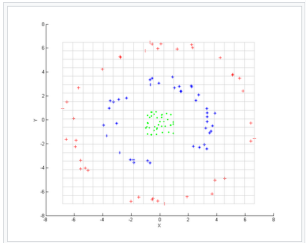
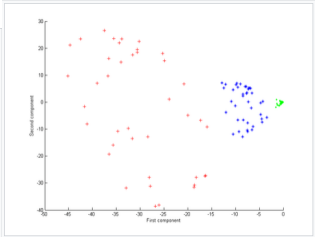
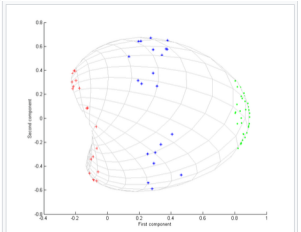
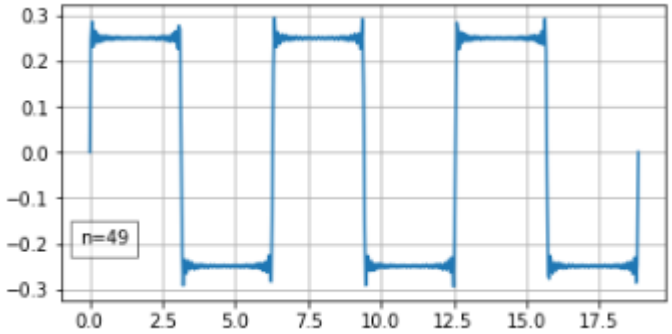
14차 일석이조 조별보고서	
작성일: 2023년 12월 5일	작성자: 김동규
조 모임 일시: 2023년 12월 5일 8교시	모임장소: 구글미트
참석자: 이준용, 유정훈, 김동규, 이학빈, 탁성재	조원: 이준용, 유정훈, 김동규, 이학빈, 탁성재

구 분	내 용
학습 범위와 내용 (조별 모임 전에 조장이 공지)	6.7 오토인코더 – 규제 오토인코더 – 적층 오토인코더 6.8 매니폴드 학습 – IsoMap – LEE – t-Sne – 귀납적 학습 모델과 트랜스덕티브 학습 모델
논의 내용 (모임 전 공지된 개별 학습 범위에서 이해된 것과 못한	Q(1) 주성분 분석 중 데이터내에 비선형적 관계가 있으면 잘 작동하지 않는다고 한다. 이때 어떻게 해결해야될지 궁금합니다. A(1)

것들)	비선형적인 관계라면 Kernal PCA가 사용되어진다. Kernel PCA란 기존의 PCA에 Kernel trick을 적용시킨 것이다. Kernel trick은 기존 formulation의 내적 부분을 kernel function을 통해 만든 kernel matrix로 대체하는 것이다. PCA를 unsupervised learning에 적용할 경우, 주어진 데이터를 우리가 정한 basis에 projection시키고, 각 basis들에 project된 값이 주어진 데이터의 feature가 되는 구조이다. 이것에 kernel trick을 적용하였기 때문에 기존의 linear projection이 nonlinear한 projection으로 바뀌게 된다. 결과는 아래와 같다.  커널 PCA 이전의 입력 지점  커널 PCA 후 출력$k(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}+1)^2$. 세 그룹은 첫 번째 구성 요소만 사용하여 구별할 수 있습니다.  가우스 커널을 사용하여 커널 PCA 후 출력 . Q2 오토인코더를 이용하면 특징 추출과 분류를 따로 학습시키게 되므로 특징 추출과 분류를 한꺼번에 시키는 CNN 에 비해 성능이 더 좋을 것 같은데, 이 이론이 맞는지, 맞다면 대부분의 이미지 인식 네트워크에서 CNN 을 사용하는 특별한 이유가 있는지 알고 싶다. A2 AutoEncoder는 Unsupervised Pretraining 기법이다. 즉, 클래스 라벨로부터 역전파를 통해 학습하는 것이 아니라 입력 값을 출력 값으로 하여 재구성할 수 있는 특징을 찾아내는 과정이다. 이는 CNN도 유사한 개념이라고 할 수 있으나, CNN은 Supervised Traning 형식이다.이전과 달리 vanishing gradient 문제를 해결한 ReLU의 등장과 Label된 데이터의 양이 증가하면서 Unsupervised Pretraining의 중요성이 감소하였고, 이후 CNN이 등장할 때에는 이미 데이터의 양이 충분했기 때문에 거의 대부분의 CNN이 Pretraining 없이 바로 Supervised Learning을 통한 학습이 가능하게 되었다. Q3 희소 코딩이란? A3 실 생활에 발생하는 모든 종류의 신호는 기저 함수 또는 기저 벡터의 선형 결합으로 되어있

다는 가정 하에 사용하는 방법입니다.
희소코딩은 어떠한 신호 또는 정보의 합성보다는 분석에 조금 더 강점을 가지는 방식입니다. 우선 훈련집합을 가지고 해당 신호를 분석하기에 최적화 된 기저 벡터를 추출합니다.
예를 들어 훈련집합으로 사각파를 입력받았다고 가정해보겠습니다.



해당 사각파는 원점을 기준으로 하는 기함수의 형태를 가지기에, 푸리에 급수를 가지고 계산할 시, sin함수와 그 정수배의 sin함수를 기저함수로 가지게 됩니다.
이후 다른 진동수를 가진 사각파가 입력될 시, 프로그램은 기저함수로 가지고 있는 sin함수를 가지고 사각파를 분석할 것입니다. 하지만 대다수의 sin함수는 0의 값을 가질 것이고, 일부만 실제 값을 가지게 될 것입니다.
따라서 해당 기저함수를 가지고 원 데이터를 표현할 경우, 수식으로 아래와 같이 표현됩니다.

$$\hat{\mathbf{D}}, \hat{\mathbf{A}} = \underset{D,A}{\operatorname{argmin}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda\phi(\mathbf{a}_i)$$

각 데이터들은 첫 항(만들어진 값과 원 데이터의 차이)과 규제항으로 구성됩니다.

Q4 오토 인코더에 대해 알아보았습니다.

A4 오토인코더는 5가지 특징을 가집니다.
1. 비지도학습이다.
2. 입력노드와 출력노드가 유사하게 학습하는 모델이다. (입력노드 수 = 출력노드 수)
3. 은닉층의 노드개수(m) < 입력층의 노드개수(d)
4. 은닉층은 입력데이터의 핵심정보를 표현한다.
5. 활성함수에 따라 비선형성, 선형성 모두 가질 수 있다.

코더의 목적함수를 수식으로 표현해보겠습니다.
X = { $\mathbf{x}1, \mathbf{x}2, \cdots, \mathbf{x}n$ } 일 때, 알아내야 하는 매개변수 *f* 와 *g*라는 매핑함수입니다.

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} L\left(\mathbf{x}_i, g\left(f\left(\mathbf{x}_i\right)\right)\right)$$

$$L\left(\mathbf{x}_i, g\left(f\left(\mathbf{x}_i\right)\right)\right) = \left\|\mathbf{x}_i - g\left(f\left(\mathbf{x}_i\right)\right)\right\|_2^2$$

참고로, 선형 매핑을 사용하고, 은닉층의 개수가 입력층의 개수보다 작으며, 위의 식을 목적함수로 사용한다면, 오토인코더가 찾아주는 가중치는 PCA가 찾아내는 주성분과 같습니다. 만약, 비선형 매핑을 사용한다면 더 POWERFUL 한 PCA라고 할 수 있습니다.

[오토인코더 규제]

오토인코더에도 규제기법을 줄 수 있습니다. 예를 들어, 은닉층 노드수 > 입력층 노드수 일 경우 규제를 적용하여 사용하게 됩니다.

1. SAE(sparse autoencoder)규제

$$\text{SAE: } \hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} L\left(\mathbf{x}_i, g\left(f(\mathbf{x}_i)\right)\right) + \lambda\phi(\mathbf{h}_i)$$

규제항을 통해 은닉층의 노드수 = 입력층 노드수로 만듭니다.

2. DAE(Denoising autoencoder) 규제

$$\text{DAE: } \hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} L\left(\mathbf{x}_i, g\left(f(\tilde{\mathbf{x}}_i)\right)\right)$$

[적층 오토인코더]

원래 오토인코더는 은닉층이 한개인 얕은 신경망입니다. 그에 반해 적층 인코더는, 은닉층을 여러 개 쌓아 깊은 구조로 확장시킨 것입니다.

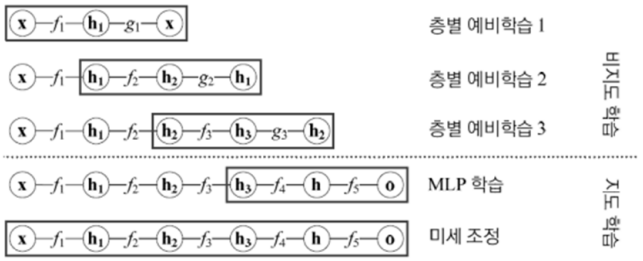


그림 6-29 적층 오토인코더의 학습 과정

앞의 층별 예배학습의 경우, 비지도 학습이고 MLP학습과 미세조정의 경우 지도학습입니다. 층별 예비학습을 통해 얻은 은닉층을 최종으로 쌓은 후(Gradient vanishing 없음), 미세조정을 통해 적층 오토인코더를 만드는 것입니다. 단, 현재는 GPU성능이 향상되어, 한꺼번에 multi-layer 알고리즘을 구현할 수 있으므로 더 이상 층별오토인코더는 쓰지 않습니다.

[확률 오토인코더]
확률모델에서는 인코딩 과정의 $\mathbf{h}=f(\mathbf{x})$ 가 $P(\mathbf{h}|\mathbf{x})$, 디코딩 단계의 $\mathbf{x}=f(\mathbf{h})$ 가 $P(\mathbf{x}|\mathbf{h})$ 로 대치됩니다. 확률모델의 가장 큰 장점은 생성모델로 활용할 수 있다는 것이 특징입니다.

Q5 ISO Map에 대해서 알아보았습니다.
A5
고차원 데이터를 저차원 매니폴드로 사상(mapping)하는 비선형 차원 감소 기법 중 하나이다. IsoMap은 주어진 데이터의 내재된 구조를 파악하고, 데이터 포인트들 간의 거리를 보존하여 저차원에서의 구조를 보다 정확하게 표현하는 데 사용된다. IsoMap은 다음과 같은 단계로 작동한다.
1. 이웃 그래프(Neighborhood Graph): 데이터 포인트들 간의 거리를 측정하여 이웃 그래프를 구성한다. 이 그래프는 데이터 간의 연결을 나타내며, 각 데이터 포인트가 서로 가까운 이웃들과 어떻게 연결되어 있는지를 표현한다.
2. 최단 경로 거리 계산(Shortest Path Distance): 이웃 그래프를 기반으로 데이터 포인트들 간의 최단 경로 거리를 계산한다. 이 과정에서 각 데이터 포인트 간의 거리를 계산할 때, 이웃 데이터 포인트들을 거쳐서 가장 짧은 경로를 찾는다.
3. 다차원 척도법(Multidimensional Scaling, MDS): 계산된 최단 경로 거리를 이용하여 저차원 매니폴드로 사상하는데, 다차원 척도법(MDS)을 사용한다. MDS는 데이터 포인트 간의 거리를 유지하면서 고차원 데이터를 저차원으로 사상하는 기법이다.

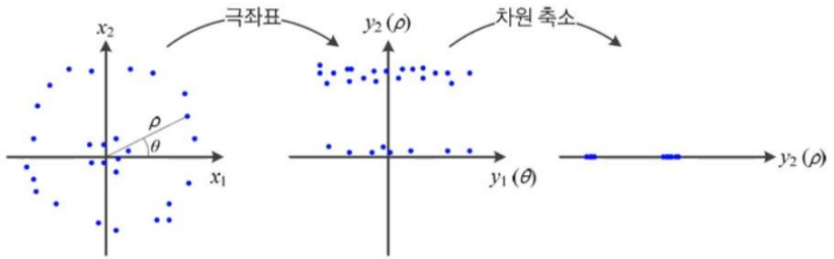
p은 선형적인 방법이 아닌 데이터의 비선형적인 구조를 잘 파악하여 차원을 축소하므로, 주로 데이터의 시각화나 저차원 특성을 추출하는 데 유용하다. 하지만 데이터의 이웃 그래프를 만드성이나 최단 경로 거리를 계산하는 과정에서 연산 비용이 크고 계산이 복잡할 수 있다.
이러한 IsoMap은 주로 이미지 분류, 시각화, 뇌 영상 데이터 분석 등 다양한 분야에서 활용되며, 데이터의 내재된 구조를 파악하여 저차원에서의 효과적인 데이터 표현을 가능하게 한다.

Q5 군집화의 주관성을 극복할 수 있는 구체적 방법에 대해서 알아보았습니다.

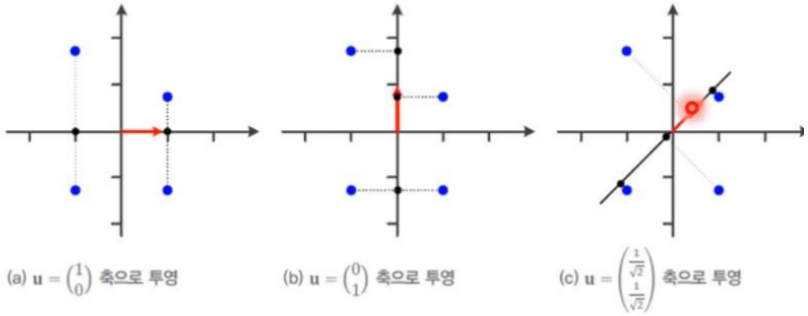
A5 **다양한 알고리즘 사용**: 여러 군집화 알고리즘을 시도하여 결과를 비교하면 주관성을 줄일 수 있습니다. K-means, 계층적 군집화, DBSCAN 등을 고려해볼 수 있습니다.
1. **하이퍼파라미터 튜닝**: 알고리즘의 매개변수를 조정하여 최적의 결과를 얻을 수 있도록 노력할 수 있습니다. 이를 통해 주관성을 줄일 수 있습니다.

	2. **표준화 및 정규화**: 입력 데이터를 표준화하거나 정규화하여 다양한 군집화 알고리즘이 민감하게 반응하는 것을 완화할 수 있습니다. 3. **앙상블 군집화**: 여러 군집화 알고리즘의 결과를 결합하여 더 강력하고 안정적인 군집을 형성할 수 있습니다. 4. **실루엣 분석**: 각 데이터 포인트의 군집 할당에 대한 품질을 측정하는 실루엣 분석을 활용하여 군집화의 품질을 평가할 수 있습니다. 5. **도메인 전문가 참여**: 군집화 결과를 해석하는 데 도움이 되는 도메인 전문가와 협력하여 주관성을 줄일 수 있습니다. 6. **앙상블 학습**: 여러 모델의 예측을 결합하여 안정적이고 일반화된 군집화를 얻을 수 있는 앙상블 학습을 적용해볼 수 있습니다. 이러한 방법들을 조합하여 주관성을 극복하고 더 신뢰성 있는 군집화 결과를 얻을 수 있습니다.
질문 내용	1. 데이터셋의 차원을 축소시키고 나서 이 작업을 되돌리는 것이 가능한지 궁금합니다. 2. 어떤 데이터셋에 적용한 차원 축소 알고리즘 성능은 어떻게 평가할 수 있는지 궁금합니다. 평가할 수 있는 모델같은 것들이 있는지 아니면 지표같은 것이 있는 것일까요?
기타	없습니다

학과	컴퓨터 전자시스템공학	학번	201904458	이름	이준용
구분	내용				
학습 범위	기계학습 6장 비지도 학습 6.5절 공간 변환의 중요성				

	6.6절 PCA, ICA, sparse 코딩 6.7절 AUTO ENCODER 6.8 절 매니폴드 개념과 isomap, LLE, T-SNE 매니폴드 학습기법
학습 내용	**기계학습 6장 비지도 학습** **6.5절 공간 변환의 중요성**  그림 6-16 공간 변환의 예 실제 문제에서는 비지도 학습을 통해 최적의 공간 변환을 자동으로 알아내야 함 원래 공간을 다른 공간으로 변환하는 인코딩, 변환 공간을 원래 공간으로 역변환 하는 디코딩. 데이터 압축의 경우, 역변환으로 얻은 X'는 원래 신호 X와 가급적 같아야 함 데이터 가시화에서는 2차원 또는 3차원의 공간으로 변환. 디코딩 불필요 선형 인자 모델 - 선형 연산을 이용한 공간 변환 기법 선형 연산을 사용하므로 행렬 곱으로 인코딩(f), 디코딩(g) 과정을 표현 $f: \mathbf{z} = \mathbf{W}_{enc}\mathbf{x} + \boldsymbol{\alpha}_{enc}$ $g: \mathbf{x} = \mathbf{W}_{dec}\mathbf{z} + \boldsymbol{\alpha}_{dec}$ A는 데이터를 원점으로 이동하거나 잡음을 추가하는 등의 역할 Z에 확률 개념이 없고, α를 생략하면 주성분 분석 주성분 분석 (PCA : Principal Component Analysis) 데이터를 원점 중심으로 옮기는 전처리를 먼저 수행 : $X_i = X_i - \mu$ (μ: 평균) 변환 행렬 W는 d * q로서 주성분 분석은 d차원의 x를 q차원의 z로 변환(q < d)

▪ 예, 2차원을 1차원으로 변환하는 상황($d = 2, q = 1$)



주성분 분석의 목적
손실을 최소화하면서 저차원으로 변환하는 것 - 변환된 훈련집합의 분산이 클수록 정보 손실이 적다고 판단.

디코딩 과정

역변환은 $x = (W^T)^{-1}z$ 인데, W가 정규직교 행렬이므로 X' = Wz 가 됨.
q = d로 설정하면 W가 d*d이고 X'는 원래 샘플 X와 같게 됨 - 원래 공간을 단지 일정한 양만큼 회전하는 것에 불과

실제로는 q<d로 설정하여 차원 축소를 꾀함

1. 데이터 압축
2. q = 2 또는 q = 3으로 설정하여 2차원 또는 3차원으로 축소하여 데이터 가시화
   - 고유얼굴 기법: 256*256 얼굴 영상(d = 65536)을 7차원(q =7) 으로 변환하여 얼굴 인식(정면 얼굴에 대해 96% 정확률) -> 상위 몇 개의 고유벡터가 대부분의 정보를 가짐

6.6절 PCA, ICA, sparse 코딩
독립 성분 분석 (ICA : Independent Component Analysis)

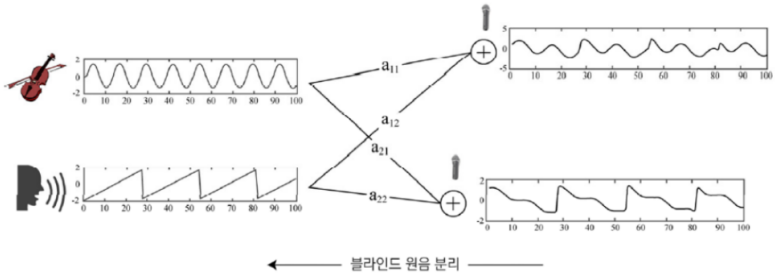


그림 6-21 블라인드 원음 분리 문제

마이크로 측정한 혼합 신호로부터 원음(음악과 목소리)를 복원할 수 있나? -> 블라인드 원음 분리 문제라 부르며, 독립 성분 분석 기법으로 해결 가능

혼합 신호 x를 원래 신호 z의 선형 결합으로 표현 가능($Z_1(t)$와 $Z_2(t)$가 독립이라는 가정)

$x_1 = a_{11}z_1 + a_{12}z_2,\ x_2 = a_{21}z_1 + a_{22}z_2$

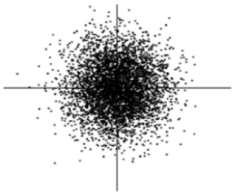
행렬 표기로 쓰면 x = AZ

블라인드 원음 분리 문제란 A를 구하는 것. A를 알면, z` = Wx, 이때 W= A^-1 를 이용해 원음 복원

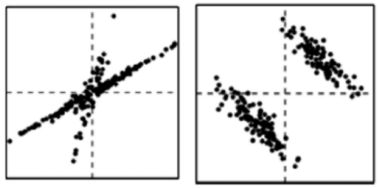
정수 하나를 주고 어떤 두 수의 곱인지 알아내라는 문제와 비슷함 -> 추가 조건을 주면 유일해가 가능. 독립성 가정과 비가우시안 가정을 이용하여 x=Az의 해를 찾음.

독립성 가정 -> 원래 신호가 서로 독립이라는 가정 (음악과 대화는 서로 무관하게 발생함)

비가우시안 가정 - > 원래 신호가 가우시안이라면 혼합 신호도 가우시안이 되므로 분리할 실마리 없음. 비가우시안이면 실마리가 있음.



(a) 확률변수가 가우시안일 때



(b) 확률변수가 비가우시안일 때

그림 6-22 서로 독립인 두 확률변수의 결합 분포

ICA의 문제 풀이 -> 원래 신호의 비가우시안인 정도를 최대화하는 가중치를 구하는 전략 사용

ICA와 PCA 비교

ICA

1. 비가우시안과 독립성 가정
2. 주로 블라인드 원음 분리 문제 해결
3. 4차 모멘트까지 사용
4. ICA로 찾은 축은 수직 아님

PCA

1. 가우시안과 비상관 가정
2. 주로 차원 축소 문제 해결
3. 2차 모멘트까지 사용
4. PCA로 찾은 축은 서로 수직

희소 코딩 - 기저함수 또는 기저벡터의 선형 결합으로 신호를 표현
푸리에 변환 또는 웨이블릿 변환 등 희소 코딩이 다른 변환 기법과 다른 점
비지도 학습이 사전(기저벡터)를 자동으로 알아냄 (푸리에 변환은 삼각함수를 사용함)
➔ 희소 코딩은 데이터에 맞는 기저 벡터를 사용하는 셈
사전의 크기를 과잉 완벽하게 책정 (m > d)

희소 코드 a를 구성하는 요소 대부분이 0값을 가짐
희소 코딩 구현 - 최적의 사전과 최적의 희소 코드를 알아내야 함

$$\hat{\mathbf{D}}, \hat{\mathbf{A}} = \underset{D,A}{\operatorname{argmin}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda\phi(\mathbf{a}_i)$$

, ϕ는 희소 코드의 희소성을 강제하는 규제항

6.7절 AUTO ENCODER

오토인코더 -> 특징 벡터 x를 입력받아 동일한 또는 유사한 벡터 x'를 출력하는 신경망
단순 복사하는 단위 행렬은 무의미

병목 구조 오토인코더의 동작 원리

m < d 인 구조 (ex, 256*256 영상을 입력 받아 256*256 영상을 출력하는 경우 d=65536인데 m=1024로 설정)
은닉층의 h는 훨씬 적은 메모리로 데이터 표현, 필요한 경우, 디코더로 원래 데이터 복원
h는 x의 핵심 정보를 표현 -> 특징 추출, 영상 압축 등의 응용

여러 형태의 오토인코더

은닉 노드 개수에 따라 m<d, m=d, m>d 구조 / 활성함수에 따라 선형과 비선형 구조

오토인코더의 학습

주어진 데이터는 훈련집합 X = {x1,x2,…xn} 알아내야 하는 매개변수는 f와 g라는 매핑 함수 즉 가중치집합 θ = {W, V}

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} L\left(\mathbf{x}_i, g(f(\mathbf{x}_i))\right)$$

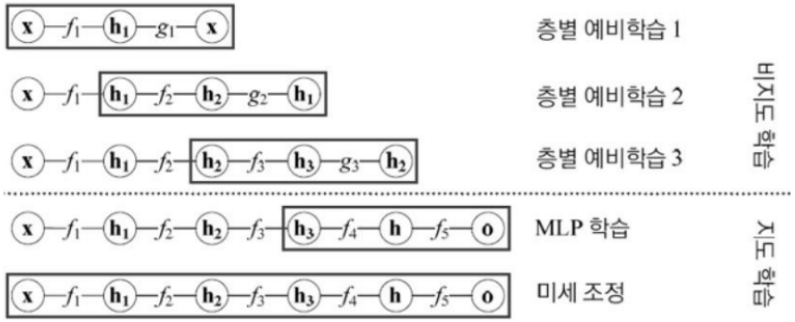
$$L\left(\mathbf{x}_i, g(f(\mathbf{x}_i))\right) = \|\mathbf{x}_i - g(f(\mathbf{x}_i))\|_2^2$$

규제 오토인코더

여러 규제 기법을 적용 - m>d인 상황에서도 단순 복사를 피할 수 있음
SAE (sparse autoencoder) - 은닉 벡터 hi가 희소하도록 강제화(0이 아닌 요소의 개수를 적게 유지
DAE (denoising autoencoder) - 잡음을 추가한 다음 원본을 복원하도록 학습하는 원리
CAE (contractive autoencoder) - 인코더함수 f의 야코비안 행렬의 프로베니우스 놈을 작게 유지CAE는 공간을 축소하는 효과

적층 오토인코더

은닉층이 하나인 경우 표현력에 한계가 있다. -> 여러 층으로 쌓으면 용량이 커짐
층별 예비학습을 이용하여 깊은 신경망을 만듦



적층 오토인코더를 지도학습(분류)에 활용하는 경우의 학습 과정 ->
층별 예비학습을 필요한 만큼 수행 ->마지막 층의 출력을
입력으로 하여 MLP 를 학습한다. ->신경망 전체를 한꺼번에 추가로 학습한다.

매니폴드 - >고차원 공간에 내재한 저차원 공간
도로가 매니폴드에 해당
자동차 위치를 3차원 데이터로 나타낼 수 있으나, 기준점에서의 거리 라는 1차원(저차원) 공간, 즉 매니폴드로 표현할 수 있음.

보통 매니폴드는 비선형 공간이지만 지역적으로 살피면 선형 구조
매니폴드 가정 -> 고차원 공간에 주어진 실제 세계의 데이터는 고차원 입력 공간 R
d에 내재한 훨씬 저차원인 dM차원 매니폴드의 인근에 집중되어 있다.

6.8절 매니폴드 개념과 isomap, LLE, T-SNE 매니폴드 학습기법
매니폴드를 어떻게 구할까?
IsoMap = 최근접 이웃 그래프 구축
1. 각 점은 k-최근접 이웃을 구하여 거리를 n*n행렬 M에 채움
2. 빈 곳은 최단 경로의 shortest path 길이로 채움
M의 고유 벡터를 계산하고, 큰 순서대로 $dlow$개의 고유 벡터를 선택
- 이들 고유 벡터가 새로운 저차원 공간 형성

- i번째 샘플의 k번째 좌표는 $\sqrt{\lambda_k v_k^i}$. M이 너무 크다는 문제점

LLE (locally linear embedding)
거리 행렬 M대신에 함수 ϵ를 최소로 하는 가중치 행렬 W를 사용함.

$$\epsilon(\mathbf{W}) = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{\mathbf{x}_j \in \{\mathbf{x}_i\text{의 이웃}\}} w_{ij}\mathbf{x}_j \right\|_2^2$$

t-SNE (stochastic neighbor embedding)
현재 t-SNE는 매니폴드 공간 변환 기법 중에서 가장 뛰어남
원래 공간에서 유사도 측정
변환된 공간에서의 유사도는 스튜던트 t 분포로 측정
$\mathbf{y}_i$와 $\mathbf{y}_j$는 변환된 공간에서의 점

$$q_{ij} = \frac{\left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|_2^2\right)^{-1}}{\sum_{k \neq i}(1 + \|\mathbf{y}_i - \mathbf{y}_k\|_2^2)^{-1}}$$

원래 데이터와 변환된 데이터의 구조가 비슷해야 하므로, 확률 분포 P와 Q가 비슷할수록 좋음
비슷한 정도를 측정하기 위해 아래의 KL 다이버전스를 사용

$$J(\mathbb{X}') = KL(P \parallel Q) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}\log\left(\frac{p_{ij}}{q_{ij}}\right)$$

Transductive 학습 모델
훈련집합 이외의 샘플을 처리할 능력이 없는 모델
t-SNE, LLE, IsoMap 모두 Transductive 모델
데이터 가시화라는 목적에 관한 한 PCA나 오토인코더와 같은 귀납적 모델보다 성능이 뛰어남

귀납적 모델 (inductive model, buttom-up)
훈련집합 이외의 새로운 샘플을 처리할 능력이 있는 모델
t-SNE, LLE, IsoMap 를 제외한 지금까지 공부한 모든 모델

학과	컴퓨터 전자시스템공학	학번	201800615	이름	김동규
구분	내용				
학습 범위	기계학습 6장 비지도 학습 6.5절 공간 변환의 중요성 6.6절 PCA, ICA, sparse 코딩 6.7절 AUTO ENCODER 6.8 절 매니폴드 개념과 isomap, LLE, T-SNE 매니폴드 학습기법				
학습 내용	공간 변환 - 기본적으로 주어진 데이터들의 분포를 특정한 좌표 공간으로 변환해 데이터들의 추출이 가능해짐 주성분 분석 - 손실을 최소화하며 저차원으로 바꾸는 것 - 변환된 훈련집합의 분산이 클수록 정보손실이 적다고 판명함 pca 최적화 • 라그랑주 함수 • 모든 데이터에 대한 효용함수 • 모든 데이터에 대한 코스트 총합 함수 • 각 데이터에 대해서 최대한의 분산을 가지는 값 계산 각 데이터별로 편미분을 통해 최대가 되는 라그랑주 함수를 추측함 • 주성분 분석의 알고리즘 1. 훈련집합의 공분산 행렬 계산 2. 라그랑주 함수를 풀어서 고윳값, 고유벡터를 구함 3. 고윳값이 큰 순서대로 훈련집합을 정렬함 4. 이중 q개의 데이터를 선택해서 행렬에 채움 데이터의 압축 • 훈련집합을 그린 공간을 일정한 크기로 분할시킨 후, 해당 공간안의 값들을 대표값 하나로 표현시킴 (퀀타이징) • 주성분 분석을 통해 얻은 주성분 행렬로 압축 표현이 가능함 데이터들의 손해값이 적어지기에 더 적은 비트로 대표값이 추출 가능해짐 • 디코딩 원 벡터와 수직이 되도록 만듦 • pca와 ica • pca				

- 각 훈련집합은 서로 연관이 있는 경우
- 각 데이터들을 가장 잘 표현할 수 있는 벡터를 찾음

훈련집합은 서로 연관이 없음
  각 데이터들의 독립성이 최대가 되는 벡터를 찾음
non negative matrix factorization
  어떠한 결과의 행렬을 구하기 위해서, 2개의 다른 행렬로 표현함
희소 코딩
- 비지도 학습을 통해 기저 벡터를 자동으로 알아냄
- 최선의 기저 벡터, 최적의 희소코드를 알고 있으면 편해짐
- 최대한 적은 갯수의 벡터로 데이터를 표현하기 위함

인코더
  특징 벡터 x를 입력받고, 동일 또는 유사한 벡터로 다시 복원시키는 신경망

$$J(W_1, b_1, W_2, b_2) = \sum_{i=1}^{m} \left(\tilde{x}^{(i)} - x^{(i)} \right)^2$$
$$= \sum_{i=1}^{m} \left(W_2 z^{(i)} + b_2 - x^{(i)} \right)^2$$
$$= \sum_{i=1}^{m} \left(W_2 (W_1 x^{(i)} + b_1) + b_2 - x^{(i)} \right)^2$$

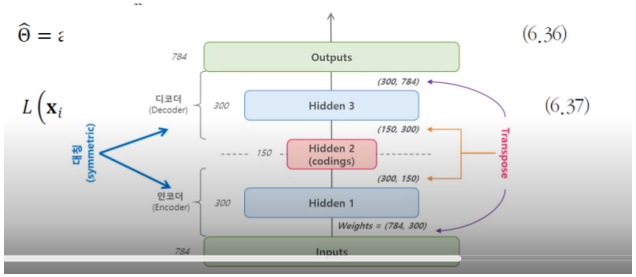
$z^{(i)} = W_1 x^{(i)} + b_1$ <- compressed data
$\tilde{x}^{(i)} = W_2 z^{(i)} + b_2$ <- recontructed/approximate data

- 입력층과 출력층에 비해 작은 은닉층의 규모

여러 층의 은닉층을 사용해 단계적으로 인코딩, 디코딩 가능함

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} L\left(\mathbf{x}_i, g\left(f(\mathbf{x}_i)\right)\right) \qquad (6.36)$$

$$L\left(\mathbf{x}_i, g\left(f(\mathbf{x}_i)\right)\right) = \|\mathbf{x}_i - g(f(\mathbf{x}_i))\|_2^2 \qquad (6.37)$$



g(f(x)) = 분산이 최소가 되는 벡터를 구하고 오차를 최소화 하는 벡터를 구함

규제 오토인코더

- 은닉층 노드 수 > 입력층 노드 수인 경우, 은닉층 = 입력층 노드수로 조절함

오토인코더

은닉층을 여러개를 쌓아서 만듦

매니폴드 공간 변환

- 고차원 공간에 포함된 저차원 공간

고차원 데이터가 존재할 경우, 해당 데이터들을 아우르는 저차원 공간이 있다고 가정한 후 학습을 진행함(스위스 롤)

T-SNE(T분포)

- xi와 xj의 유사도를 가지고 조건부 확률로 변환
- 유사도가 높을수록 확률이 올라감
- 변환된 공간은 가우시안 분포 대신 t분포로 분포를 표현함

- 각 확률 변수들은 kl 다이버전스를 목적함수로 분별됨