
  - Translate decision problems into optimization models.
  - Solve optimization models using analytics tools.

Matrix

- Data are stored in matrices: $\mathbf{A_{n\times m}} = (\mathbf{A_{i,j}})_{\mathbf{n\times m}}$
  - n data points, m covariates.

- Example:

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{pmatrix} = \begin{pmatrix} 2 & 0 & -1 \\ 3 & 1 & 2 \end{pmatrix}$$

- Two matrices $\mathbf{A} = \mathbf{B}$ if each entry of $\mathbf{A}$ is the same as the corresponding entry of $\mathbf{B}$, i.e., $A_{ij} = B_{ij}$ for all $1 \le i \le n, 1 \le j \le m$.

- Transpose: $\mathbf{B} = \mathbf{A}^T$, then $B_{ij} = A_{ji}$.
  - If $\mathbf{A}$ is n by m, then $\mathbf{A^T}$ is m by n.

Summation of Matrix

- Two matrices of the same dimensions $(n \times m)$, $\mathbf{A}$ and $\mathbf{B}$

$$\mathbf{A}+\mathbf{B}=\begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} & \cdots & A_{1m}+B_{1m} \\ A_{21}+B_{21} & A_{22}+B_{22} & \cdots & A_{2m}+B_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ A_{n1}+B_{n1} & A_{n2}+B_{n2} & \cdots & A_{nm}+B_{nm} \end{pmatrix}$$

- $\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A}$, $(\mathbf{A}+\mathbf{B})+\mathbf{C}=\mathbf{A}+(\mathbf{B}+\mathbf{C})$, $\mathbf{A}+\mathbf{O}=\mathbf{A}$, where all the entries of $\mathbf{O}$ is 0.

- Matrix substraction: $\mathbf{A}-\mathbf{B}=\mathbf{A}+(-\mathbf{B})$.

Multiplication of Matrix

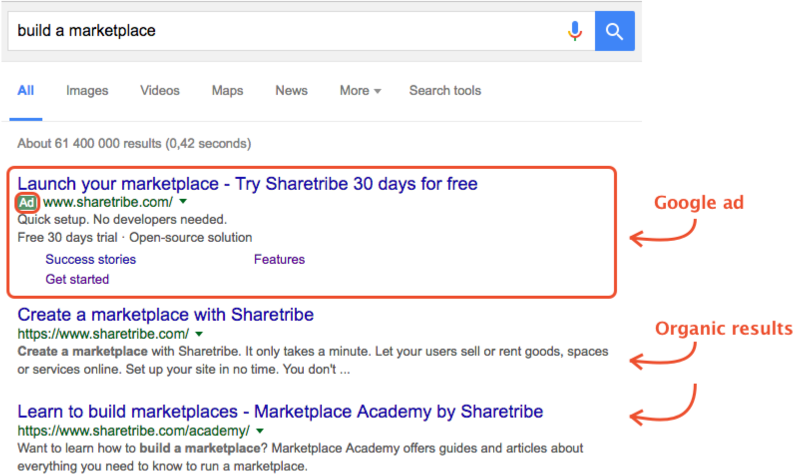
- Two matrics $\mathbf{A}_{p\times q}$ and $\mathbf{B}_{q\times r}$: $\mathbf{M}_{p\times r} = \mathbf{AB}$
  - $M_{ij} = \sum_{k=1}^{q} A_{ik}B_{kj}$
- $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$, as long as the dimensions match.
- In general, $\mathbf{AB} \neq \mathbf{BA}$.
  - The dimensions may not even match.
- How do we represent the following in matrices:
  - $f(x) = \sum_{i=1}^{n} c_i x_i$
  - $\sum_{j=1}^{n} A_{ij}x_j \leq B_i$, for $i = 1, 2, \cdots, m_1$
  - $\sum_{j=1}^{n} a_{ij}x_j = b_i$, for $i = 1, 2, \cdots, m_2$

Google AdWords

Business Model of Google

- Although Google has more than 1 billion unique monthly visitors, its search engine is free to use.
- How does Google make money?
  - Google generates more than 110 billion USD in 2017.
- Google AdWords: The sponsored online ads system.
  - 97% of Google's revenues come from AdWords.

Sponsored Ads on Google



Google AdWords

- Why do companies advertise on Google?
  - Google receives very heavy traffic.
  - Search pages are formatted in a very clean manner.
  - Companies can choose the queries their ads will be displayed so as to target the desired audience.

- Advertising on Google:
  1. Advertisers place bids for different queries in an auction (generalized second-price auction).
  2. Based on bids and quality score (fit of advertiser and ad to the queries), Google decides price-per-click of each advertiser and each query.
  3. Google then decides how often to display each ad for each query.

Price-Per-Click (PPC) and Budget

- Price-Per-Click: How much the advertiser pays Google when a user clicks its ad for that query.

Advertiser	Query 1 "4G LTE"	Query 2 "largest LTE"	Query 3 "best LTE network"
AT&T	$5	$5	$20
T-Mobile	$10	$5	$20
Verizon	$5	$20	$25

- Budget: The maximum total payment of the advertiser to Google.
  - Each time user clicks on the ad, the budget is depleted by PPC amount.

Advertiser	Daily Budget
AT&T	$170
T-Mobile	$100
Verizon	$160

Click-Through Rate (CTR)

- Click-through rate (CTR): The probability a user clicks on an ad.
  - CTR can be viewed as "click per user".

Advertiser	Query 1 "4G LTE"	Query 2 "largest LTE"	Query 3 "best LTE network"
AT&T	0.10	0.10	0.08
T-Mobile	0.10	0.15	0.10
Verizon	0.10	0.20	0.20

  - If we have 100 users who search "largest LTE" (all allocated to T-Mobile), how many clicks will T-Mobile's ad receive on average?
- PPC×CTR: Average money an advertiser pays Google for each display of its ad for the query.
  - If we allocate 100 "best LTE network" queries to AT&T, what is the expected payment from AT&T to Google.

Advertiser	Query 1 "4G LTE"	Query 2 "largest LTE"	Query 3 "best LTE network"
AT&T	$\$5 \times 0.10 = \$0.5$	$\$5 \times 0.10 = \$0.5$	$\$20 \times 0.08 = \$1.6$
T-Mobile	$\$10 \times 0.10 = \$1$	$\$5 \times 0.15 = \$0.75$	$\$20 \times 0.1 = \$2$
Verizon	$\$5 \times 0.10 = \$0.5$	$\$20 \times 0.20 = \$4$	$\$25 \times 0.2 = \$5$

Query Estimates

- Google cannot control how many times a query will be requested – Driven by users!
- Instead, Google estimates the number of times a query will be requested over a given day.

Query	Estimated # of Requests
"4G LTE"	140
"largest LTE"	80
"best LTE network"	80

Google's Problem

- Key Question: How many times to display each ad for each query to maximize revenue?
- Decisions: For each ad and each query, the number of times the ad will be displayed for that query.
- Constraints:
  - Budget: Average amount paid by each advertiser cannot exceed the budget
  - Query estimates: Total ads for a given query cannot exceed the number of requests for that query

First Approach: Greedy Strategy

Greedy Strategy: Display the most profitable ad.

- Most profitable (feasible) allocation: Q_3 to Verizon (\$5).

	Q_1	Q_2	Q_3	Budget
AT&T				170
T-Mobile				100
Verizon			32	160-32$\times$5=0
Numbers	140	80	80-32=48	

- Most profitable (feasible) allocation: Q_3 to T-Mobile (\$2).

	Q_1	Q_2	Q_3	Budget
AT&T				170
T-Mobile			48	100-48$\times$2=4
Verizon			32	0
Numbers	140	80	48-48$\times$1=0	

- Most profitable (feasible) allocation: Q_1 to T-Mobile (\$1).

	Q_1	Q_2	Q_3	Budget
AT&T				170
T-Mobile	4		48	4-4$\times$1=0
Verizon			32	0
Numbers	140-4=136	80	0	

First Approach: Greedy Strategy

- Most profitable (feasible) allocation: Q_1 to AT&T (\$0.5).

	Q_1	Q_2	Q_3	Budget
AT&T	136			170-136$\times 0.5$=102
T-Mobile	4		48	0
Verizon			32	0
Numbers	136-136=0	80	0	

- Most profitable (feasible) allocation: Q_2 to AT&T (\$0.5).

	Q_1	Q_2	Q_3	Budget
AT&T	136	80		102-80$\times$0.5=62
T-Mobile	4		48	0
Verizon			32	0
Numbers	0	80-80=0	0	

- Revenue of the greedy strategy:

$$170 + 100 + 160 - 62 = \$368$$

- Question: Can we do a better job?

Modeling the Problem

- **Decisions**:

Advertiser	Query 1 "4G LTE"	Query 2 "largest LTE"	Query 3 "best LTE network"
AT&T	X_{A1}	X_{A2}	X_{A3}
T-Mobile	X_{T1}	X_{T2}	X_{T3}
Verizon	X_{V1}	X_{V2}	X_{V3}

- **Revenue**:

$$0.5X_{A1} + 0.5X_{A2} + 1.6X_{A3} + X_{T1} + 0.75X_{T2} + 2X_{T3} + 0.5X_{V1} + 4X_{V2} + 5X_{V3}$$

- **Constraints**:
  - Budget for AT&T: $0.5X_{A1} + 0.5X_{A2} + 1.6X_{A3} \leq 170$
  - Budget for T-Mobile: $X_{T1} + 0.75X_{T2} + 2X_{T3} \leq 100$
  - Budget for Verizon: $0.5X_{V1} + 4X_{V2} + 5X_{V3} \leq 160$
  - Number of Q_1: $X_{A1} + X_{T1} + X_{V1} \leq 140$
  - Number of Q_2: $X_{A2} + X_{T2} + X_{V2} \leq 80$
  - Number of Q_3: $X_{A3} + X_{T3} + X_{V3} \leq 80$
  - Non-negativity: $X_{A1}, X_{A2}, X_{A3}, X_{T1}, X_{T2}, X_{T3}, X_{V1}, X_{V2}, X_{V3} \geq 0$
  - We ignore the integer constraints for now.

Linear Programming (LP) for Google AdWords

$$\mathit{max}\ 0.5X_{A1} + 0.5X_{A2} + 1.6X_{A3} + X_{T1} + 0.75X_{T2} + 2X_{T3} + 0.5X_{V1} + 4X_{V2} + 5X_{V3}$$

Subject to

$$\begin{aligned}
0.5X_{A1} + 0.5X_{A2} + 1.6X_{A3} &\leq 170 \\
X_{T1} + 0.75X_{T2} + 2X_{T3} &\leq 100 \\
0.5X_{V1} + 4X_{V2} + 5X_{V3} &\leq 160 \\
X_{A1} + X_{T1} + X_{V1} &\leq 140 \\
X_{A2} + X_{T2} + X_{V2} &\leq 80 \\
X_{A3} + X_{T3} + X_{V3} &\leq 80 \\
X_{A1}, X_{A2}, X_{A3}, X_{T1}, X_{T2}, X_{T3}, X_{V1}, X_{V2}, X_{V3} &\geq 0
\end{aligned}$$

Solving the Linear Program in Python

- Use the the package "cvxopt".
  - pip install cvxopt
- Demonstration.
- Optimal Ad Display Strategy:

Advertiser	Query 1 "4G LTE"	Query 2 "largest LTE"	Query 3 "best LTE network"
AT&T	$X^*_{A1} = 40$	$X^*_{A2} = 40$	$X^*_{A3} = 80$
T-Mobile	$X^*_{T1} = 100$	$X^*_{T2} = 0$	$X^*_{T3} = 0$
Verizon	$X^*_{V1} = 0$	$X^*_{V2} = 40$	$X^*_{V3} = 0$

- Optimal Revenue=\$428
  - 16.3% higher than the greedy strategy (\$368)
- In practice, the problem scale is much larger.
  - Hundreds of thousands of bidders, over \$100 billion.
  - Gains from optimization models at this scale become enormous.

Linear Programming Triplet for Google AdWords

- Decision variables: How many ads to display for each advertiser and each query?

- Objective function: The revenue of Google per day.

- Constraints:
  - The total revenue from one advertiser is within its budget
  - The total displays for one query cannot exceed its total number.

Linear Programming: General Formulation

$$\max_{(x_1,x_2,\cdots,x_n)} c_1x_1 + c_2x_2 + \cdots c_nx_n$$

subject to

$$\begin{cases} G_{11}x_1 + G_{12}x_2 + \cdots + G_{1n}x_n & \leq h_1 \\ G_{2,1}x_1 + G_{2,2}x_2 + \cdots + G_{2n}x_n & \leq h_2 \\ \cdots\cdots & \\ G_{m_1,1}x_1 + G_{m_1,2}x_2 + \cdots + G_{m_1,n}x_n & \leq h_{m_1} \\ A_{11}x_1 + A_{12}x_2 + \cdots + A_{1n}x_n & = b_1 \\ A_{2,1}x_1 + A_{2,2}x_2 + \cdots + A_{2n}x_n & = b_2 \\ \cdots\cdots & \\ A_{m_2,1}x_1 + A_{m_2,2}x_2 + \cdots + A_{m_2,n}x_n & = b_{m_2} \\ x_1 \geq 0, x_2 \geq 0, \cdots, x_n \geq 0 & \end{cases}$$

Notes on LP

- Maximization and minimization problems are easily transformable.
  - $\max f(x)$ is equivalent to $\min[-f(x)]$
- In the constraints, "$\geq$" is equivalent to "$\leq$"; "$>$" is equivalent to "$<$".
  - $g(x) \geq 0$ is equivalent to $-g(x) \leq 0$; $g(x) > 0$ is equivalent to $-g(x) < 0$
- If all the constraints are strict inequalities ("$>$" and "$<$"), we cannot find an optimal solution.
  - When building LP models, try not to include strict inequalities as constraints.

Homework

- Submit your choice of final topic to me by 10:00pm, Sunday, April 14.
- Review the questions discussed today.
- Read *Analytics Edge*, Chapter 12.2-12.4.
- Finish Homework 9 (NO need to submit it).
- Read the required reading for Session 10.