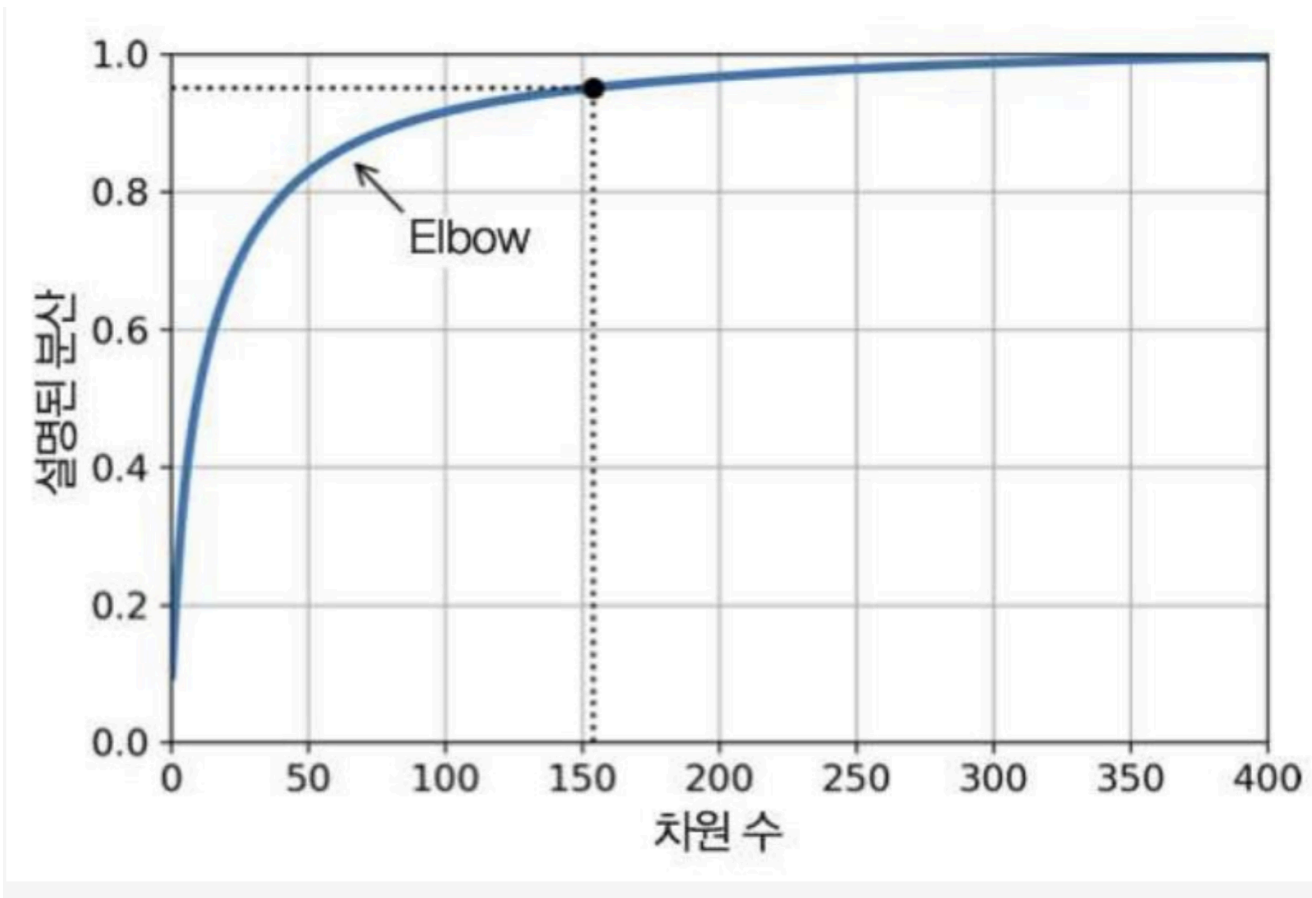
8주차_차원축소

1. 차원 축소 기법
    a. projection
        - 사영한 공간의 축을 적절하게 찾는 것이 주요 과제
        - 경우에 따라 복잡한 결과가 나올 수 있다
    b. manifold learning
2. PCA
    - 아이디어
        - 훈련 데이터에 가장 가까운 hyperplane에 데이터셋 projection
        - 분산 보존 개념 & 주성분 개념 활용
    - SVD
        - SVD 이용하면 데이터셋의 주성분 쉽게 계산 가능
        - hyperplane으로의 사영도 쉽게 계산됨
    - 적절한 차원
        - 밝혀진 분산 비율의 합이 95% 정도 되도록 하는 주성분들로 구성
        - 시각화 목적이면 2개 또는 3개
    - 설명 분산 비율 활용
        - 주목할 부분: elbow (설명 분산의 비율 합의 증가가 완만하게 변하는 지점)



- 랜덤 PCA
    - SVD 알고리즘을 확률적으로 작동하도록 만드는 기법
    - 보다 빠르게 지정된 개수의 주성분에 대한 근삿값 찾음
- 점진적 PCA
    - 훈련세트를 미니배치로 나눈후 IPCA(incremental PCA)에 하나씩 주입 가능
    - 온라인 학습에 적용 가능
    - partial_fit() 활용에 주의

3. 임의 사영
    - 존슨-린덴슈트라우스 정리
        - 고차원의 데이터를 적절한 크기의 저차원으로 임의로 사영해도 데이터셋의 정보를 많이 잃어버리지 않음을 보장
        - 오른쪽 부등식을 만족하는 d를 사영 공간의 차원으로 지정

m: 훈련셋 크기
ε: 허용된 정보손실 정도

$$d \geq \frac{4\log(m)}{\frac{1}{2}\varepsilon^2 - \frac{1}{3}\varepsilon^3}$$

- 사이킷런의 임의 사영 모델

a. GaussianRandomProjection

```
gaussian_rnd_proj = GaussianRandomProjection(eps=0.1, random_state=42)
X_reduced = gaussian_rnd_proj.fit_transform(X)
```

b. SparseRandomProjection

- sparse matrix 사용하는 GaussianRandomProjection 모델
- 대용량 데이터셋에 유용

```
gaussian_rnd_proj = SparseRandomProjection(eps=0.1, random_state=42)
X_reduced = gaussian_rnd_proj.fit_transform(X)
```

4. LLE(국소적 선형 임베딩)
    - 대표적인 다양체 학습 기법
    - 전체적으로 비선형인 다양체이지만 국소적으로는 데이터가 선형적으로 연관되어 있음
    - 국소적 관계가 가장 잘 보존되는 훈련 세트의 저차원 표현 찾을 수 있다
5. 기타 차원 축소 기법
    - 다차원 스케일링(MDS)
    - Isomap
    - t-SNE
    - 선형 판별 분석(LDA)
    - 커널 PCA