22년도 1학기

데이터마이닝팀

4팀

장이준
이선민
김영호
김현우
박시언

1

클린업

클린업 1주차

What is DataMining?

1. 데이터 마이닝이란?

2. 모델링은 무엇인가?
    A. 정의: Train data, Test data
    B. 지도학습, 비지도학습
    C. 지도학습: Variance-Bias Trade-off, MSE Decomposition, KNN

3. 과적합을 방지하는 방법
    A. 교차 검증 (Cross Validation)
    B. 차원의 저주: Feature selection, Feature Extraction

클린업 2주차

Tree based model & non-linear model

1. 트리 기반 모델
    - A. Decision Tree Regressor, Decision Tree Classifier
    - B. Avoid Overfitting in Tree based Models
    - C. Ensemble Methods
        - i. Bagging: Random Forest
        - ii. Boosting: GBM

2. Non-linear 모델
    - A. Basis Function
    - B. Cubic Spline
    - C. GAM(Generalized Additive Model)

클린업 3주차

Clustering & recommendation system

1. Clustering
    - A. Silhouette Method, Elbow Point Method
    - B. Non-hierarchical Clustering
        - i. K-means
        - ii. K-medoids
        - iii. DBSCAN
    - C. Hierarchical clustering

2. Recommendation system
    - A. Content-based filtering
    - B. Collaborative filtering

2

주제분석

주제선정 배경

현 서울특별시의 제로웨이스트 주요 사업으로는 "포장폐기물저감/제로웨이스트 인증제/빈병무인회수기설치/찐플리마켓/자원순환교육" 등이 존재함. 그러나 현 서울시의 제로웨이스트 프로젝트 중 "빈병 무인회수기 설치" 공약의 구체적인 실행 계획이 전무하고, 『자원순환보증금관리센터』에 의하면 현 서울시에는 24개의 무인 공병 회수기만이 존재.

<서울시 무인 공병 회수기 현황>

현 서울시에는 24개의 무인 공병 회수기만이 존재

<38대 서울특별시장 공약이행현황(2021년 12월 말 기준)>

3 **빈병 무인회수기 설치**

○ **빈병무인회수기 설치 확대 추진 : 시민들의 빈병 반환 편의성 제고**

- 시민중심 제로웨이스트 및 재활용 문화 확산
- ('22년) 자원순환보증금관리센터와 긴밀히 협력하여 **5개소 설치**
- ('23~'26년) 미설치된 자치구 위주로 **40개소 추가설치**

▼

서울시의 제로 웨이스트 프로젝트의 성공적인 이행을 위해 "빈병 무인회수기"가 우선적으로 필요한 지역을 선정하고, 무인공병회수기의 문제를 해결할 수 있는 추가 프로젝트를 진행할 예정.

2 최종 분석 흐름(1) - 무인공병회수기 입지선정

1. 무인공병회수기 입지선정

데이터 전처리/파생변수 생성

1. 행정동별 쓰레기 배출량 비율

강서구의 행정동별 음식점 수, 거주 인구 수 등 쓰레기 배출량 관련 변수들을 이용해 회귀분석을 진행한 후 타 행정동의 쓰레기 배출량을 예측 **(회귀분석 진단 & 처방, PCA, 변수 선택법)**

2. 행정동별 거주인구 비율

3. 행정동별 공병회수공간 비율

4. 공시지가&고령자 비율&수급자 비율

5. 구별 환경인식지수

ANOVA의 RCBD 모델을 통해 개인탄소감축량과 월간 버즈량(환경 관련 단어 언급 수)의 구별 차이가 유의한지 확인 후 파생 변수 생성

최종 4개 행정동 선정

1. 무인공병회수기 클러스터링

Silhouette method를 통해 최적의 클러스터 개수를 정하고 K-means clustering을 통해 16개 행정동 선정

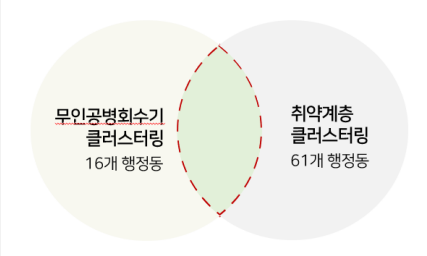
2. 취약계층 클러스터링

위와 동일한 방법으로 클러스터링 진행하고, 최종 61개 행정동 선정

3. 최종 4개 행정동 선정

무인공병회수기 타겟과 취약계층 타겟에서 동시에 선정된 행정동을 최종 행정동으로 선정.

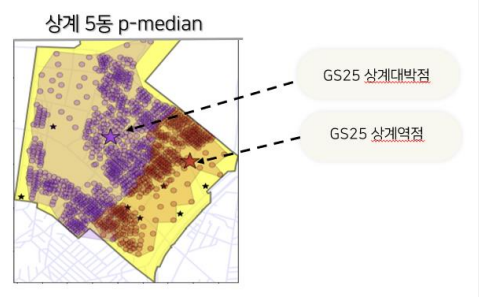
→ 상계5동, 중계2.3동, 등촌3동, 가양3동



최적 입지(편의점) 선정

1. 각 행정동별 최적의 설치장소 선정

선정된 상계5동, 중계2.3동, 등촌3동, 가양3동을 대상으로 무인공병회수기 입지 선정 진행. 입지 최적화 기법인 P-median을 사용하여 각 행정동에 무인공병회수기를 설치할 편의점을 선정.



최종 분석 흐름(2) - 훼손된 바코드 복원

2. 토이 프로젝트: 훼손된 바코드 복원

훼손된 바코드 복원 모델

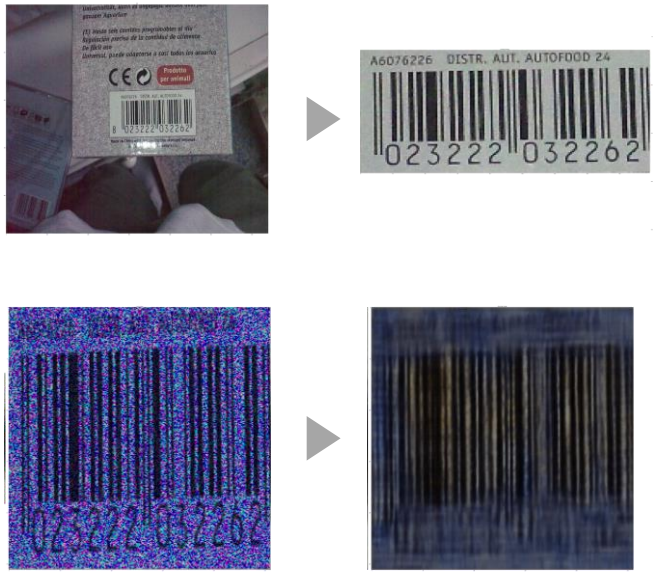
무인공병회수기의 다른 문제점으로 바코드 인식률 문제가 존재. 공병 바코드가 훼손되거나 변질된 경우에는 회수기가 일련번호를 인식하지 못해 공병을 회수하지 못하는 문제가 발생. 이를 해결하고자 훼손된 바코드를 원본 바코드로 복원하는 프로젝트 진행.

1. 사물에서 바코드 객체를 detect 후 바코드 부분만 이미지 crop

Object detection: FastRCNN

2. crop된 바코드 이미지에 노이즈를 부여하여 이를 원본 이미지로 복원

Gaussian, Salt&pepper noise method, DAE



2 파생변수 생성

행정동별 쓰레기 배출량 예측 & 구별 환경인식지수 생성

강서구를 제외한 타 행정동의 쓰레기 배출량 데이터와 환경인식 관련 변수는 현재 존재하지 않는 상태. 클러스터링에 쓰일 주요 파생변수인 행정동별 쓰레기 배출량과 구별 환경인식지수를 생성.

Train data(x) 강서구의 지역적 특성을 담은 변수

시군구	행정동	주택 수	상업지수		면적
강서구	가양1동	7921	2.215503	…	1.048
강서구	가양2동	7085	0.925131	…	-0.427
…	…	…	…	…	…

Train data(y) 강서구의 행정동별 쓰레기 배출량

행정동	주택 쓰레기 배출량
가양1동	925,380
가양2동	31,240
…	…

각 변수의 유의성을 검정해 보고, 회귀분석 진단과 그에 맞는 처방 진행.

→ 최종적으로 주택 수를 유의한 변수로 채택. 이를 통해 타 행정동의 쓰레기 배출량도 예측

강서구 행정동별 쓰레기 배출량을 예측하는 **다중선형회귀모델 생성**

자치구	MM시도언급량	MM탄소배출량	MM재활용률	환경인식지수
강남구	0.595709	0.614535	0.385465	0.398322
강동구	0.657864	0.838101	0.482759	0.434174
…	…	…	…	…
중랑구	0.092328	0.284325	0.610345	0.472783

환경인식지수

$$\frac{b + (1 - c) + r}{3}$$

b = Min-Max 스케일링한 시도 언급량

c = Min-Max 스케일링한 탄소 배출량

r = Min-Max 재활용률

2 클러스터링

무인공병회수기 클러스터링 & 취약계층 클러스터링

Silhouette method를 통해 최적의 클러스터 개수를 정하고 다양한 클러스터링 기법을 시도. 무인공병회수기 클러스터링에는 16개, 취약계층 클러스터링에는 61개의 행정동이 선정됨.

쓰레기 배출량 비율(10m)	거주 인구 비율	무인 공간 개수 비율
61.809871	0.403	1.78

공시지가	고령자 비율	수급자 비율
61.809871	0.403	1.78

* 쓰레기 배출량 비율(10m): 새로운 면적(10m기준) 비율을 고려해준 쓰레기 배출량 비율

▼

K-means | K-medoids | GMM | Hierarchical

▼

시군구	행정동
중구	청구동
...	...
성북구	길음 1동

무인공병회수기 클러스터링에서 최종 타겟으로 선정된 16개 행정동의 일부 데이터

시군구	행정동
종로구	창신1동
...	...
강동구	상일2동

취약계층 클러스터링에서 최종 타겟으로 선정된 61개 행정동의 일부 데이터

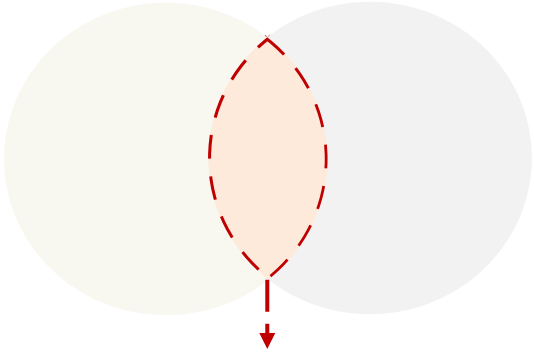
최종 클러스터링 및 행정동 선정

최종 클러스터링 및 행정동 선정

무인 공병 회수기를 설치할 최종 행정동은 무인 가능 공간 클러스터링과 취약계층 클러스터링에 모두 포함된 행정동으로 선정

무인공병회수기 클러스터링

16개 행정동



취약계층 클러스터링

61개 행정동

최종 4개 행정동 선정 노원구 중계 2.3동, 노원구 상계5동, 강서구 등촌3동, 강서구 가양3동

서울시 전체 425개 행정동 평균

	고령자비율	전체거주인구	공시지가	수급자비율
mean	0.168	22854	4528835	0.041

최종 4개 행정동 평균

	고령자비율	전체거주인구	공시지가	수급자비율
mean	0.223	26063	3348531	0.140

2 최종 입지 선정

P-median을 통한 대상 편의점 선정

선정된 상계5동, 중계2.3동, 등촌3동, 가양3동을 대상으로 무인공병회수기 입지 선정 진행. 입지 최적화 기법인 P-median을 사용하여 각 행정동에 무인공병회수기를 설치할 편의점을 선정.

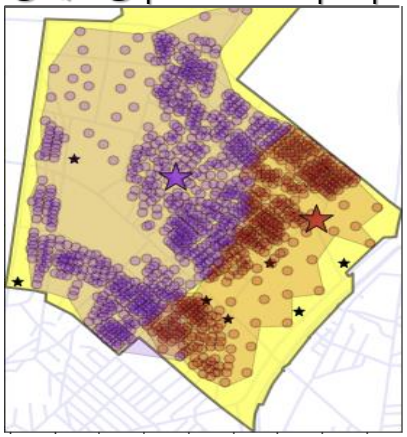
$$minimize \sum_i \sum_j h_i d_{ij} y_{ij}$$

$\sum_j y_{ij} = 1, \sum_j x_j = p\ , y_{ij} \leq x_j\ , y_{ij} \in (0,1)\ ,\ x_j \in (0,1)\ \ for\ all\ i,j$

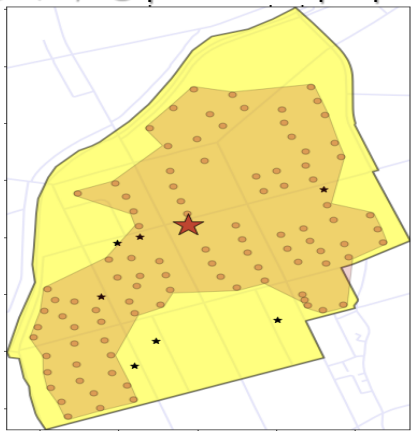
$where\ i$ = 수요지, j = 시설 입지 후보, h_i = 수요지 i의 수요량(가중치), d_{ij}= 수요지와 시설 입지 후보 간 거리,

p=시설물의 수, x_j = 시설 입지 후보 j의 시설물 설치 여부 , y_{ij} = 시설물의 총 수요 충족 여부

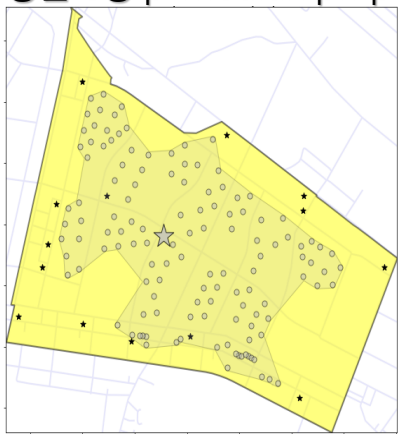
상계 5동 p-median pmp



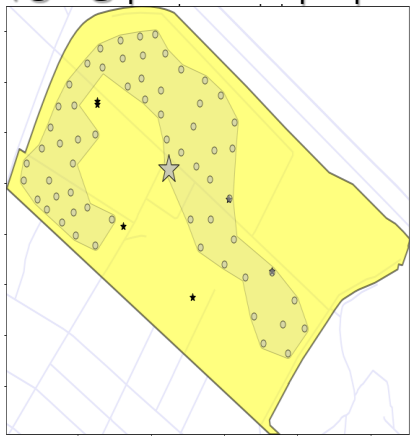
중계 2,3동 p-median pmp



등촌 3동 p-median pmp



가양 3동 p-median pmp



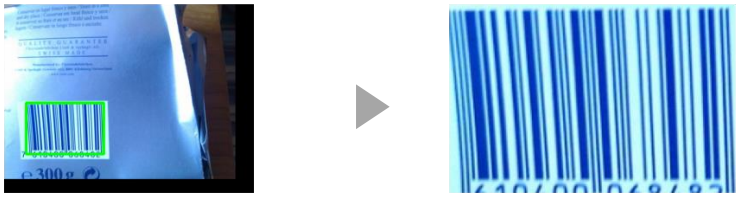
토이프로젝트

바코드 object detection & 바코드 훼손된 이미지 복원

무인공병회수기의 다른 문제점으로 바코드 인식률 문제가 존재. 공병 바코드가 훼손되거나 변질된 경우에는 회수기가 일련번호를 인식하지 못해 공병을 회수하지 못하는 문제가 발생. 이를 해결하고자 훼손된 바코드를 원본 바코드로 복원하는 프로젝트 진행.

1. 사물 이미지에서 바코드 객체를 detection

정확도를 위해 Object detection 모델 중 two-stage 모델에 해당되는 FastRCNN 사용



epoch=3, train_set: 300개 image, valid_set: 65개 이미지

Train Loss: 0.5343

Valid Loss: 0.7890

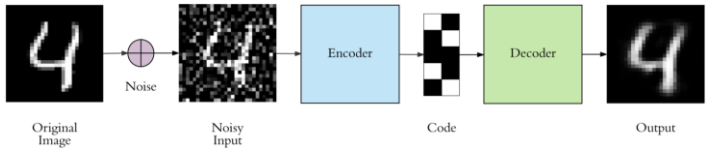
2. crop된 바코드 이미지에 노이즈를 부여하여 이를 원본 이미지로 복원

1) Noise method

Gaussian Method

Salt and Pepper Method

2) DAE(denoising auto encoder)



Final model: conv – pooling - conv – pooling – conv - pooling