3.1 분석 모형 설계

개념 1

● 통계 기반 모형

기술통계, T-test, 카이제곱검정, 분산분석 등..

개념 2

● 데이터 마이닝(머신러닝 모형)

분류, 추정, 예측, 연관분석, 군집분석 등..

개념 3

● 그리드 서치를 이용한 모델 선택

모든 하이퍼파라미터를 이용해서 최적의 파라미터를 찾는 방법이다.

개념 4

● 랜덤 서치를 이용한 모델 선택

하이퍼파라미터 값의 범위를 지정하고 지정한 범위 내에서 랜덤 샘플링 통해 하이퍼파라미터 조합을 생성하는 방법이다.

3.1.2 분석 모형 정의

개념 5

● 지도학습과 비지도학습

1. 지도학습

Label이 있는 데이터를 통해서 학습하는 방법을 지도학습이라고 말한다.

분류모형) 로지스틱 회귀모형, 신경망 모형, 결정트리, 앙상블모형, SVM, 나이브 베이즈, KNN

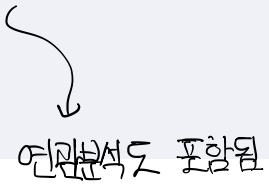
회귀모형) 회귀모형, 결정트리, SVR, 신경망 모형, 릿지, 라쏘

2. 비지도학습

Label이 없는 데이터를 통해서 학습하는 방법을 비지도학습이라고 말한다.

군집모형) K-means, SOm, DBSCAN, 병합 군집, 계층 군집

차원축소) PCA, LDA, SVD, MDS



3.1.3 분석 모형 구축 절차

개념 6

● 폭포수 모델

요구사항분석부터 설계, 구현, 테스트, 유지보수의 과정을 순차적으로 진행하는 방법

개념 7

● 모형

1. 통계 기반 분석 모형

기술통계, 상관분석, 회귀분석, 분산분석, 주성분 분석, 판별분석 등이 있다.

2. 데이터 마이닝 기반 분석 모형

분류 모형 : 트리, 앙상블 모형, 회귀 모형 등 머신러닝 모형

예측 모형 : 분류 모형에 사용되는 모형들이 사용된다.

3. 강화학습 모형

선택가능한 행동 중 보상을 최대화 하는 행동 혹은 행동순서를 선택하는 방법

개념 8

● 독립변수와 종속변수의 데이터 유형에 따른 분석 모형

1. 독립변수 : 연속형, 종속변수 : 연속형

회귀모형, 인공신경망 모형, K-최근접 이웃기법, 의사결정나무

2. 독립변수 : 연속형, 종속변수 : 이산형

로지스틱 회귀모형, 판별분석, K-최근접 이웃기법, 의사결정나무

3. 독립변수 : 이산형, 종속변수 : 연속형

회귀모형, 인공신경망 모형, 의사결정나무

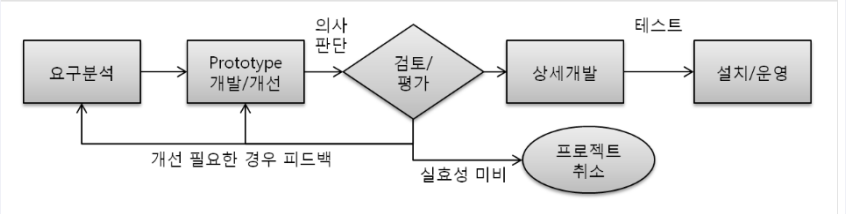
4. 독립변수 : 이산형, 종속변수 : 이산형

인공신경망 모형, 의사결정 나무, 로지스틱 회귀모형

개념 9

● 프로토타이핑 모델

폭포수 모델의 단점인 피드백에 의한 반복이 어렵다는 점을 극복하기 위해 만든 점진적 프로세스 모델이다.



1. 장점

요구사항이 모호한 경우 유용, 변경이 용이, 빠른 요구 발견 가능

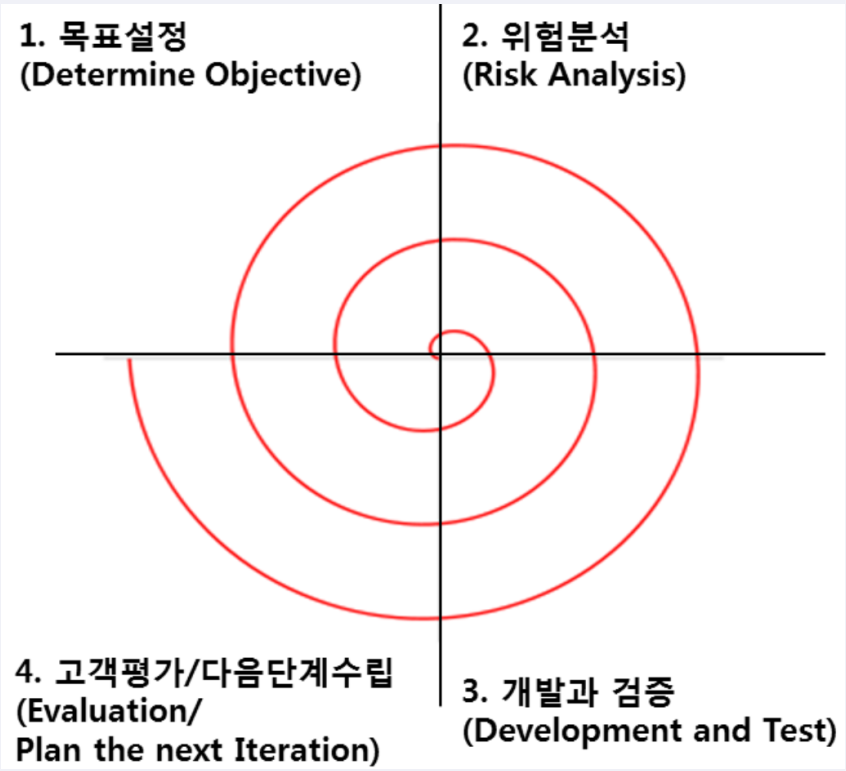
2. 단점

프로젝트를 포기할 경우 비경제적, 문서화하기 어려움, 변경으로 인해 개발 시간 및 비용이 증가할 수 있음

개념 10

● 나선형 모델

시스템을 개발하면서 발생하는 위험을 최소화하기 위해 개발 단계를 반복적으로 수행하며 완벽한 시스템을 개발하는 모델이다.



1. 장점

정확한 사용자의 요구사항 파악 가능, 위험 최소화, 대규모 프로젝트에 적합

2. 단점

많은 시간 소요, 프로젝트 관리가 어려움

3.2.1 데이터 분석 선정

개념 11

● 데이터 분석 도구

1. 엑셀

사용하기 쉽고, 메뉴들로 통해 간단한 통계분석이 가능하다.

2. R

데이터 분석을 위해 만들어진 프로그램이기에, 많은 패키지들이 있다. 다만, 메모리 및 속도가 너무 느리다는 단점이 있다.

3. 파이썬

직관적인 언어라서 배우기가 쉽고, 다양한 용도로 사용이 가능하다. 속도가 느리다는 단점이 있다.

4. 하둡

빅데이터를 저장 및 처리하는데 사용되는 프레임워크이다. 단점으로는 저장된 데이터를 변경할 수 없고 실시간 데이터 분석에는 부적합하다는 단점이 있다.

5. 맵리듀스

대용량 데이터를 분산 병렬 컴퓨팅을 통해 처리하는 프로그램이다.

6. SAS

기업체에서 주로 사용하는 대표적인 통계 프로그램이다. 간단한 명령문만으로도 여러가지 통계분석을 실행할 수 있다.

7. SPSS

사회과학 분야에서 사용하는 통계분석 프로그램

8. Stata

통계분석 프로그램

3.2.2 데이터 분할

개념 12

● 데이터 분할

전체 데이터를 학습 데이터, 검증 데이터 그리고 테스트 데이터로 분할 하는 과정을 말한다.

1. 학습 데이터

모형을 학습하는데 사용되는 데이터

2. 검증 데이터

모형이 과적합이 이루어져 있는지 확인할때 사용하는 데이터, 하이퍼 파라미터를 정할때도 사용된다.

3. 테스트 데이터

모형의 성능을 최종적으로 판단하는데 사용되는 데이터

개념 13

● 과적합(overfitting)

과적합이란 모형이 학습 데이터를 과하게 학습하여, 학습 데이터에 대한 정확도는 높지만 새로운 데이터(테스트 데이터)에 대한 정확도가 낮은 것을 의미한다.

개념 14

● 과적합

1. 과적합의 원인

훈련 데이터에 다양한 데이터가 포함되어 있지 않을때

모델이 과도하게 복잡할때

Train 데이터에 노이즈가 있을때

2. 과적합 해결방안

Train 데이터를 늘린다

라쏘, 릿지 penalty term 이용한다

모형의 모수의 수를 줄인다

개념 15

● 홀드아웃

데이터를 Train 데이터, Validation 데이터 그리고 Test 데이터로 나누어서 학습하는 방법을 말한다.

개념 16

● k-폴드 교차검증

데이터 셋을 k개로 나눈 다음 k-1개의 데이터를 학습 데이터로 사용하고, 1개의 데이터를 검증 데이터로 사용하는 방법이다.

이를 k번 반복한 다음 성능들을 평균하여 최종 성능으로 도출한다.

개념 17

● 부트스트랩

부트스트랩은 중복을 허용해서 표본추출을 허용하는 방법이다.

米 기출 정리

학생들의 교육 표준 치수를 정하기 위해 학생들의 팔길이, 키, 가슴둘레를 기준으로 학교 쓰는 방법. ⇒ 군집

페이스북 사진으로 사람을 분류.