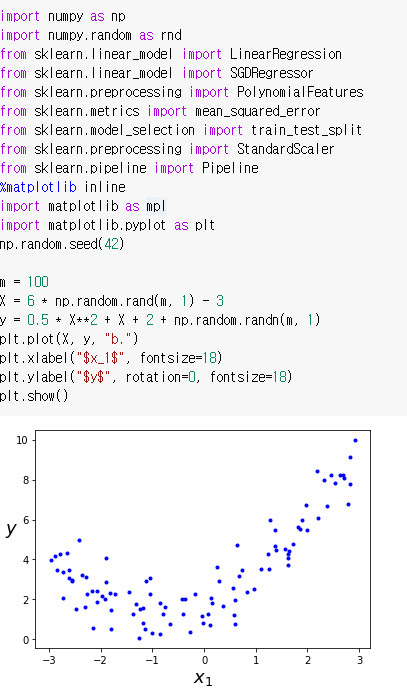
**1. 다항 회귀**

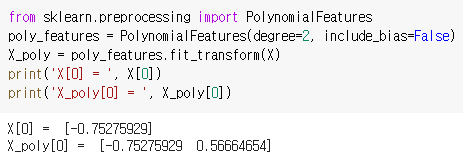
**(1) 간단한 2차방정식으로 비선형 데이터를 생성하고 그래프를 그린다.**



잡음이 포함된 비선형 데이터가 생성되었다.

직선은 이 데이터에 맞지 않을 것이다.

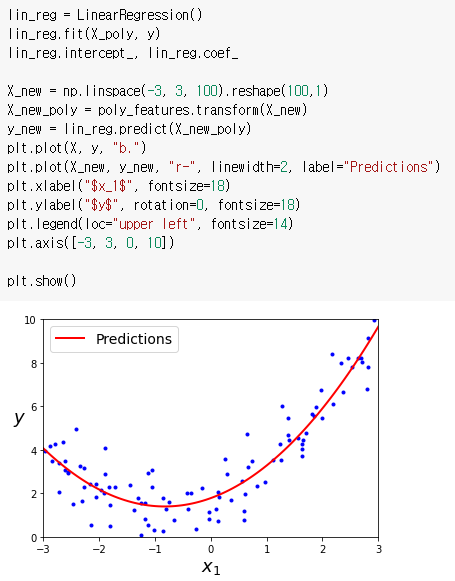
**(2) 훈련 데이터를 2차 다항으로 변환한다.**



훈련 세트에 있는 각 특성을 제곱(2차 다항)하여 새로운 특성으로 추가하였다.

X\_poly는 원래 특성 X와 이 특성의 제곱을 포함한다.

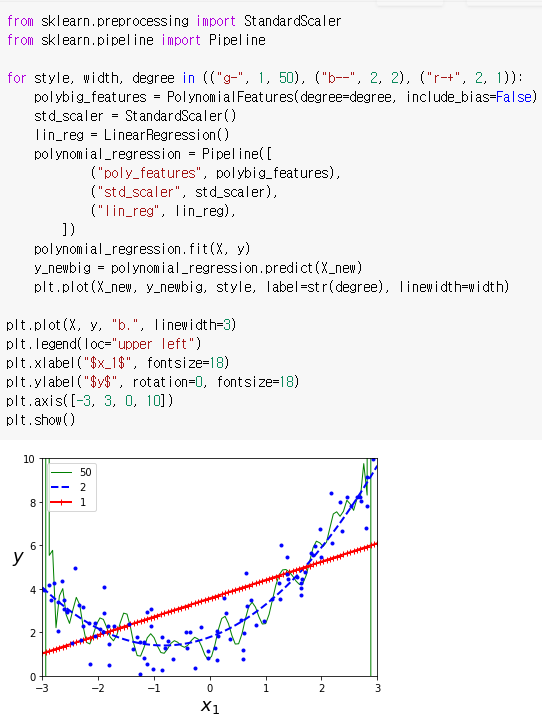
**(3) 선형 회귀를 적용하여 예측하고 훈련 데이터와 예측 그래프를 중첩하여 그래프로 그린다.**



Linear Regression을 적용하였다.

원래 함수와 예측된 모델간에 차이가 있긴 하지만 비슷한 결과를 얻었다.

**(4) 1차, 2차, 50차 다항 변환을 하고, 각 예측 값을 비교하여 그래프로 그린다.**



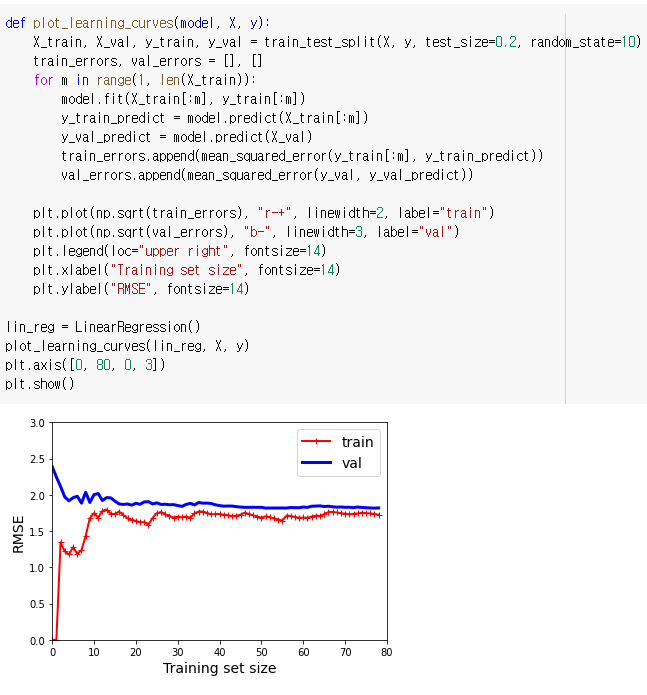
고차 다항 회귀 모델은 과대적합 되었다.

반면에 선형 모델은 과소 적합 되었다.

가장 일반화가 잘된 모델은 2차 다항회귀라고 볼 수 있다.

(2차 방정식이었기 때문에 당연하다)

**(5) 단순 선형 회귀 모델의 학습 곡선을 그래프로 그려라.**



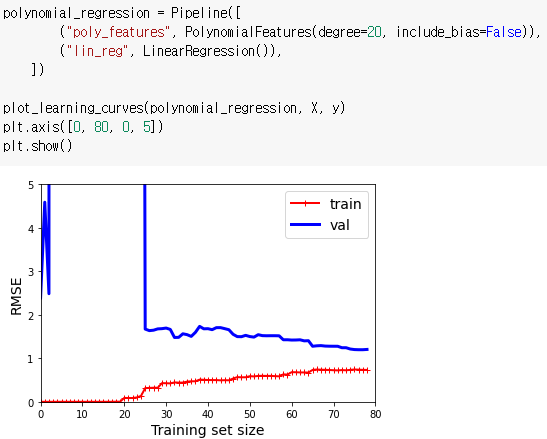
우선 과소적합된 경우라고 볼 수있다.

훈련 데이터의 그래프가 0에서 시작하므로 훈련 세트에 하나 혹은 두 개의 샘플이 있을 때는 모델이 완벽하게 작동한다. 하지만 훈련 세트에 샘플이 추가됨에 따라 잡음도 있고, 비선형이기 때문에 곡선이 어느 정도 수평이 될때까지 오차가 계속 상승한다.

검증 데이터의 경우 적은 수의 훈련 샘플로 훈련될 때는 제대로 일반화될 수 없어서 검증 오차가 초기에 매우 크다.

모델에 훈련 샘플이 추가됨에 따라 학습이 되고 검증 오차가 천천히 감소하지만, 선형 회귀의 직선은 데이터를 잘 모델링할 수 없으므로 오차의 감소가 완만해져서 훈련 세트의 그래프와 가까워진다.

**(6) 20차 다항 회귀의 학습 곡선을 그래프로 그려라.**

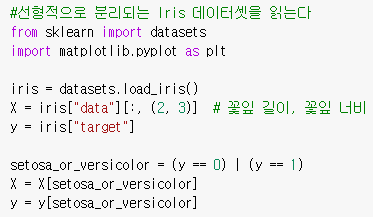


우선 훈련 데이터에서 모델 성능이 검증 데이터보다 훨씬 좋으므로 과대 적합된 모델이라고 볼 수 있다.

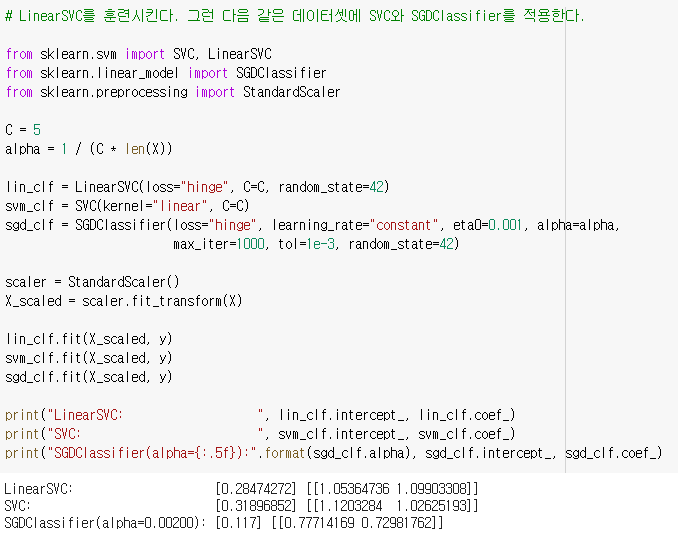
또한, 훈련데이터의 오차가 이전에 비해 선형회귀 모델보다 매우 낮다.

**2. 서포트 벡터 머신**

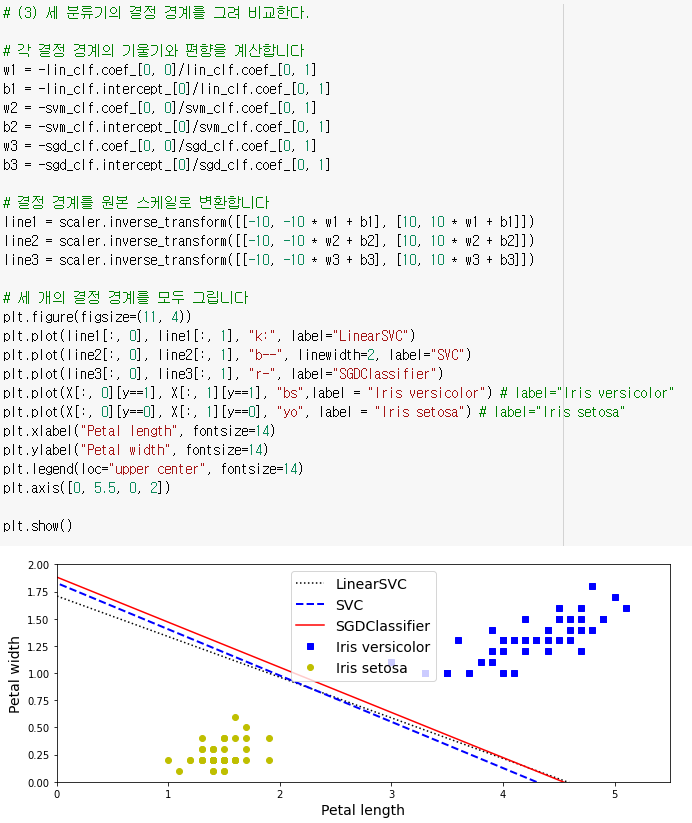
**(1) 선형적으로 분리되는 Iris 데이터셋을 읽는다. Iris Setosa와 Iris Versicolor 클래스는 선형적으로 구분이 가능하다.**

****

**(2) LinearSVC를 훈련시킨다. 그런 다음 같은 데이터셋에 SVC와 SGDClassifier를 적용한다.**

****

**(3) 위 세 분류기의 결정 경계를 그림으로 그려 비교한다.**

****

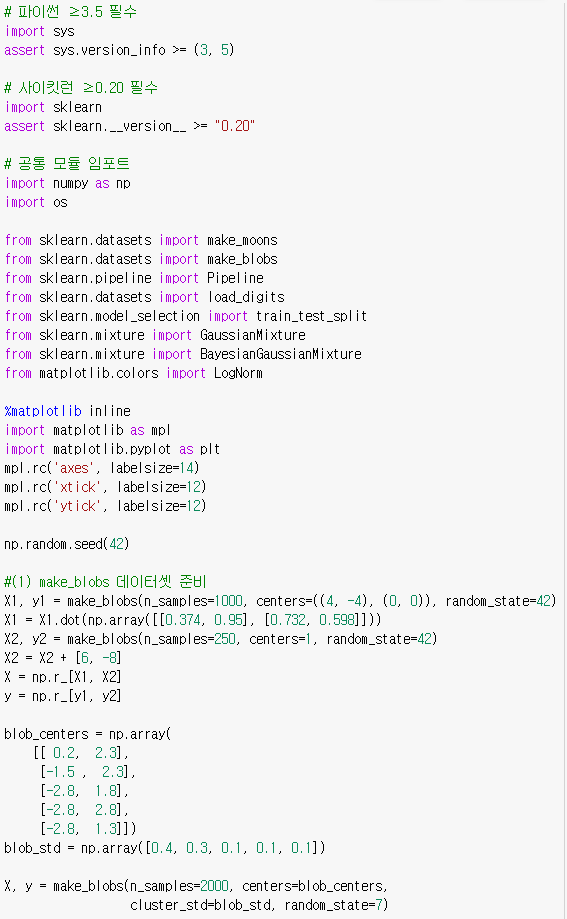
3개 모델의 결정 경계를 그려본 결과 매우 비슷한 결과를 얻었다.

모두 선형인 경계를 얻을 수 있었다.

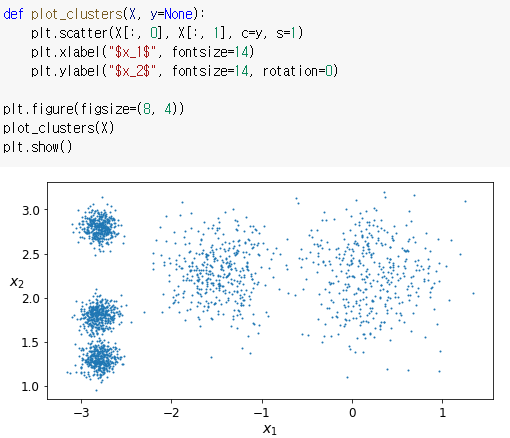
모델은 오른쪽 위에 있는 꽃들을 Iris versicolor, 왼쪽 아래에 있는 꽃들을 Iris setosa라고 판단할 것이다.

**3. 비지도 학습: 베이즈 가우시안 혼합 모델**

**(1) make\_blobs 레이블이 없는 데이터셋을 생성한다. (주교재의 그림9-2)**

****

**(2) (1)의 데이터셋을 그림으로 그린다.**

****

샘플 군집 5개를 확인 할 수 있다.

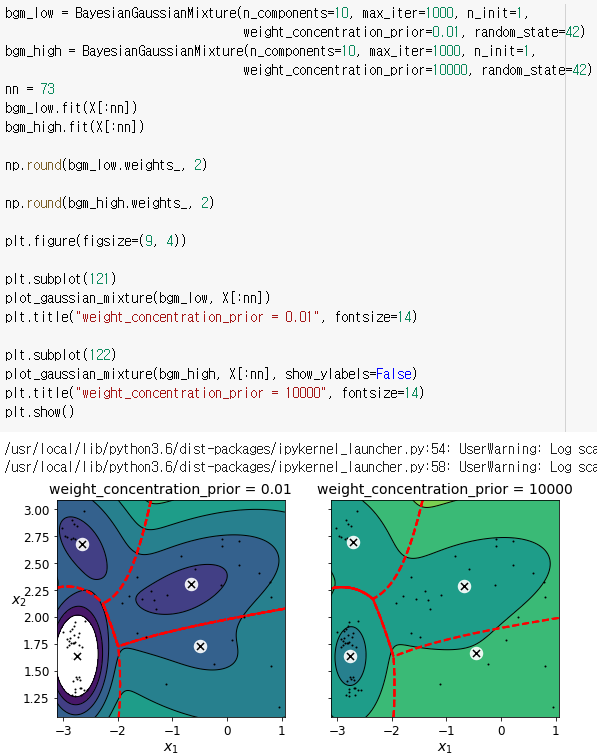
라벨이 지정되어 있지 않은 data set으로, 비지도학습을 실행하기에 적절하다. 하지만, 왼쪽의 세 개 군집은 잘 밀집되어 분류하기 쉬운 방면에 오른쪽의 두 군집은 분산이 커서 분류하기 어려울 수 있다.

**(3) 베이즈 가우시안 혼합 모델을 사용하여 농도 가중치(weight\_concentration\_prior) 파라미터 값을 조정하면서 최적 클러스터 개수를 정한다.**

베이즈 가우시안 혼합 모델을 사용했기 때문에 가중치에 따라서 클러스터의 개수 k가 달라진다.

베이즈 가우시안 혼합 모델에 의한 클러스터링을 확인할 수 있는데, k가 5로 되어 있는 것이 최적이라고 볼 수 있다.

**(4) 센트로이드와 결정 경계를 포함한 클러스터링 결과를 그림으로 그려 비교한다.**

****

4번에서는 가중치에 따른 모델의 차이를 알아볼 수 있다. 센트로이드의 위치는 달라지지 않았으나 군집의 등고선이나 인식하는 표본의 정도가 달라지는 것을 확인하였다.