1장

3. 넘파이

- 장점: 선형대수 기반의 프로그램을 쉽게 만들 수 있고, 루프를 사용하지 않고 대량 데이터의 배열 연산이 가능
- 단점: 수행 성능에 제약이 있어서 C/C++과 통합

- np.array(): 인자를 입력 받아서 ndarray로 변환

  adarray 생성하는 함수

- np.arange(n): 0부터 n-1까지 연속 숫자 값으로 구성된 1차원 ndarray
- np.zeros(): 0으로 채운 해당 shape 가진 ndarray
- np.ones(): 1으로 채운 해당 shape 가진 ndarray

- reshape(): 차원 및 크기 변경

  인덱싱

- 단일 값 추출
- 슬라이싱: 연속된 인덱스상의 ndarray 추출
- 팬시 인덱싱: 일정한 인덱싱 집합을 반환
- 불린 인덱싱: 특정 조건에 해당하는지에 따라 반환

  행렬의 정렬

- np.sort()
- np.argsort()

- np.dot()

- np.transpose()

4. 데이터 핸들링 - 판다스

- 판다스의 핵심 개체는 DataFrame
- DataFrame 을 넘파이 ndarray, 리스트, 딕셔너리로 변환할 수 있고 그 반대도 가능하다
- pd.drop(): 데이터 삭제

데이터 셀렉션

- [ ]
- ix[ ]
- iloc[ ]
- loc[ ]

- sort_values(): DataFrame과 Series 정렬
- groupby()

결손 데이터 처리

- isna(): 결손 데이터 여부 확인
- fillna(): 결손 데이터 대체

2장

4. model Selection 모듈 소개

- train_test_split(): 학습 및 데이터 세트를 분리해서 학습 데이터 세트를 기반으로 테스트 세트를 예측한다
- K 폴드 교차 검증: K개의 데이터 폴드 세트를 만들어서 K번만큼 각 폴트 세트에 학습과 검증 평가를 반복적으로 수행
- Stratified K 폴드: 불균형한 분포도를 가진 데이터 집합을 위한 K 폴드 방식
- cross_val_score(): 폴드 세트를 설정하고, for 루프를 통해 반복으로 학습 및 테스트 데이터의 인덱스를 추출해서 학습과 예측을 반복하는 과정을 한꺼번에 수행
- GridSearchCV()

5. 데이터 전처리

- LabelEncoder: 레이블 인코딩, 범주형 변수를 코드형 숫자 값으로 변환
- One-Hot Encoding: 원핫인코딩, 피처 값의 유형에 따라 새로운 피처를 추가해 고유 값에 해당하는 칼럼에만 1을 표시. 나머지는 0.
- StandardScaler: 개별 피처를 평균이 0이고 분산이 1인 값으로 변환
- MinMaxScaler: 0과 1 사이의 값으로 변환

3장

1. 정확도(Accuracy)

- 정확도 = 예측 결과가 동일한 데이터 건수 / 전체 예측 데이터 건수

2. 오차 행렬

- 오차 행렬: 이진 분류의 예측 오류가 얼마인지와 더불어 어떠한 유형의 예측 오류가 발생하고 있는지를 나타내는 지표
- 정확도 = 예측 결과와 실제 값이 동일한 건수 / 전체 데이터 수

  = (TN + TP) / (TN + FP + TP + FN)

3. 정밀도와 재현율

- 정밀도 = TP / (FP + TP)
- 재현율 = TP / (TP + FN)
- 정밀도가 중요한 경우는 실제 NEGATIVE 음성 데이터 예측을 POSITIVE 양성으로 잘못 판단하게 되는 경우
- 재현율이 중요한 경우는 실제 POSITIVE 양성 데이터를 NEGATIVE로 잘못 판단하게 되는 경우
- 정밀도/재현율 트레이드오프: 업무의 특성에 따라 정밀도 또는 재현율을 임계값(threshold)을 조정해서 강조하는 것. 어느 한쪽을 높이면 다른 하나는 떨어질 가능성이 있다
-

4. F1 스코어

- $F1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} = 2 \times \frac{precision \times recall}{precision + recall}$

5. ROC 곡선과 AUC

- ROC 곡선 (Receiver Operation Characteristic curve): 수신자 판단 곡선. FPR(False Positive Rate)이 변할 때 TPR(True Positive Rate)이 어떻게 변하는지 나타내는 곡선
- FPR = FP / (FP + TN) = 1 - TNR = 1 - 특이성