



Introduction to data mining

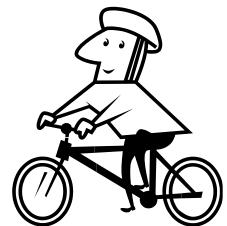
Hyerim Bae

Department of Industrial Engineering, Pusan National University
hrbae@pusan.ac.kr

What is learning

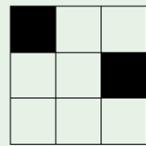
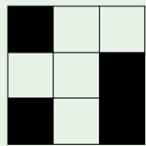
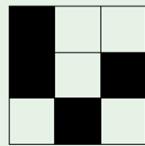
- Learning
 - A process that allows an agent to adapt its performance through **instruction** or **experience**
 - Considered fundamental to intelligent behavior
 - May be
 - Simple association task
 - A specific output is required when given some input
 - Acquisition of a skill

changes in a system that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population more efficiently and more effectively **next time**.

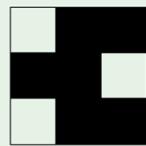
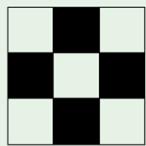
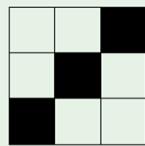


- Why?
 - Very active and large area of AI
 - Biological and cognitive perspective
 - Desire to understand more about our selves
 - Get machines to perform tasks that serve us in some way

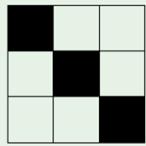
Quiz #1: What is f?



$$f = -1$$



$$f = +1$$



$$f = ?$$

Quiz #2: What is A, and what is B?

"A people bow but B people shake hands



Quiz #3: What is y?

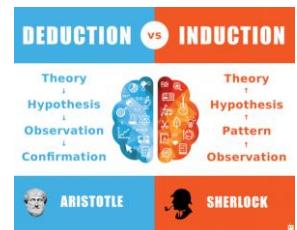
- $y=f(x) = x!$
 - $f(3) = ?$
- $y=f(x) = 3x+2$
 - $f(1)=? ?$
- $f(1)=5, f(2)=8,$
 - $f(3)=? ?$

Inductive vs. Deductive

- Induction(귀납): Specific to General
 - A dies, B dies, C dies, ...
 - Everybody dies.



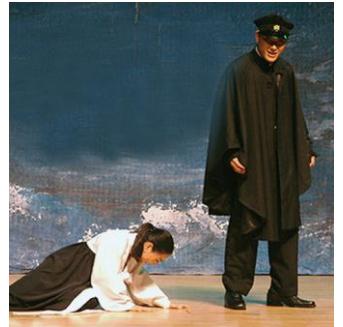
- Deduction(연역): General to specific
 - Every man dies. Socrates is a man.
 - Socrates dies.



Deductive

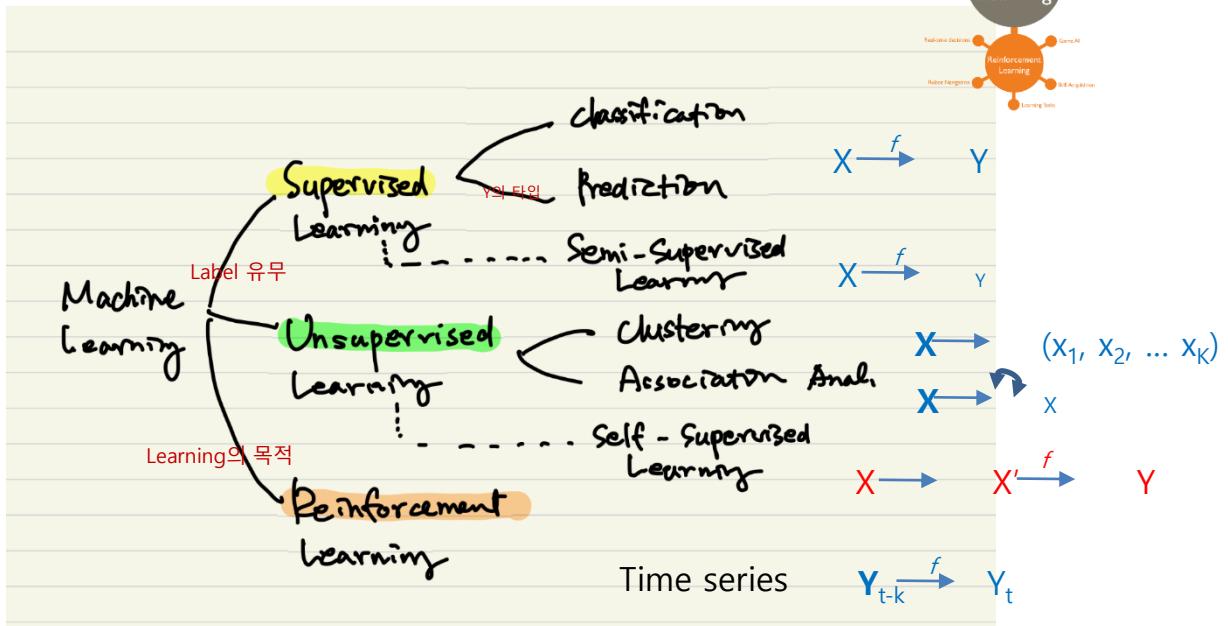
```
girl(sunae).  
boy(jungbae).  
rich(jungbae).  
pretty(sunae).  
likes(X, Y):- girl(X), boy(Y), rich(Y).  
likes(X, Y):- boy(X), girl(Y), pretty(Y).
```

- |?- likes(jungbae, sunae).



Learning method

- Traditional learning



- Deep learning
- Reinforcement learning

- Supervised vs. Unsupervised
 - Supervised learning
 - learning from training instances of known classification
 - Unsupervised learning
 - learning from unclassified training data
 - conceptual clustering or category formation
 - Reinforcement learning

What is (Machine) learning?

- Finding ' f ' such that

$$Y = f(X)$$

rule
pattern
knowledge

- We use X and Y to find ' f '

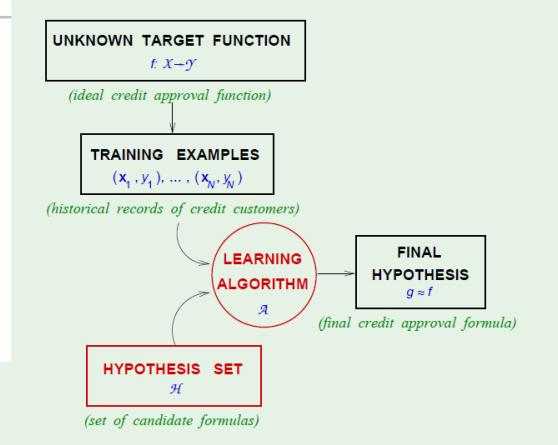
Components of learning

Formalization:

- Input: \mathbf{x} (*customer application*)
- Output: y (*good/bad customer?*)
- Target function: $f : \mathcal{X} \rightarrow \mathcal{Y}$ (*ideal credit approval formula*)
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)$ (*historical records*)



- Hypothesis: $g : \mathcal{X} \rightarrow \mathcal{Y}$ (*formula to be used*)



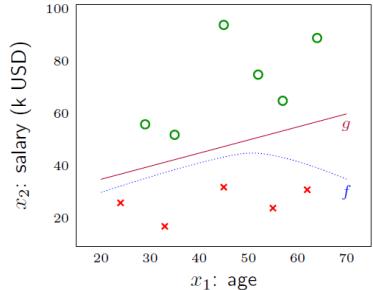
- Elements of learning
 - Algorithm
 - Define the process that is used for learning
 - Transform input data into a particular form of useful output
 - Target function
 - The product of learning
 - Training

$$\mathbf{W} \leftarrow \mathbf{W} + c(d - f)\mathbf{X}$$

$$\text{Weigh update} = \frac{\text{Direction reducing err. (decent)}}{-\eta \nabla_{\theta} J(\theta)} \times \frac{\text{Size of one step (learning rate)}}{\eta} \times \frac{\text{slope (gradient)}}{\nabla_{\theta} J(\theta)}$$

PLA (Perceptron)

n	x_1	x_2	y
1	29	56k	approve
2	64	89k	approve
3	33	17k	deny
4	45	94k	approve
5	24	26k	deny
6	55	24k	deny
7	35	52k	approve
8	57	65k	approve
9	45	32k	deny
10	52	75k	approve
11	62	31k	deny



For input $\mathbf{x} = (x_1, \dots, x_d)$ ‘attributes of a customer’

Approve credit if $\sum_{i=1}^d w_i x_i > \text{threshold}$,

Deny credit if $\sum_{i=1}^d w_i x_i < \text{threshold}$.

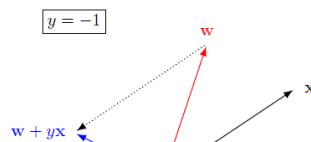
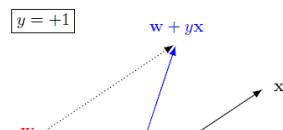
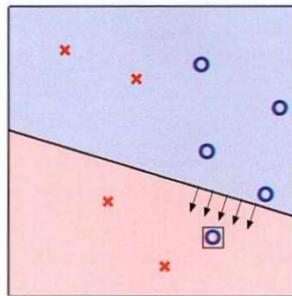
This linear formula $h \in \mathcal{H}$ can be written as

$$h(\mathbf{x}) = \text{sign}\left(\left(\sum_{i=1}^d w_i x_i\right) - \text{threshold}\right)$$

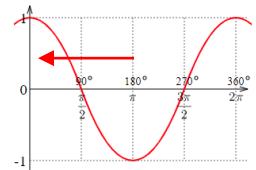
How to make ‘P’ learn

- PLA (Perceptron Learning Algorithm)

$$\mathbf{w}(t+1) = \mathbf{w}(t) + y(t)\mathbf{x}(t).$$



$$\mathbf{w}^T \mathbf{x}$$



The weight update rule in (1.3) has the nice interpretation that it moves in the direction of classifying $\mathbf{x}(t)$ correctly.

- Show that $y(t)\mathbf{w}^T(t)\mathbf{x}(t) < 0$. [Hint: $\mathbf{x}(t)$ is misclassified by $\mathbf{w}(t)$.]
- Show that $y(t)\mathbf{w}^T(t+1)\mathbf{x}(t) > y(t)\mathbf{w}^T(t)\mathbf{x}(t)$. [Hint: Use (1.3).]
- As far as classifying $\mathbf{x}(t)$ is concerned, argue that the move from $\mathbf{w}(t)$ to $\mathbf{w}(t+1)$ is a move ‘in the right direction’.

The Widrow-Hoff Procedure

- Weight update procedure:
 - Using $f = s = \mathbf{W} \cdot \mathbf{X}$
 - Data labeled 1 → 1, Data labeled 0 → -1

- Gradient: if $f = s$,

$$\frac{\partial \varepsilon}{\partial \mathbf{W}} = -2(d - f) \frac{\partial f}{\partial s} \mathbf{X} = -2(d - f) \mathbf{X}$$

- New weight vector

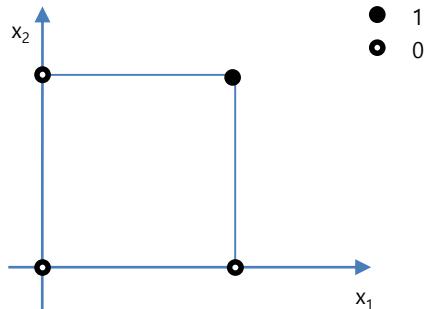
$$\mathbf{W} \leftarrow \mathbf{W} + c(d - f)\mathbf{X}$$

- Widrow-Hoff (delta) rule
 - $(d - f) > 0 \rightarrow$ increasing $s \rightarrow$ decreasing $(d - f)$
 - $(d - f) < 0 \rightarrow$ decreasing $s \rightarrow$ increasing $(d - f)$

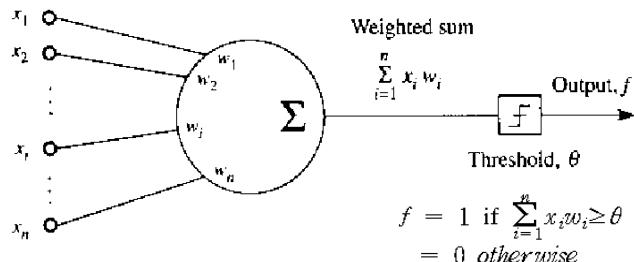
A simple classifier: Perceptron

- Dataset

input		Output (by f)		
X_0	X_1	AND	OR	XOR
0	0	0	0	0
0	1	0	1	1
1	0	0	1	1
1	1	1	1	0



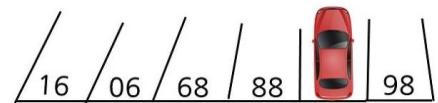
- Perceptron



We need data

- 수치형, 범주형
 - (234, 0.327, ...) (토요일, 맑음, 배혜림)
- 연속형, 이산형
 - (0.234, 0.327, ...) (0, 1)
- 정형, 비정형
 - (Table, 벡터, 리스트), (이미지, 음성, 문서)
- 균형, 비균형
 - 양, 불량
- 기계는 모든 유형의 data를 받아 들일 수 있을까요?

아래 그림에 주차된 자동차에 가려진 숫자는 무엇일까요?



Handling categorical variables
Mixed input/output



귀납적 추리 (Inductive inference)

- 사례/예제/데이터에서 명제/논제로
 - 수학적 귀납법: $k=1$ 일 때 성립함을 증명, k 일 때 성립한다고 추정, 그리고 나서 $k+1$ 일 때 성립함을 증명
 - “엔지니어들은 이러저러하다”, “비즈니스맨들은 이러저러하다.”
 - 귀납법 vs. 정형화: 사례의 수
- Cf. 연역적 추리 (deductive inference)
 - 하나의 진술에서 또 다른 진술로
 - “인간은 죽는다. 소크라테스는 인간이다. 따라서, 그는 죽는다.”



What is DM (for)?

- 의미 있는 패턴이나 법칙을 찾기 위해 다량의 데이터를 탐구하거나 분석하는 과정

Data Mining: the process of discovering new patterns from large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. (from Wikipedia)

- 소비자를 더 잘 이해함으로써 판매량과 이윤 그리고 회사의 가치를 높이는 것
- 경쟁력



언제, 어디서, 왜

- 데이터로부터 비즈니스에 관련 있는 중요한 정보/지식을 유도하는 과정
- 다량의 데이터를 분석하기 위한 일련의 강력한 테크닉/모델들

배경

- 다량의 데이터를 생성
- 다량의 데이터를 저장
- 저렴한 연산력
- 강력한 데이터마이닝 패키지 개발
- 기업들 간의 치열한 경쟁

Brief history

- 1960년대 SAS, SPSS 패키지와 더불어 통계분석에 관련됨
- 1980년대 기계학습, 데이터베이스 엔지니어링 방법
- 1990년대 데이터마이닝=‘데이터 준설’ 또는 “자백할 때까지 데이터 때리기”
- 2000년대 (예측) 분석



Application

- Traditionally
 - 통신 (SKT)
 - 금융 (Charles Swab, Capital One, Bank of America) 사기 & 마케팅
 - 보험 (State Farm)
 - 소매 (현대, Walmart, Costco)
 - 통신판매
 - 정부 (미국 FBI, CIA, 재무부-돈세탁)
 - 항공 (United Air, 대한항공)
- Recently
 - (Smart) Manufacturing?

Manufacturing?

DM의 출발점: Data(base)

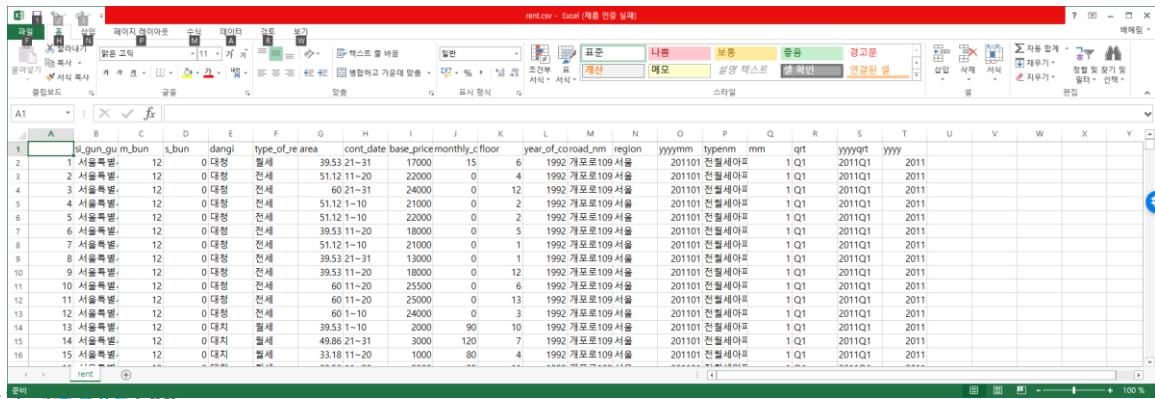
- 생산 DB
 - 시간 t에서의 기온, 압력
- 통신 소비자 DB
 - 나이, 직업, 구매 (양, 빈도, 최근성), 폰 태입
- 대출 신청인 DB
 - 나이, 직업, 수입, 자산, 대출한도, 기준 대출
- 증권 DB
 - 가격, 이자, 여타 기준들

기본 개념들

- 알고리즘(Algorithm)
 - A specific procedure used to implement a particular data mining technique: classification tree, discriminant analysis, etc.
- 독립변수 (independent variable), 종속변수 (dependent variable)
 - Predictor, input variable, field, attribute
 - Response, output variable, target variable, outcome variable
- 계수 추정 또는 파라미터 추정
- 잔차, 오류
- 학습 오류 vs. 평가 오류

Terminology and notation

- Case, Observation, Record, Pattern, Row
 - Unit of analysis on which the measurements are taken.
- Supervised learning vs. unsupervised learning
 - The process of providing an algorithm with records in which an output variable of interest is known
 - Test Data
 - Training Data
- Confidence
 - “If A and B are purchased, C is also purchased”
 - The conditional probability that C will be purchased if A and B are purchased

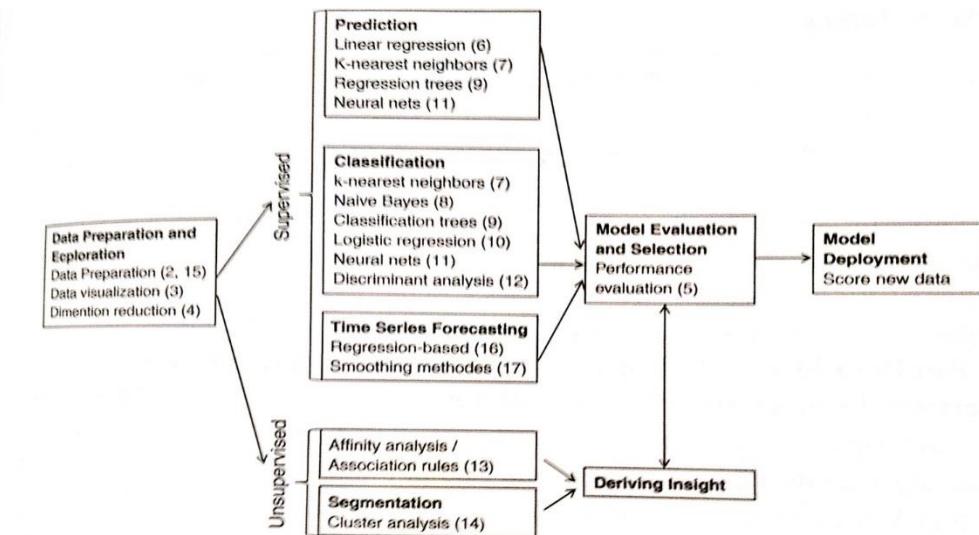


The screenshot shows a Microsoft Excel spreadsheet titled "rent.csv - Excel (제작자: 빅데이터 분석과 제작자: 김민수)" with 16 rows of data. The columns are labeled from A to Y. The data includes various variables such as address components (id, gun, gu, m_bun, s_bun, dangi), date (type_of_re_area, cont_date), price (base_price, monthly_c), floor (c_floor), location (year_of_conroad_nm, region), and time (yyymm, typenn, mm, qrt, yyyyqrt, yyyy). The last two columns show dates (date1, date2).

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	id_gun_gu_m_bun_s_bun_dangi				type_of_re_area	cont_date	base_price	monthly_c	c_floor	year_of_conroad_nm	region	yyymm	typenn	mm	qrt	yyyyqrt	yyyy							
2	1	서울특별시	12	0	대청	전세	39.53	11~31	17200	15	6	1992	개포로109	서울	201101	전월세이포	1	Q1	201101	2011				
3	2	서울특별시	12	0	대청	전세	51.12	11~20	22000	0	4	1992	개포로109	서울	201101	전월세이포	1	Q1	201101	2011				
4	3	서울특별시	12	0	대청	전세	60.21~31		24000	0	12	1992	개포로109	서울	201101	전월세이포	1	Q1	201101	2011				
5	4	서울특별시	12	0	대청	전세	51.12~10		21000	0	2	1992	개포로109	서울	201101	전월세이포	1	Q1	201101	2011				
6	5	서울특별시	12	0	대청	전세	51.12~10		22000	0	2	1992	개포로109	서울	201101	전월세이포	1	Q1	201101	2011				
7	6	서울특별시	12	0	대청	전세	39.53~11~20		18000	0	5	1992	개포로109	서울	201101	전월세이포	1	Q1	201101	2011				
8	7	서울특별시	12	0	대청	전세	51.12~10		21000	0	1	1992	개포로109	서울	201101	전월세이포	1	Q1	201101	2011				
9	8	서울특별시	12	0	대청	전세	39.53~21~31		13000	0	1	1992	개포로109	서울	201101	전월세이포	1	Q1	201101	2011				
10	9	서울특별시	12	0	대청	전세	39.53~11~20		18000	0	12	1992	개포로109	서울	201101	전월세이포	1	Q1	201101	2011				
11	10	서울특별시	12	0	대청	전세	60.11~20		25500	0	6	1992	개포로109	서울	201101	전월세이포	1	Q1	201101	2011				
12	11	서울특별시	12	0	대청	전세	60.11~20		25000	0	13	1992	개포로109	서울	201101	전월세이포	1	Q1	201101	2011				
13	12	서울특별시	12	0	대청	전세	60.1~10		24000	0	3	1992	개포로109	서울	201101	전월세이포	1	Q1	201101	2011				
14	13	서울특별시	12	0	대청	월세	39.53~11~20		2000	90	10	1992	개포로109	서울	201101	전월세이포	1	Q1	201101	2011				
15	14	서울특별시	12	0	대청	월세	49.86~21~31		3000	120	7	1992	개포로109	서울	201101	전월세이포	1	Q1	201101	2011				
16	15	서울특별시	12	0	대청	월세	33.18~11~20		1000	80	4	1992	개포로109	서울	201101	전월세이포	1	Q1	201101	2011				

Data mining

- $Y=f(X)$ 에서 f찾기 문제



예측(Prediction)

- 대표적인 예측 방법: 회귀분석(Regression)
- 독립변수 또는 예측변수로 종속변수를 예측
- 독립변수는 연속적이거나 또는 범주적일 수 있음
- 선형 회귀분석, k-근접 이웃 기법, 신경망
- 예
 - 마케팅 캠페인에 응할 것인가?
 - 다음 6개월 안에 통신 서비스를 그만둘 것인가?
 - 다음 주 주가 = $f(\text{최근 가격}, \text{경제지표})$

보르도 와인 생산과정



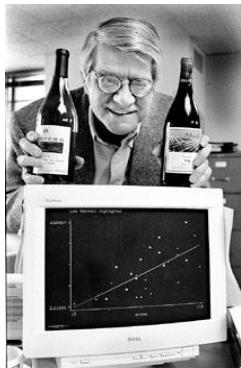
18개월 후



6개월 후



기상조건에 따른 보르도 와인의 품질은?



기온, 강수량 등등

X: 1952~1980 보르도 날씨 (월별 날씨, 핫볕, 강수량)

Vs.

Y: 보르도 와인 평균 가격 (또는 품질)

보르도 와인



[Orley Ashenfelter]

품질 = 12.145

+ 0.00117 * 겨울 강우량

+ 0.06140 * 생장기 평균 기온

- 0.00386 * 추수기 강우량

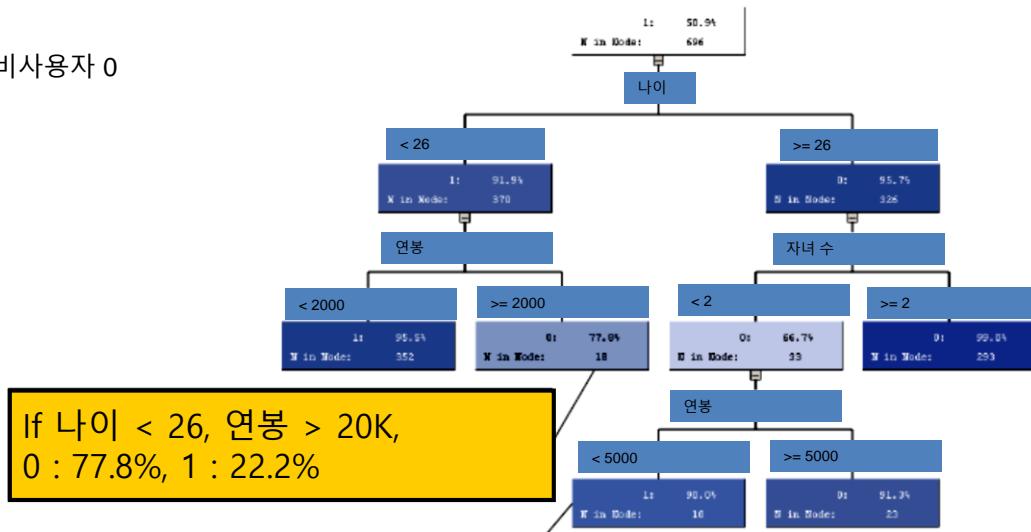
이것은 언제 계산될 수 있는가? Vs. 전문가들?

분류 (Classification)

- 범주형 종속변수를 가진 회귀분석
- 학습 데이터, 평가 데이터, 일반화
- 의사결정 나무, 신경망
- 예
 - 대출 승인 = $f(\text{개인정보})$
 - 증권가격 이동 방향 = $f(\text{최근 가격, 경제지표})$

의사결정 나무: 포털 사용자 중에서 서비스 사용자 예측

사용자 1: 비사용자 0



If 나이 > 26, 자녀 수 < 2, 연봉 < 50K ,
0 : 10.0%, 1 : 90.0%

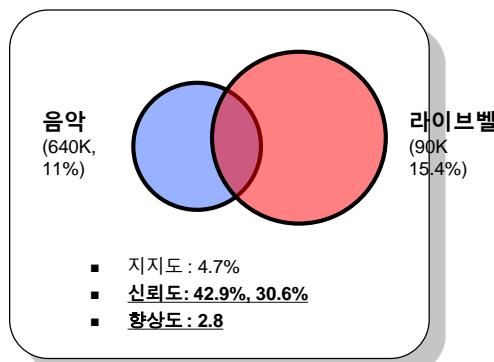
장바구니 분석

- 어떤 항목들이 함께 구매되는가?
- A Priori 알고리즘
- 예
 - A라는 영화를 본 사람들은 B라는 영화도 본다.
 - 우유와 버터는 함께 구매된다.
 - 또한, 시간관계

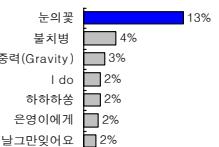
추천 시스템

- 내용 연관분석(Nate)

규칙: 음악 – 라이브벨



“눈의 꽃”
음악 구매자 중에서
라이브벨 구매



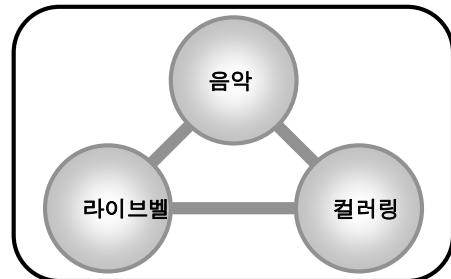
순위	음악 & 라이브벨	소비자 수
1	눈의꽃 & 눈의꽃	19,160
2	불치병& 불치병	10,300
3	눈의꽃 & 불치병	6,134
4	중력(Gravity) & 중력(Gravity)	5,705
5	불치병 & 눈의꽃	4,925
6	삭제 & 삭제	4,427
7	은영이에게 & 은영이에게	4,190
8	눈의꽃 & 중력(Gravity)	4,051
9	I do & I do	3,874
10	중력(Gravity) & 눈의꽃	3,732

* Data Source : '04.11 시점 MLB Transaction 데이터'

NATE에서 내용

- 음악 ~ 라이브벨
 - 라이브벨 ~ 컬러링
 - 컬러링 ~ 음악
-
- 이들은, 노래 레벨에서, 함께 팔리는 것으로 판명되었다!
 - 패키징과 인터페이스가 제안된다.

- 19~24세 사람들 사이에서 같은 노래 구매. 패키징과 인터페이스 디자인을 위한 비즈니스 정보를 제공할 수 있음



지식

“19~24세 사람들은 같은 노래의
각각 다른 포맷들을 구매하는 경향이 있다.”



교차판매 기회

패키징 **노래 레벨 라이트 패키지 제공**

\$4에 3~5곡을 담는 “해비” 패키지와 더불어, 노래 레벨 “라이트” 패키지(음악, 라이브렐 그리고 컬러링)가 제공된다.

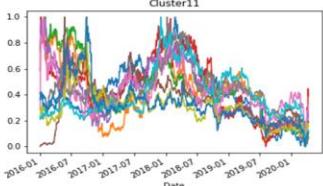
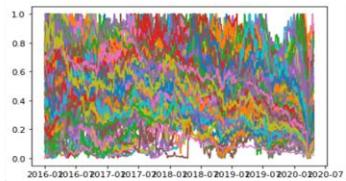
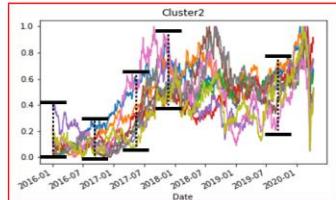
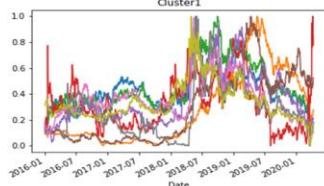
인터페이스 **같은 노래의 각각 다른 포맷들을 구매하기에 용이**

- 라이브렐을 사면, 컬러링 구매가 추천됨
- 라이브렐 구매에서 컬러링 구매로 직접 링크
- 번잡했었음: 인터페이스 위계를 오르락내리락

군집화 (Clustering)

- 비슷한 대상들의 무리
- 비교사, 탐색적 지식 발견
- C-means, **SOM**, 경쟁학습
- 예
 - 타깃 마케팅을 위한 시장 세분화
 - 체형

금융상품의 지수 변화에 따른 시계열 패턴 군집화



각각 다른 형태의 체형들

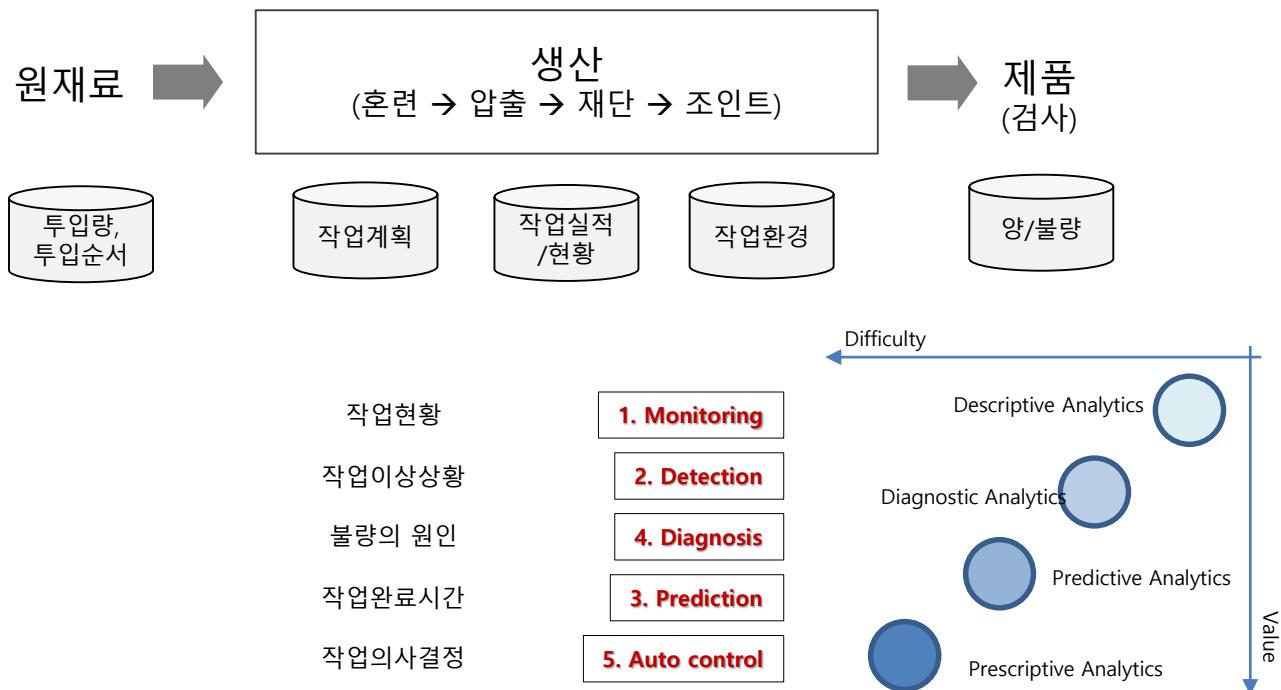
- 미 육군은 제공되는 유니폼 사이즈의 수를 줄이기 원함
- 코넬 대학 연구 팀이 여군 3,000명의 100개 이상 치수를 만듦
- 군집화의 결과 더 적은 수의 체형이 나왔고, 따라서 제공되는 유니폼의 수도 적어짐



마케팅 태스크, 데이터마이닝 태스크, 응용 분야

- 응답 모델링: 분류, 보험회사
- 고객 프로파일링: 데이터 탐색, 신용카드회사
- 시장 세분화: 군집화, 신용카드회사
- 개인화: 분류, Amazon, NYTimes
- 교차판매: 분류, MBA, 신용카드회사
- 상향판매: 분류, MBA , 신용카드회사
- 이탈예측: 분류, 회귀분석, 통신회사
- 위기예측: 분류, 회귀분석, 통신회사 또는 신용카드회사

데이터 기반 접근법의 단계



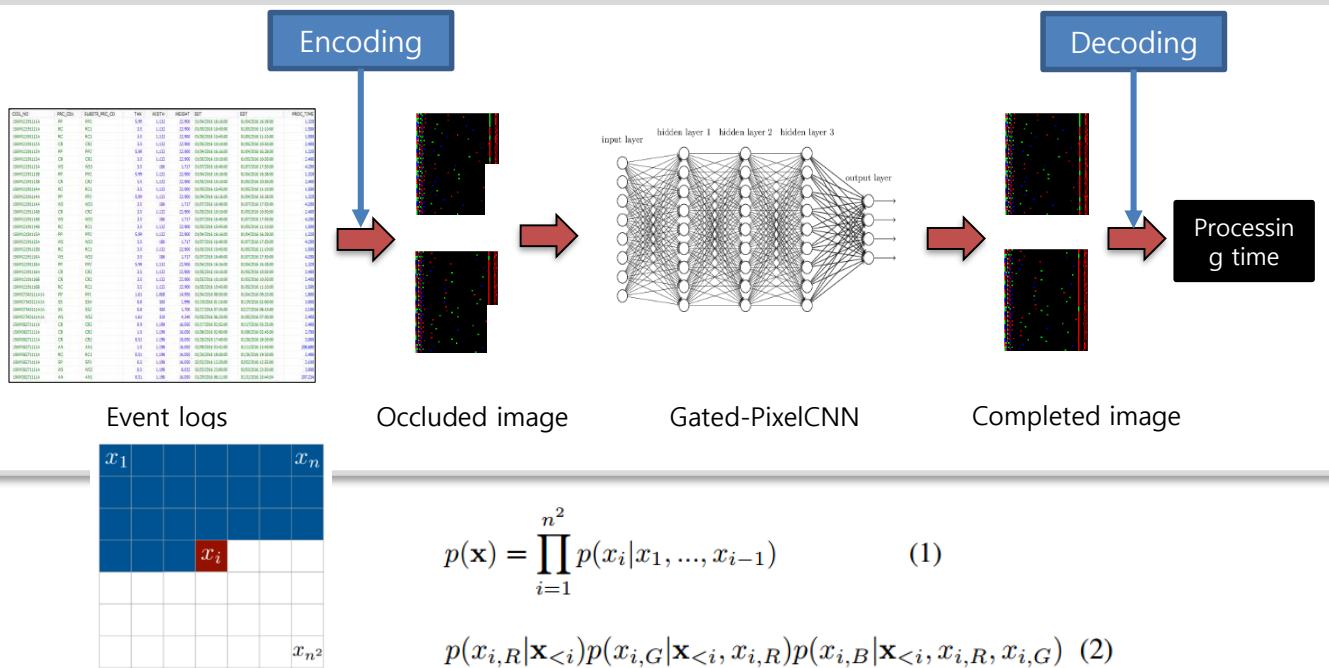
항만에서의 모니터링



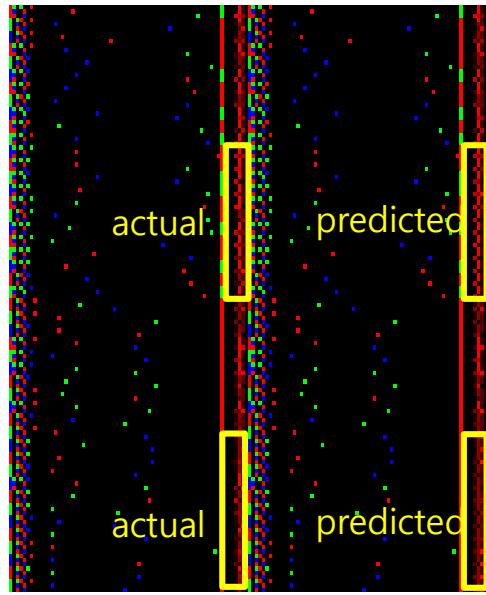
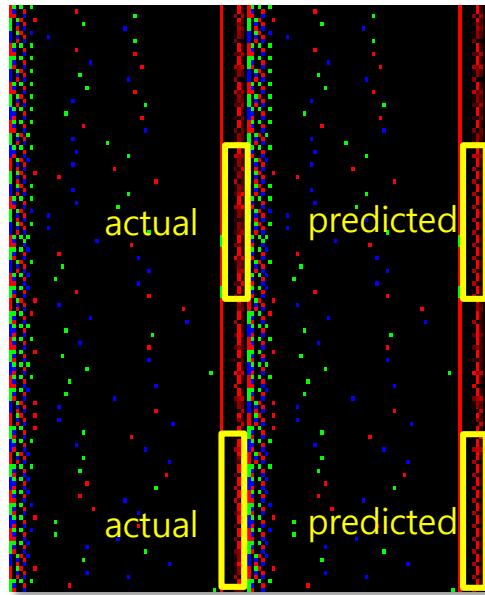
▪ SafePort 3D 모니터링 시스템



작업 완료시간 예측

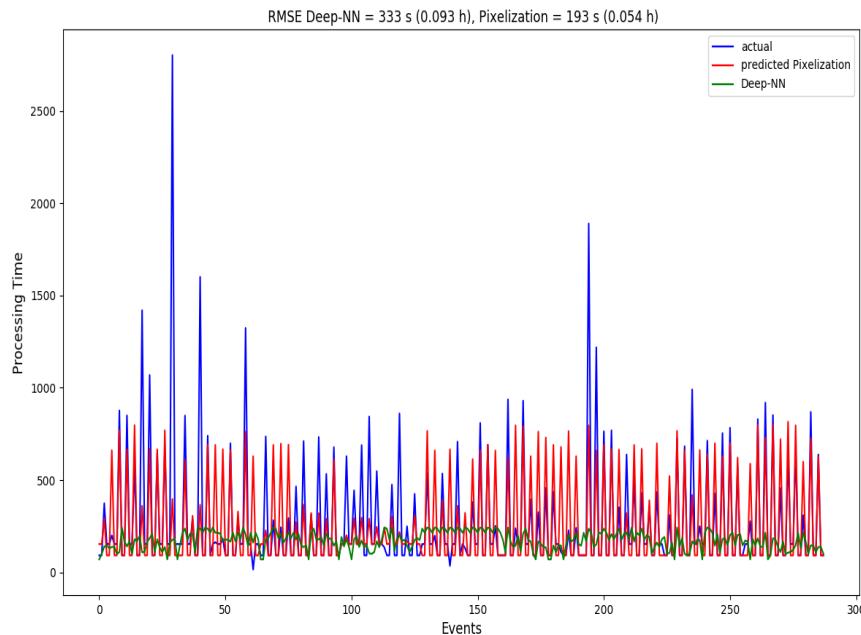


Experimental Result (1)



Experimental Result (2)

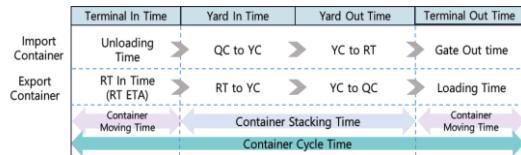
- Comparison between Deep-Neural Network and Our Approach



Deep-NN cannot perform well with categorical data !

ML for Smart Operation

가. 미래 장치장 예측 정의



- 수출입 컨테이너에 따른 Container Cycle Time, Moving Time, Stacking Time 정의
- 야드 적재량 예측을 통한 **야드 운영 계획 효율성 증대 및 미래 훈련도 계산 기반 마련**

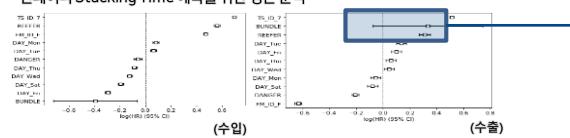
나. 분석 방법

컨테이너 특성별 Cycle Time에 대한 현황 분석

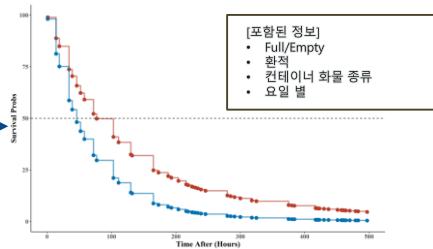


컨테이너 특성에 따른 컨테이너 물량 및 Cycle time 현황

컨테이너 Stacking Time 예측을 위한 생존 분석

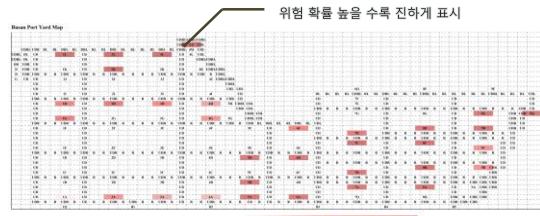


■ 생존 분석을 활용한 컨테이너 속성 별 HR Ratio



환적의 영향이 가장 크기 때문에, 환적의 유무로 HR Ratio가 표현

다. 분석 결과



생존 분석을 통한 야드 적재량 확률을 색의 진함에 따라 Grid화 표시

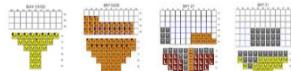
ML for Smart Operation

가. QC 작업 패턴 분석 정의

- 현재 TOS에서는 본선 작업 계획(Crane Working Plan)을 구성할 때, 양/적하 작업 수량에 따른 단순 산술 형식(Ex, 양하 * 1.5분, 적하 * 2.5분)으로 본선 작업을 계획하고 있음
- 이러한 단순 산술 형식은 실제 본선 작업 중에 찾은 Schedule 변경과 Yard Crash 등의 항만 생산성 저하 문제를 야기하고 있음
- 본선 작업 시에 더 정교한 CWP를 구성할 수 있다면, 항만 터미널 생산성을 증가시킬 수 있는 효과가 있음
- 본선 작업 시 컨테이너의 특성을 반영하여 QC(CC)의 작업 예상완료시간(ETC)을 더 정교하게 예측함을 통해, 본선 작업 예상시간(ETW)의 정확성을 높여 항만 생산성을 증가시키는 것이 목적

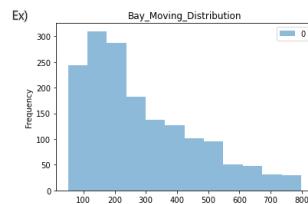
나. QC 작업 패턴 분석 방법

▪ 본선 작업 스케줄 정리 기법



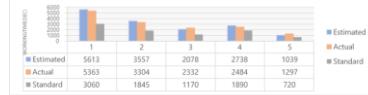
각 선박에 대한 Bay Cluster 추출

▪ 본선 작업에서의 QC의 작업 특성 별 ETC 분포 도출



Type, F/M, 작업자 변경, Bay 변경 등의 조건을 고려한 QC의 ETC 분포 기반 ETC 예측

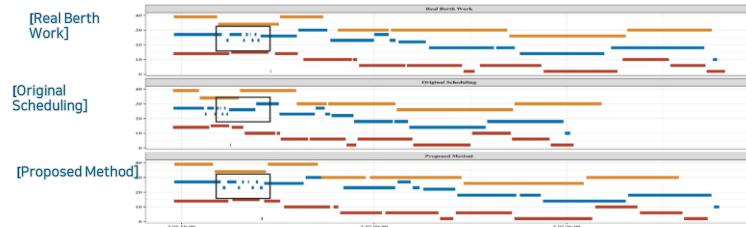
▪ ETW를 기반으로 본선 작업의 총 시간 예측



[성능 지표]

- 371개 선박 Schedule에 대한 평가 진행
- 예측 오차 평균 50.8% 감소
- 기존 산술 방식 : Median 기준 3.46 시간의 오차 발생
- 제안된 방법론 : Median 기준 1.56 시간의 오차 발생
- Yard Crash 발생 예측 정확성 향상

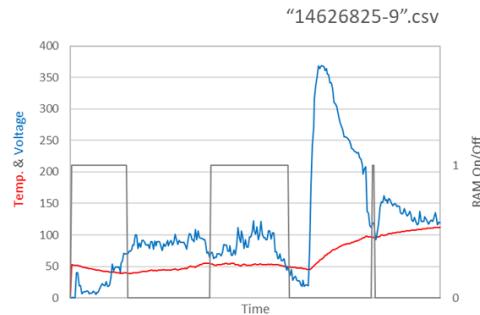
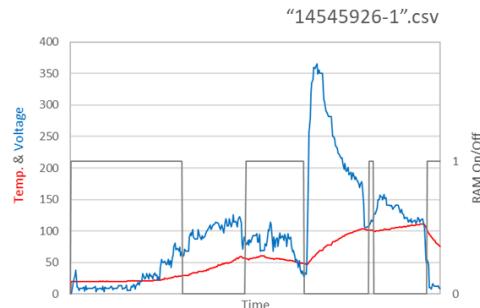
▪ 도출된 스케줄의 예상 작업시간 완료를 기준으로 ETW 예측



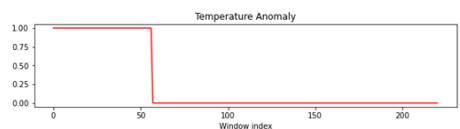
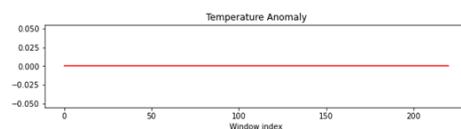
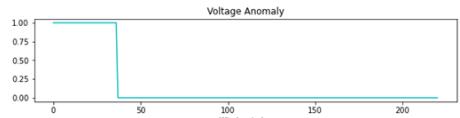
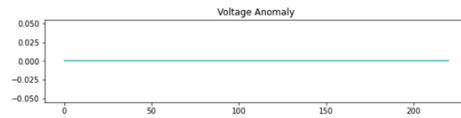
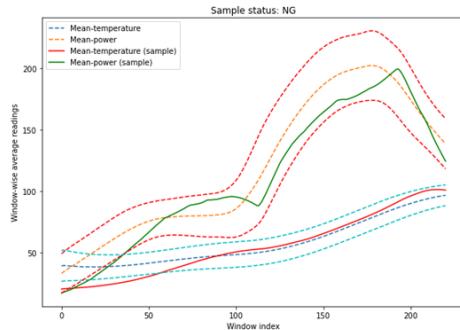
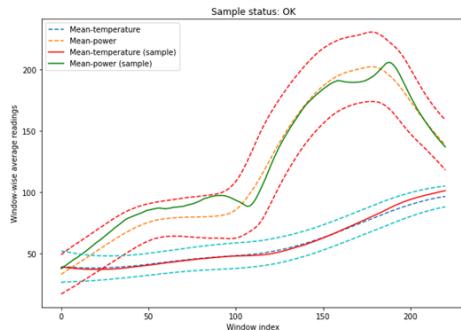
시간 대 별로 시각화 했을 때, Original Scheduling에 비해 본 연구의 제안 방법이 탐지 및 예측력이 우수함

DRB problem

- 기계의 현황
- 재고의 현황
- 현재까지의 작업 실적



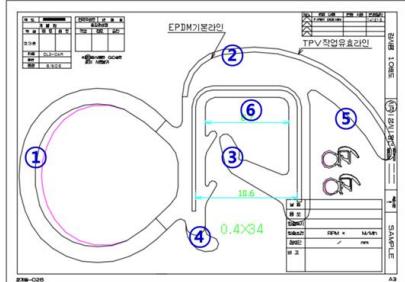
작업의 이상상황



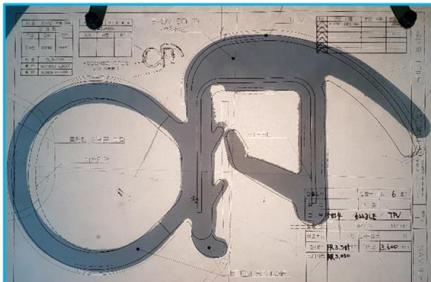
불량의 원인

1. 형상 불량 발생 유형 및 주요 제어 조건

▷ 형상 검사 기준



▷ 10배 투영 검사 결과 : 한도견본 수립후 육안검사



주요 관리 부위	불량 유형	주요 제어 조건	사용원재료
① 튜브	튜브 형상 크기/두께 산포, 튜브 상/하단부 변형	AIR 주입 압력, 스폰지발포, 가이드를 부착위치	스폰지고무(발포)
② 캐리어 상단	두께 산포, 내부 철심금 위치 산포	압출구금 마모, 압출기스크류 RPM, 프리포밍각도	슬리드고무
③ 그리프	두께/길이 산포, 끝단부 상/하 각도 산포	압출기 스크류 RPM, 단면 이송용 지지를 간섭, 엔보롤간섭	
④ 수밀립	두께/길이/각도 산포	오븐벨트 속도	
⑤ 트림립	두께/길이 산포, 끝단부 상/하 각도 산포	엔보롤 간섭, 가이드를 부착위치, 압출고무 무늬	
⑥ 벤딩 내폭	내폭 산포	사이드 벤딩룰 세팅 간격	