The 2021 MIE Master Class Series

Computational Optimization for Data Analytics

Chun-An (Joe) Chou

Assistant Professor in Industrial Engineering
Mechanical and Industrial Engineering
Northeastern University

07/12/2021

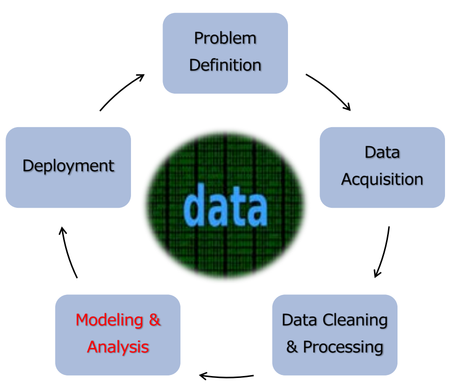
Introduction

- In this course,
  - We will introduce the concepts of **Data Science** and **Mathematical Optimization** with applications
  - Students will learn how to apply the knowledge to solve the real-world problems via data analysis, problem formulation, and result interpretation
  - Students will practice in **Python** interfaced with **Gurobi** (optimization solver)

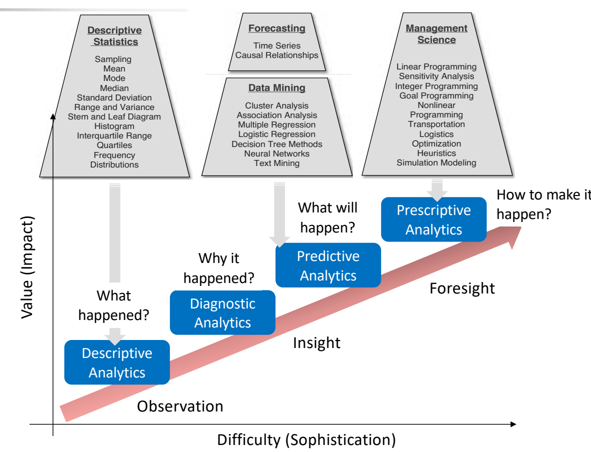
Data → Decision

- What are products to recommend for next purchase?
- What type of client is potentially default?
- Which route is faster to get to campus?
- What courses do I take?
- Does this student perform well in class?
- Who is at risk for infection with COVID19?

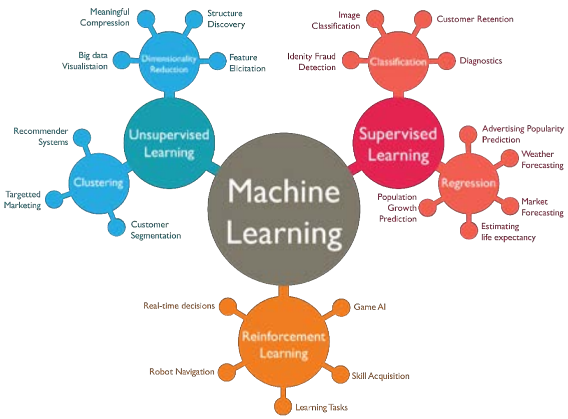
A Process (Cycle) From Data to Decision



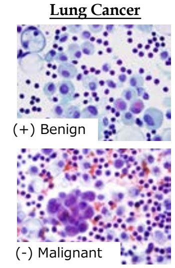
Role of Analytics Tools



Machine Learning Problem Categories

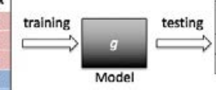


Supervised Learning for Prediction Application

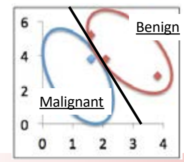


Extracted and selected features:
1) Texture_Correlation_OrigGray_3_45
2) Intensity_MassDisplacement_OrigGreen
3) Texture_InfoMeas1_OrigBlue_3_135
4) RadialDistribution_RadialCV_OrigGray_1of4
5) Texture_InfoMeas1_OrigBlue_3_45

patient	Class (θ)	Tumor size (x)	Mass index (x)
S1	+	3.8	2.8
S2	+	1.6	5.2
S3	+	2.1	3.8
S4	-	1.6	3.8
S5	-	2.1	1.0



patient	Class	Tumor size	Mass index
n1	?	2.5	2
n2	?	4.2	2

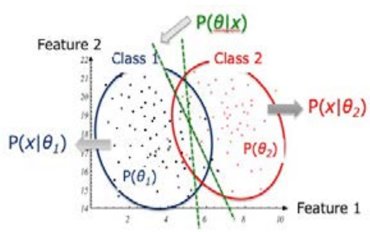


$$\underbrace{p(\theta\,|\,x)}_{\text{Posterior prob}} = \frac{\overbrace{p(x\,|\,\theta)}^{\text{Likelihood}}}{\underbrace{\int p(x\,|\,\theta)p(\theta)}_{\text{Evidence } = p(x)}} \cdot \underbrace{p(\theta)}_{\text{Prior prob}}$$

Bayes Rule for Decision

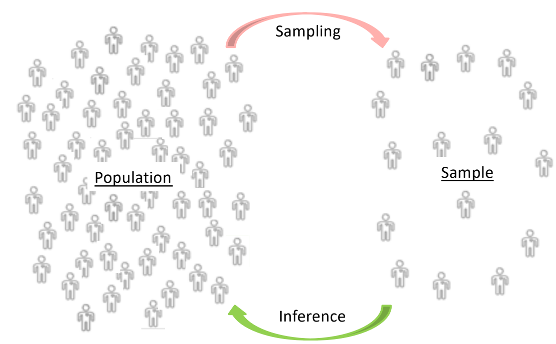
- Decision rule based on prior probability
  - Decide θ_1 if $P(\theta_1) > P(\theta_2)$; θ_2 otherwise
  - But miss sample information in each class
- Decision rule based on posterior probability
  - Decide θ_1 if $P(\theta_1|x) > P(\theta_2|x)$; θ_2 otherwise
  - Decision boundary then comes after $P(\theta|x)$ is determined

$$\underbrace{p(\theta\,|\,x)}_{\text{Posterior prob}} = \frac{\overbrace{p(x\,|\,\theta)}^{\text{Likelihood}}}{\underbrace{\int p(x\,|\,\theta)p(\theta)}_{\text{Evidence } = p(x)}} \cdot \underbrace{p(\theta)}_{\text{Prior prob}}$$



Data Sampling (Collection) is Key for Inference & Predicition

- **Sampling** is the selection of a subset (statistical sample) from within a statistical population of individuals to estimate characteristics of the whole population. (Wikipedia)



Basic Theorems of Probability - Foundation of Supervised Learning

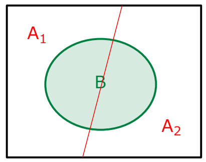
- **Probability** is defined as a likelihood of something being the case
  - A probability law for a random experiment is a rule that assigns probabilities to the events in the experiment
- **Conditional Probability**
  - Given A and B are two events, the probability of event A occurring when we already know that event B has occurred is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- **Law of Total Probability**
  - $S = A_1 \cup A_2$ and $A_1 \cap A_2 = \emptyset$
  - Sample B can be represented as

    $B = B \cap A = B \cap (A_1 \cup A_2) = (B \cap A_1) \cup (B \cap A_2)$
  - $P(B) = P(B \cap A_1) + P(B \cap A_2)$
- **Bayes' Theorem**

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)}{P(B \cap A_1) + P(B \cap A_2)} P(A)$$



Sample space S

Exercise: Probabilistic Inference

- **Example**: Employment information of males versus females in a town of 10,000 residents are sampled as follows:

	Employed	unemployed	Total
Male	460	40	500
Female	140	260	400
Total	600	300	900

  - What is the probability that you meet a man in the town?
  - What is the probability that the person you meet in the town is employed?
  - What is the probability that an employed person is a woman?
  - What is the probability that an unemployed person is a man?

Exercise: Bayes Rule for Disease Prediction

- According to a very reliable test, 1 in 1000 people carries a disease
  - Probability of carrier testing negative (false negative) is 1% (so probability of carrier testing positive is 99%)
  - Probability of non-carrier testing positive (false positive) is 5%
- Question:
  - A person just tested positive. What is the chance (s)he is a carrier of the disease?

Association Rule Mining - Introduction

- Originated with a study of customer transaction databases to determine associations among items purchased
  - "**Market Basket Analysis**" - Mine customers' behaviors
- Objective: determine associations between groups of items bought by customers
  - The frequency-based model is mostly popular to use

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

- Association Rule in an "IF-THEN" format

$$\begin{array}{ll} \{\texttt{Beer, Diapers}\} \longrightarrow \{\texttt{Milk}\}, & \{\texttt{Beer, Milk}\} \longrightarrow \{\texttt{Diapers}\}, \\ \{\texttt{Diapers, Milk}\} \longrightarrow \{\texttt{Beer}\}, & \{\texttt{Beer}\} \longrightarrow \{\texttt{Diapers, Milk}\}, \\ \{\texttt{Milk}\} \longrightarrow \{\texttt{Beer,Diapers}\}, & \{\texttt{Diapers}\} \longrightarrow \{\texttt{Beer,Milk}\}. \end{array}$$

  - What are the most "supportive" and "strong" rules for marketing?

Association Rule Mining - Algorithm

- ❑ Computational challenge occurs as data is big
  - $3^{|k|} - 2^{(|k|+1)} - 1$ association rules need to be generated for a database of *k* itemsets (or variables)
- ❑ Objective
  - Systematic find itemsets with high frequency and then association rules with high confidence
- ❑ Two-step approach
  1. Frequent Itemset (Candidate) Generation
     - Enumerate all rules (with support ≥ *minsup*)
     - Methods: Apriori algorithm, FP-Growth, etc.
  2. Rule Generation
     - Generate strong rules (with confidence ≥ *minconf* from frequent itemsets)

Association Rule Mining - Evaluation Metrics

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

- ❑ **Relative Support (*s* or *rsup*)**
  - Estimate the joint probability (or frequency) of items *X* and *Y*:

$$rsup(X \longrightarrow Y) = \frac{sup(XY)}{|\mathbf{D}|} = P(X \wedge Y)$$

|D|: total # of records

  - A symmetric measure, *e.g., sup*(*X* → *Y*) = *sup*(*Y* → *X*)
- ❑ **Confidence (*c*)**
  - Measure how often items in Y appear in records that contain X

$$c = conf(X \longrightarrow Y) = P(Y|X) = \frac{P(X \wedge Y)}{P(X)} = \frac{sup(XY)}{sup(X)}$$

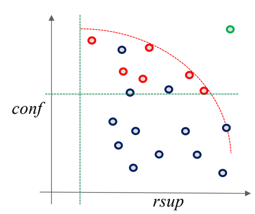
- ❑ **Lift (*l*)**
  - Measure the strength of the association (*X* → *Y*)

$$lift(X \longrightarrow Y) = \frac{P(XY)}{P(X) \cdot P(Y)} = \frac{rsup(XY)}{rsup(X) \cdot rsup(Y)} = \frac{conf(X \longrightarrow Y)}{rsup(Y)}$$

Association Rule Mining - Evaluation Metrics

- ❑ A rule is frequent if *sup*(*XY*) ≥ *minsup*
  - *minsup* is pre-determined by users
- ❑ A rule is strong if *conf*(*XY*) ≥ *minconf*
  - *minconf* is pre-determined by users

Rule Distribution



Possible rules

{Beer, Diapers} ⟶ {Milk}, {Beer, Milk} ⟶ {Diapers},
{Diapers, Milk} ⟶ {Beer}, {Beer} ⟶ {Diapers, Milk},
{Milk} ⟶ {Beer,Diapers}, {Diapers} ⟶ {Beer,Milk}.

Rule 1: {Beer} -> {Diapers}
=> *s* = 3/5 = 0.6
=> *c* = 3/3 = 1

Rule 2: $\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

Connect Association Rule to Bayesian Classification

- ❑ Objective
  - Determine "good/strong" decision rules for prediction (classification)
  - Based on Prob(*Y*=1) or Prob(*Y*=0)
- ❑ Recall *Bayesian Theorem*

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)}{P(B \cap A_1) + P(B \cap A_2)} P(A)$$

- ❑ Recall *confidence* (*c*) in association rule mining

$$c = conf(X \longrightarrow Y) = P(Y|X) = \frac{P(X \wedge Y)}{P(X)} = \frac{sup(XY)}{sup(X)}$$

- ❑ Consider a dataset $\{a_{ij}\}$ that contains
  - 17 samples of 5 binary variables
  - 2 classes

	f1	f2	f3	f4	f5	Class
0	1	0	1	0	1	1
1	1	1	1	0	0	1
2	1	0	1	0	1	1
3	1	1	1	0	0	1
4	1	0	0	1	0	1
5	1	0	1	1	1	1
6	1	1	1	1	1	1
7	1	1	0	0	0	0
8	0	0	1	0	0	0
9	0	0	0	0	0	0
10	0	1	0	0	0	0
11	0	0	0	0	0	0
12	1	1	1	0	0	0
13	1	1	1	1	0	0
14	0	0	0	0	0	0
15	0	0	1	1	0	0
16	0	0	0	0	0	0

Association Rule Generation and Selection for Classification

	itemsets	support	confidence	lift	size	Tcovered	Ncovered
0	[f1]	0.411765	0.70	1.700000	1	7	3
1	[f3]	0.352941	0.60	1.457143	1	6	4
2	[f5]	0.235294	1.00	2.428571	1	4	0
3	[f1, f3]	0.352941	0.75	1.821429	2	6	2
4	[f1, f5]	0.235294	1.00	2.428571	2	4	0
5	[f3, f5]	0.235294	1.00	2.428571	2	4	0
6	[f1, f3, f5]	0.235294	1.00	2.428571	3	4	0

Association Rule Generation and Selection for Classification

Data matrix $\{a_{ij}\}$

	f1	f2	f3	f4	f5	Class
0	1	0	1	0	1	1
1	1	1	1	0	0	1
2	1	0	1	0	1	1
3	1	1	1	0	0	1
4	1	0	0	1	0	1
5	1	0	1	1	1	1
6	1	1	1	1	1	1
7	1	1	0	0	0	0
8	0	0	1	0	0	0
9	0	0	0	0	0	0
10	0	1	0	0	0	0
11	0	0	0	0	0	0
12	1	1	1	0	0	0
13	1	1	1	1	0	0
14	0	0	0	0	0	0
15	0	0	1	1	0	0
16	0	0	0	0	0	0

Rule presentation matrix $\{b_{jk}\}$

	0	1	2	3	4	5	6
f1	1	0	0	1	1	0	1
f2	0	0	0	0	0	0	0
f3	0	1	0	1	0	1	1
f4	0	0	0	0	0	0	0
f5	0	0	1	0	1	1	1

Rule coverage matrix $\{c_{ik}\}$

	0	1	2	3	4	5	6	Class
0	1	1	1	1	1	1	1	1
1	1	1	0	1	0	0	0	1
2	1	1	1	1	1	1	1	1
3	1	1	0	1	0	0	0	1
4	1	0	0	0	0	0	0	1
5	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1
7	1	0	0	0	0	0	0	0
8	0	1	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0
12	1	1	0	1	0	0	0	0
13	1	1	0	1	0	0	0	0
14	0	0	0	0	0	0	0	0
15	0	1	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0

Indicators: sample i, feature j, and rule k

Association Rule Generation and Selection for Classification

- Decision variables: $x_i \in \{0, 1\}$ to indicate if sample i can be covered or not;
  $y_j \in \{0, 1\}$ to indicate if feature j is used in the *final* decision model or not;
  $z_k \in \{0, 1\}$ to indicate if rule k is used in the *final* decision model or not.
- Optimization Model:

$$
\begin{aligned}
\min \quad & \alpha \sum_{j=1}^{m} y_j + \beta \sum_{k=1}^{p} z_k + \gamma \sum_{i \in |I|^-} x_i - \lambda \sum_{i = \in |I|^+} x_i, \\
\text{s.t.} \quad & \sum_{k \in K} c_{ik}^{+} z_k \geq x_i \ \forall\ i \in I^+, \\
& \sum_{k \in K} c_{ik}^{-} z_k \leq M_1 x_i \ \forall\ i \in I^-, \\
& \sum_{k \in K} b_{jk} z_k \leq M_2 y_j \ \forall\ j \in J, \\
& (\sum_{i \in I^+} c_{ik} + \sum_{i \in I^-} c_{ik} - \theta_s |I|) z_k \geq 0, \ \forall\ k \in K \\
& (\frac{\sum_{i \in I^+} c_{ik}}{\sum_{i \in I^+} c_{ik} + \sum_{i \in I^-} c_{ik}} - \theta_c) z_k \geq 0, \ \forall\ k \in K \\
& (\theta_l - \sum_{j \in J} b_{jk}) z_k \geq 0 \ \forall\ k \in K, \\
& x_i, y_j, z_k \ \in \{0, 1\}.
\end{aligned}
$$

Minimize # of features and # of rules included in the decision model while ensuring that the rules are selected to minimize negative coverage and maximize positive coverage

Ensure that unplanned transfer patient *i*, if covered, is covered by at least one rule

Indicate if non-transfer patient *i* is covered by selected rules

Indicate if feature *j* is used in any selected rules

Filter out association rules that do not satisfy pre-set thresholds such as *Sup* and *Conf*

Association Rule Generation and Selection for Classification

- Recall Computational challenge occurs as data is big
  - $3^{|k|} - 2^{(|k|+1)} - 1$ association rules need to be generated for a database of *k* itemsets (or variables)
- Solution Approach
  - Set up ”proper” threshold to keep good/strong association rules
  - Find (Optimize) the best combination of association rules

Algorithm 1 Heuristic for determining strong association rules

```
procedure Run(a⁺, a⁻)
    R ← Apriori(a⁺, θs, θc, θl)          ▷ R is a set of strong association rule candidates
    Ropt ← ARSOM-R(R, a⁺, a⁻)            ▷ Ropt is a refined set of strong association rules
```

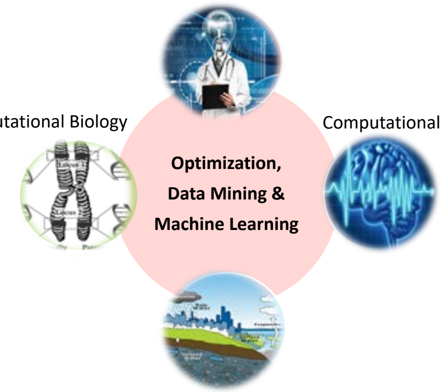
$$
\begin{aligned}
\min \quad & \alpha \sum_{j=1}^{m} y_j + \beta \sum_{k=1}^{p} z_k + \gamma \sum_{i \in |I|^-} x_i - \lambda \sum_{i = \in |I|^+} x_i, \\
\text{s.t.} \quad & \sum_{k \in K} c_{ik}^{+} z_k \geq x_i \ \forall\ i \in I^+, \\
& \sum_{k \in K} c_{ik}^{-} z_k \leq M_1 x_i \ \forall\ i \in I^-, \\
& \sum_{k \in K} b_{jk} z_k \leq M_2 y_j \ \forall\ j \in J,
\end{aligned}
$$

Associative Classification

	itemsets	support	confidence	lift	size	Tcovered	Ncovered
0	[f1]	0.411765	0.70	1.700000	1	7	3
1	[f3]	0.352941	0.60	1.457143	1	6	4
2	[f5]	0.235294	1.00	2.428571	1	4	0
3	[f1, f3]	0.352941	0.75	1.821429	2	6	2
4	[f1, f5]	0.235294	1.00	2.428571	2	4	0
5	[f3, f5]	0.235294	1.00	2.428571	2	4	0
6	[f1, f3, f5]	0.235294	1.00	2.428571	3	4	0

	f1	f2	f3	f4	f5	Class
0	1	0	1	0	1	1
1	1	1	1	0	0	1
2	1	0	1	0	1	1
3	1	1	1	0	0	1
4	1	0	0	1	0	1
5	1	0	1	1	1	1
6	1	1	1	1	1	1
7	1	1	0	0	0	0
8	0	0	1	0	0	0
9	0	0	0	0	0	0
10	0	1	0	0	0	0
11	0	0	0	0	0	0
12	1	1	1	0	0	0
13	1	1	1	1	0	0
14	0	0	0	0	0	0
15	0	0	1	1	0	0
16	0	0	0	0	0	0

What I Research?



Contact me if you have questions.
Email: ch.chou@northeastern.edu
Visit us at MOCA webpage

What I Teach?

IE 7275 Data Mining in Engineering (Fall Semester)

- The learning goal is to understand and solve classification, clustering, recommendation, and feature selection problems.
- R is the major programming tool (while Python is also good!) in this course.

IE 5400 Healthcare Systems Modeling and Analysis (Spring Semester)

- The learning goal is to understand and solve healthcare operations problems using data analytics, optimization modeling, and simulation techniques.
- MATLAB, Python, Gurobi, and Arena are optionally used for course projects to solve various healthcare decision-making problems.

Acknowledgements

- Northeastern students:
  - Tushar Sharma, M.Sc.
  - Ruilin Ouyang, Ph.D.
  - Qingtao Cao (PhD Candidate)
  - Yuchun Zou, M.Sc.

- The real case study is provided by our collaborators:
  - Shao-Jen Weng (Director of Healthcare Systems Consortium and Professor at Tunghai University)
  - Che-Hung Tsai (Vice President of Taichung VG Hospital Puli Branch)

Questions?