데이터분석 실습(지도학습)

문제번호	1	데이터셋	Breast Cancer 데이터(공개데이터)
사용패키지	sklearn	영역	분류(SVM)

사이킷런의 기본 데이터셋인 breast cancer를 사용하여 유방암을 예측하는 가장 최적의 SVM모델을 찾아내는 과정을 수행하시오. [총 80점]

문제1) breast cancer 데이터셋을 트레이닝데이터셋(X_train, y_train)과 테스트데이터셋(X_train, y_train)으로 분리하시오. 단, 종속변수 데이터의 비율을 유지하고, 트레이닝 데이터와 테스트데이터 분리비율은 70:30으로 함. random_state = 1 로 고정(10점)

문제2) 커널의 형태가 rbf, 선형, 다항인 SVM모델을 각각 생성하시오.(15점)

문제3) 5번째 코드셀에 base_model을 순서대로 추가하여 실행한 결과를 다음 표로 정리하시오. (10점)

모델		rbf	선형 (linear)	다항 (poly()
최고 정확도				
최적 하이퍼 파리미터	C			
	Gamma			
	degree			

문제4) 각 모델의 하이퍼파라미터를 사용하여 커널형태가 rbf, 선형, 다항인 SVM모델을 만들고, 학습시켜 테스트데이터에 대한 예측값을 생성하시오.
예측값과 테스트데이터셋의 종속변수값(groud truth)를 분류결과레포트(classification_report)에 입력하여 나온 결과를 다음 표로 정리하시오. (30점)

모델		rbf	선형 (linear)	다항 (poly()
정확도 (accuracy)				
평가 지표	precision			
	recall			
	f1-score			

문제5) 정확도, precision, recall의 측면에서 유방암을 예측하는 가장 좋은 모델은 3가지 SVM모델 중 어떤 것일지 자신의 생각을 서술하시오.(15점)

데이터분석 실습(비지도학습)

문제번호	2	데이터셋	iris
사용패키지	KMeans	영역	군집

사이킷런의 기본 데이터셋인 iris를 군집으로 나눌 때 최적의 그룹갯수를 구하는 과정을 수행하시오. [총 20점]

문제1) 최적의 k를 구하는 순환(for문)코드를 작성하되 다음 조건을 만족하여 작성하시오. (15점)

(1) k의 범위는 1~12, 2간격으로 입력 (5점)

(2) sklearn.cluster의 KMeans()함수 사용 (4점)

(3) kmeans객체 학습(fit) (3점)

(4) Sqaured Error를 구하는 kmeans의 속성값을 사용하시오. (4점)

문제2) 4번셀의 코드를 사용하여 "Sum of Sqaured Error" 그래프를 그리고, 이 그래프를 사용하여 최적 군집의 수를 설명하시오. (4점)