**차원 축소 요약 정리**

**차원의 저주**

* 특성 수 증가 시 학습 난이도 급격히 상승
* 고차원 공간에서 샘플 간 거리 증가, 과대적합 위험 증가
* 새로운 데이터와 기존 샘플 간 거리 멀어짐, 기존 값 기반 추정 어려움
* 고차원에서 충분한 샘플 확보 사실상 불가능, 차원 축소 필요성 대두

**차원 축소 개념**

* 특성 수 감소로 학습 가능성 확보
* 정보 손실 최소화, 훈련 속도 및 성능 극대화 목표
* 저차원 부분공간에 데이터가 분포한다는 가정
* 차원 축소 후에도 정보 손실 크지 않음

**활용 예제**

* MNIST 데이터셋: 784개 픽셀 중 154개만으로도 숫자 인식 가능
* 데이터 시각화: 2~3차원 축소 후 군집 등 시각적 패턴 탐색 가능

**차원 축소 주요 기법**

**사영(Projection)**

* n차원 데이터를 d차원 평면으로 투영
* 적절한 축(z1, z2) 선택이 핵심
* 부적절한 사영 시 정보 손실 및 샘플 구분 어려움

**다양체 학습(Manifold Learning)**

* 비선형 구조 데이터(예: 롤케이크)에서 말린 구조를 펼치는 방식
* 대부분의 고차원 데이터가 저차원 다양체에 가까운 분포라는 가설
* 사영보다 다양체 학습이 더 적합한 경우 존재

**PCA(주성분 분석)**

* 데이터 분산 최대 보존 축(주성분) 찾기
* 첫째 주성분: 분산 최대 축
* 둘째 주성분: 첫째 주성분에 직교, 남은 분산 최대 보존
* SVD(특잇값 분해) 활용, 계산량 많음
* explained\_variance\_ratio\_로 각 주성분 분산 비율 확인
* 누적 분산 비율 95% 수준에서 적절한 차원 선택
* 파일 압축, 정보 손실 최소화, 시각화, 과적합 방지 목적 활용
* 랜덤 PCA: SVD 근사, 속도 향상
* 점진적 PCA: 미니배치 기반, 온라인 학습 가능

**임의 사영(Random Projection)**

* 존슨-린덴슈트라우스 정리 기반, 임의 사영으로 정보 손실 최소화 보장
* GaussianRandomProjection, SparseRandomProjection 등 모델
* 대용량 데이터에 빠르고 메모리 효율적 적용

**LLE(국소적 선형 임베딩)**

* 대표적 다양체 학습 기법
* 국소적으로 선형 관계 보존, 비선형 구조 데이터에 적합
* 사영이 아닌 다양체 학습 방식
* 예시: 롤케이크 데이터셋의 저차원 전개

**기타 차원 축소 기법**

* 다차원 스케일링(MDS), Isomap, t-SNE, 선형 판별 분석(LDA), 커널 PCA 등
* 사이킷런에서 다양한 모델 지원

**적절한 차원 선택**

* 설명 분산 비율의 합이 95% 수준이 되도록 주성분 선택
* elbow point(완만하게 변하는 지점) 참고