|  |  |
| --- | --- |
| **대표 유형** | **준비도** |
| **1. 회귀분석**  - 사용 library: statsmodel  - 표준 flow  1) 가정 check: 잔차의 정규성, 등분산성, 독립성  2) 변수 선택법  3) 회귀 모형 적용 | 코드를 조금 더 다듬는 게 필요하다. |
| **2. 특성 선택**: 분산, 단일변수 선택, 특성 중요도 계산 | 문제 유형에 보다 익숙해지는 것 필요. |
| **3. 교차분석, 분산분석, 상관분석** | 문제 유형에 보다 익숙해지는 것 필요. |
| **4. 표본 추출과 t 검정**  - 사용 library: statsmodel | 표본 추출문제는 어떻게 나오지?  t 검정은 코드를 좀 더 다듬는 게 필요하다. |
| **5. 분류분석: 정형 데이터 마이닝에서 나오는 분류 모델들**  - 사용 library: sklearn | 준비는 거의 다 되어있다.  khnes code를 가져가자.  모델을 최종적으로 선택하는 논리를 어떻게 짜야하는지 정리가 필요함. |
| **6. 시계열 분석**: ARIMA, SARIMA, MARIMA, MA, AR, 정상성처리 | 전혀 준비가 되어있지 않다.  새로 code를 파야한다. |
| **8. 타겟변수 불균형 처리해결**(언더샘플링, 오버샘플링) | 코드를 좀 더 다듬는 게 필요하다. |
| **9. 군집분석**: KNN, SOM, EM 알고리즘 활용 등 | SOM, EM 알고리즘이 뭐지?  준비도 낮음. |
| **10. 연관분석**: Aprori 알고리즘 | 준비는 거의 다 되어있다.  문제 1~2개만 더 풀어볼 것. |
| 11. Datatime, Random 등 기타 pandas 활용 함수들 |  |
| 12. 텍스트마이닝 | 전혀 준비가 되어있지 않다.  새로 code를 파야한다. |

**R로 했을 때 훨씬 편한 것들.**

1. 회귀분석 변수선택법: mlxtend 쓰면 된다.

2. 시계열 통계 분석

3. 군집분석의 SOM, EM

**핵심**

기계학습의 전반에 대한 구현능력(EDA, 전처리, 자료분할, 튜닝, 지표선정, 결과해석) 과 논리적 판단

데이터에 대한 처리능력(형태변환, 그룹핑, 피벗) 및 시각화

시계열 자료의 처리 능력

불필요한(부적당한) 자료의 선별능력(시험에서 이런 데이터도 섞여있음을 충분히 고려하자)

|  |  |
| --- | --- |
| ADP 15회  -> 17회랑 거의 비슷하게 출제된 듯? | **1. 기계학습(50점)**  1.1 머신러닝: 주택가격 예측  1.2 시계열 분석: 코로나 위험지수 <- 왜 이렇게 만들었는지 설명  **2. 통계분석(50점)**  3.1 항목별 그룹별 만족도 응답의 평균, 표준편차, 왜도, 첨도  3.2 응답항목별 차이가 있는시 분석(아마 ANOVA table)  3.3 탐색적 요인분석(Factor analysis)  3.4 신뢰성 지수 개발 |
| ADP 16회 | 인터넷에 안보이는 듯? |
| ADP 17회 | **1. 기계학습: Housing Data (집값 예측)**  1.1 EDA 및 데이터 전처리(시각화 및 통계량 제시)  1.2 Train Valid Test set으로 분할 및 시각화 제시  1.3 2차 교호작용항까지 고려한 회귀분석 수행 및 변수선택과정 제시  1.4. 벌점, 앙상블을 포함하여 모형에 적합한 기계학습 모델 3가지(MSE, MAPE, R2)  **2. 시계열 분석: Corona Data**  2.1 인구대비 코로나 확진자 비율이 가장 높은 국가 5개를 제시하고, 일일확진자,  누적확진자, 일일사망자, 누적 사망자 추이를 각각 2장씩의 시각화 그래프로 시각화  (차분을 이용함)  2.2  2.3 코로나 위험지수를 개발하고, 위험지수가 높은 국가 10개를 추려내서 막대그래프  로 시각화하기  2.4 시계열 모델링 및 비선형 모델링  **3. 통계분석: Survey Data**  \*분석 전, 역코딩을 반영해야함.  3.1 항목별 그룹별 만족도 응답의 평균, 표준편차, 왜도, 첨도  3.2 응답항목별 차이가 있는시 분석(아마 ANOVA table)  3.3 탐색적 요인분석(Factor analysis)  3.4 신뢰성 지수 개발 |
| ADP 18회 | **1. 기계학습**: 고객 등급 예측 문제  \* R의 SOM 알고리즘이 나옴. python에는 package가 없다.  **2. 비정형 데이터 분석**: 영어 데이터  **3. 시계열 분석**: ARIMA (엄청 많이 나온단다) |
| ADP 19회 | **1. 기계학습(총 50점)**  1-1. 데이터 전처리 및 시각화  1-2. 분류분석 3개 및 confusion matrix 만들기(15점)  1-3. 분류분석 3개를 앙상블  **2. 통계학습(총 50점)**  2-1. 시계열 데이터 정규성, 이분산성 분석 및 시각화(10점)  2-2. 정규성이 아니라면 고정시계열이 있는지 확인하고 처리(15점)  2-3. SARIMA 분석을 실시, 여러 파라미터를 시도한 후 가장 성능 좋은 것 제시(15점)  2-4. 위에서 제시한 모델의 잔차와 잡음에 대해 시각화하고 분석(10점) |
| ADP 20회 | **1. 시계열 분석(총 50점)**  1-1 데이터 전처리 (10점)  1-2 Random forest (15점)  1-3 SVM (15점)  1-4 최적의 모델 선택하기  **2. 군집분석(총 25점)**  2-1. 클러스터링 (10점)  2-2. Heatmap으로 시각화하기(15점)  **3. 회귀분석(25점)** |
| APD 21회 | **1. 기계학습(총 50점)**  1-1. 전처리: EDA, 결측치 처리, One-hoe encoding  1-2. 분류분석: RF, SVM, XGM -> 데이터 나누고 tuning하여 최적의 모델 선택  1-3. 모델 선택: 해당 모델을 선택한 이유와 정당성, 실제 업무에서의 활용  **2. 통계분석(총 50점)**  2.1 Linear, Ridge, Lasso regression 수행 후 최적의 alpha 선택  2.2 Multinomial regression그린 후 plot 그리기  2.3 Two-way ANOVA: 결과 통계치 및 해석 |