1. 이상 탐지 개념 및 머신러닝 기초

목차

▼ 이상탐지

1. 이상탐지 개념

   이상탐지는 데이터 마이닝을 기반으로 한 데이터 분석 기법 중의 하나로, 이상치를 탐지하는 기법입니다. 따라서 이상치, 이상 징후로 부르고, 영어로는 Anomalies, Outliers, Exceptions와 같이 표현될 수 있습니다. 대표적으로 이상탐지는 IT 보안, 의료진단, 제조공정의 모니터링 등 다양한 산업분야에 적용되고 있으며 활용분야가 점차 확대되고 있습니다.

   다음은 이상탐지를 보는 관점에 따라 이상탐지의 정의가 조금씩 달라지는 것을 알 수 있습니다.

   - **점 이상 (Point Anomaly)**
     - 데이터 내 하나의 관측 값이 나머지에 이상하다고 판단되는 경우

   - **맥락적 이상 (Contextual Anomaly)**
     - 시간의 특성을 가지는 Time-Series 분야에서 많이 나타납니다.
     - 시계열 자료에서는 시간의 연속성이 존재하여 특정 시점이 그 시점 전, 후의 값에 크게 영향을 받습니다.
     - 시계열 자료에서 비정상적인 시점을 찾는 것을 목표로 할지, 비정상적인 변화의 패턴을 찾는 것을 목표로 할지에 따라 분류합니다.

2. 이상 탐지의 적용 사례

- **Cyber-Intrusion Detection**

  컴퓨터 시스템 상에 침입을 탐지하는 사례. 주로 시계열 데이터를 다루며 RAM, file system, log file 등 일련의 시계열 데이터에 대해 이상치를 검출하여 침입을 탐지합니다.

- **Fraud Detection**

  보험, 신용, 금융 관련 데이터에서 불법 행위를 검출하는 사례. Kaggle Credit Card Fraud Detection 과 같은 공개된 Challenge도 진행되었습니다.

- **Malware Detection**

  Malware(악성코드)를 검출해내는 사례. Classification과 Clustering이 주로 사용되며 Malware 데이터를 그대로 이용하기도 하고, 이를 Gray Scale Image로 변환하여 이용하기도 합니다.

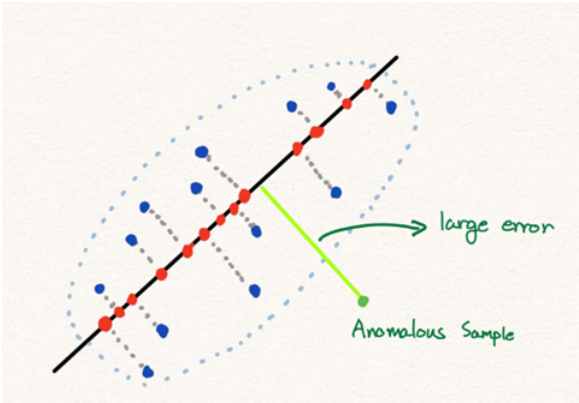
- **Medical Anomaly Detection**

  의료 영상, 뇌파 기록 등의 의학 데이터에 대한 이상치 탐지 사례. 주로 신호 데이터와 이미지 데이터를 다루며 X-ray, CT, MRI, PET 등 다양한 장비로부터 취득된 이미지를 다루고 있습니다.

1. 비지도 학습에 기반한 이상탐지 활용 사례

   - **PCA**

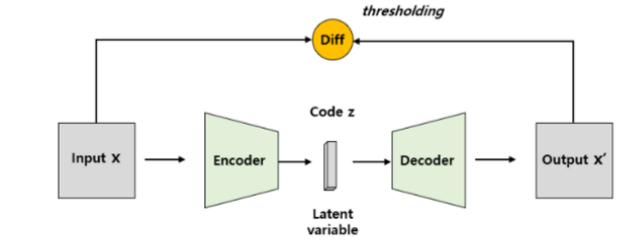
     PCA에 의해 선택된 축과 원래의 샘플 위치를 비교하여 거리가 먼 샘플은 비정상이라고 판별합니다.

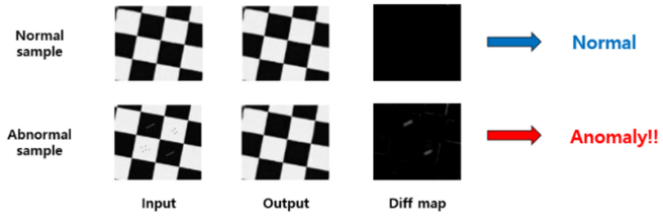


   - **Autoencoder**

복원 오차는 이상 점수 (Anomaly score) 가 되어 threshold와 비교를 통해 이상 여부를 결정합니다.

- threshold 보다 클 경우 이상
- threshold 보다 작을 경우 정상





▼ 지도 학습

먼저 지도 학습이 무엇인지에 대해 간단히 살펴볼까요?😀

지도 학습이란 훈련 데이터로부터 하나의 함수를 유추해내기 위한 기계 학습의 한 방법으로,

주어진 데이터에 대해 예측하고자 하는 값을 올바르게 추출하기 위한 목적을 가지고 있습니다.

지도 학습은 유추된 함수가 어떤 값을 띄는지에 따라 2개로 분류가 가능합니다.

1) 회귀 분석: 연속적인 값을 출력합니다.

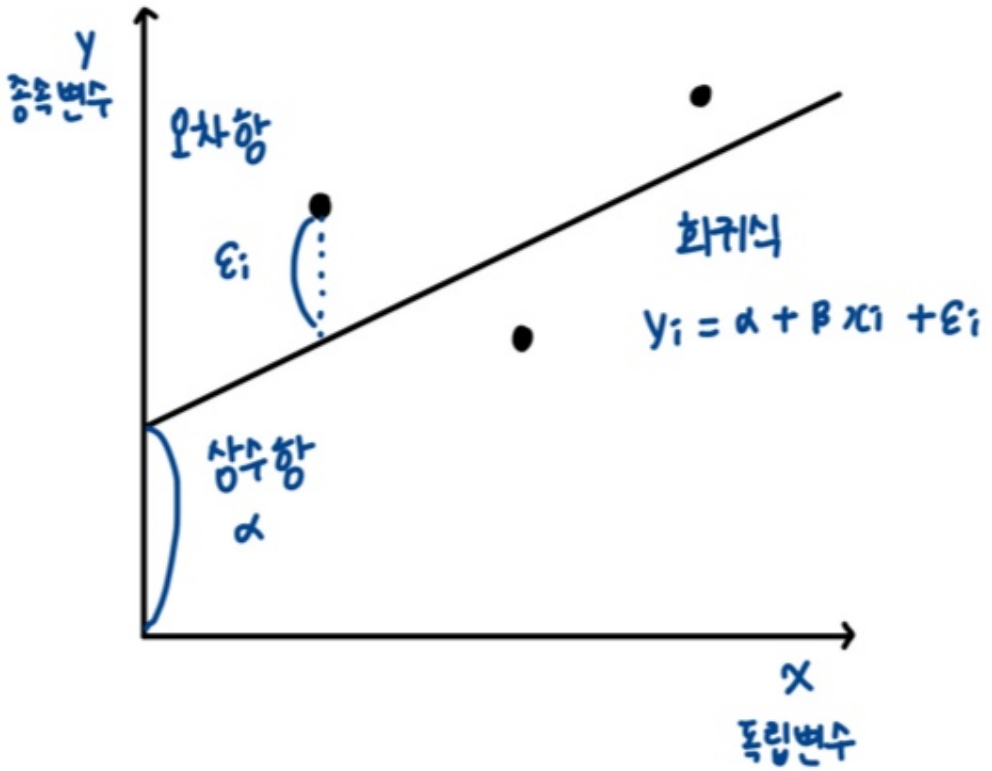
2) 분류: 비연속적인 값으로 ,어떤 종류의 값인지 출력합니다.

지도 학습을 이용한 알고리즘은 매우 다양합니다. **최근접 이웃(k-NN), 선형 모델,랜덤 포레스트**
등이 있으며 저희는 크게 선형 방법, 이웃 기반 방법, 트리 기반 방법 3가지의 기준에 따라 알고리즘을 분류해보았습니다.🤭

1.선형 방법

선형 회귀 모델이란 파라미터가 선형식으로 표현이 되는 모델을 뜻합니다. 선형 회귀 분석, 로지스틱 회귀 분석과 같은 모델이 선형 회귀 모델에 속합니다.

01)선형 회귀 분석

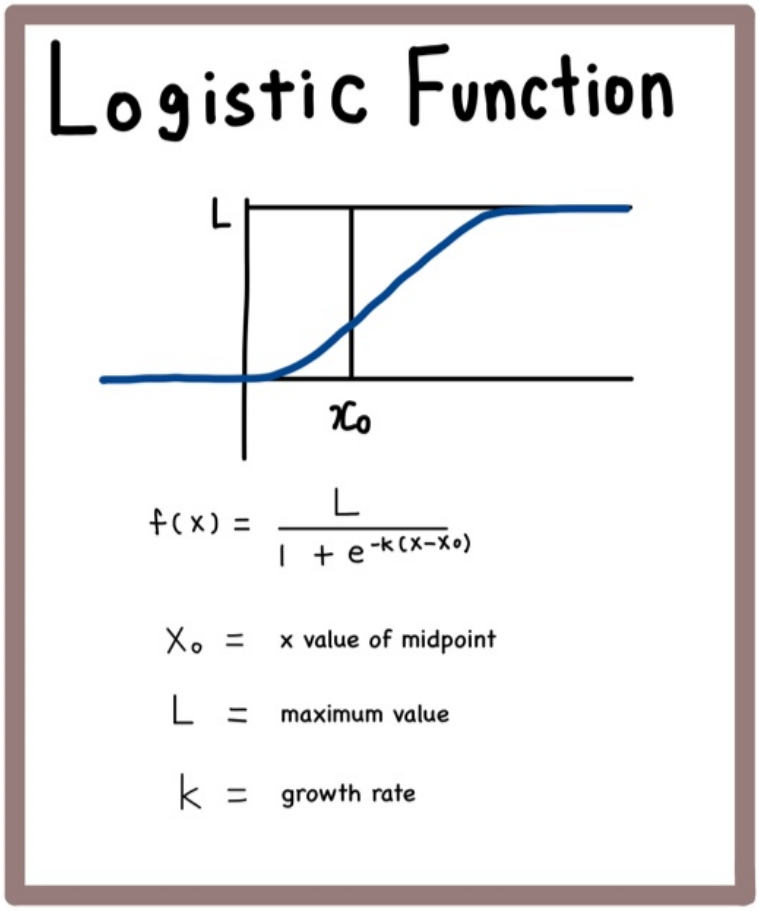


선형 회귀는 입력 변수(x)와 단일 출력 변수(y) 사이의 선형 관계를 가정하는 모델을 사용하는 가장 간단한 알고리즘으로 수치 예측 문제 (추론 문제 ,예측 문제)에 활용됩니다.

선형 회귀 분석은 독립변수의 개수에 따라 2가지로 분류되는데요. 독립변수가 1개로 이루어질 경우 단순선형 회귀 모델로 불려지며, 독립변수들로 이루어진 행렬과 종속변수가 주어졌을 때 다중선형 회귀 모델이라고 볼 수 있습니다 !

선형 회귀 분석의 장점으로는 지나치게 복잡한 관계를 모형화할 수 없기 때문에 간단하고 이해하기 쉬우며 과대적합되기 어렵다는 점입니다.그러나 입력 변수와 출력 변수간의 관계가 선형이 아닌 비선형일때 데이터에 과소적합할 위험이 존재합니다.

02)로지스틱 회귀 분석



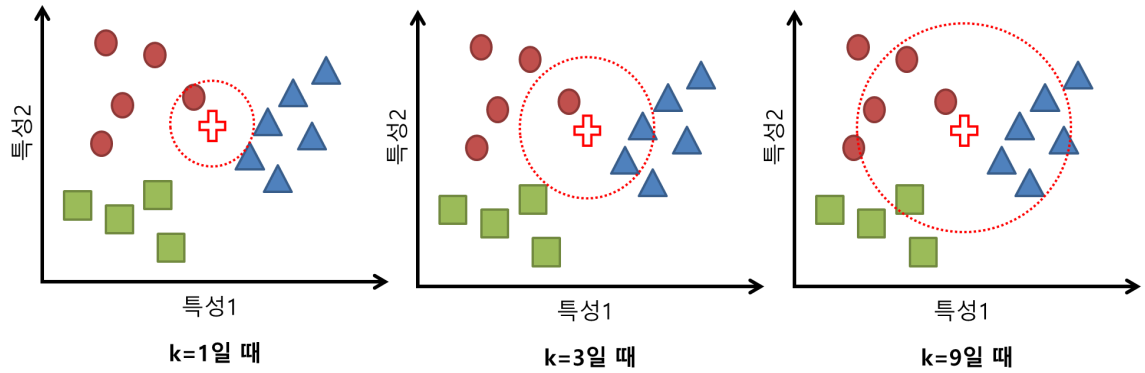
로지스틱 회귀 분석은 간단한 분류 알고리즘으로 선형 방법이지만, 종속 변수는 로지스틱 함수에 의해 반환되며 이 변환을 통해 클래스 확률을 출력합니다.

Y가 범주형 변수일때 선형 회귀 모델을 그대로 적용하게 될 경우 숫자는 아무 의미를 지니고 있지 않기 때문에 적용한 의미가 없습니다. 이때 , 로지스틱을 사용합니다.

즉, 로지스틱 회귀 분석은 종속 변수가 연속형이 아닌 범주형 데이터로 주어져 해당 데이터의 결과가 특정 분류로 나뉘게 될때 쓰입니다. 따라서 Logistic Regression은 회귀분석이지만 분류성격을 갖고 있습니다.

2.이웃 기반 방법

다음으로는 이웃 기반 방법입니다. 대표적인 알고리즘은 KNN으로 자세히 짚어보고 넘어가겠습니다.



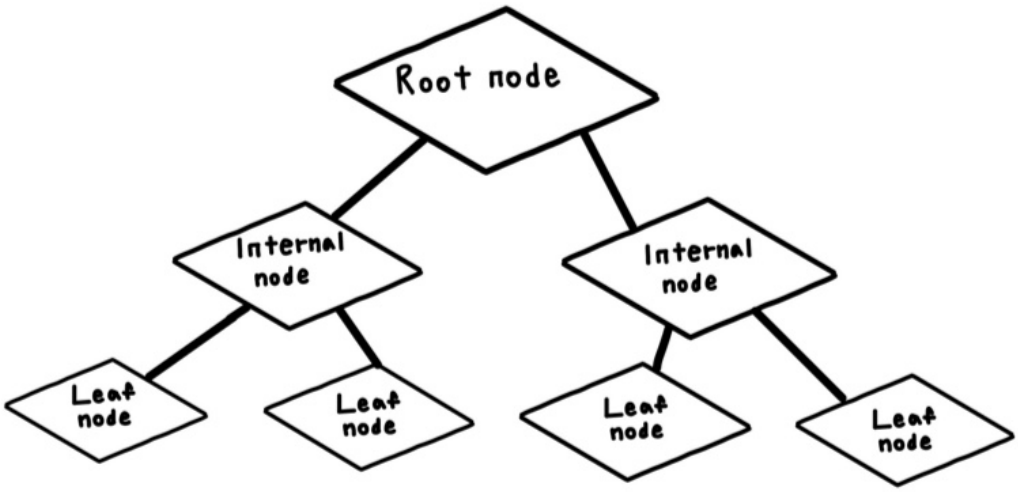
KNN은 비슷한 특성이나 속성을 가진 것들끼리 가깝게 모여있는 이웃의 속성에 따라 k개씩 분류하여 레이블링을 하는 알고리즘입니다.

각각의 새로운 데이터 포인트에 레이블을 지정하기 위해 K개 (K는 정수)의 가장 가까운 레이
블이 지정된 데이터 포인트를 보고 이미 레이블이 지정된 이웃들에게 새로운 데이터 포인트에 레이블을 지정하는 방법을 투표하게 합니다.

또한 KNN은 유클리드 거리와 맨해튼 거리를 사용합니다. KNN은 K값의 선택이 매우 중요한데 매우 작은 값으로 설정될 경우 과대적합의 위험이, 매우 큰 값으로 설정될 경우 과소적합의 위험이 있어 K값을 적당하게 선정해야 합니다.

3.트리 기반 방법

01)단일 의사 결정 트리

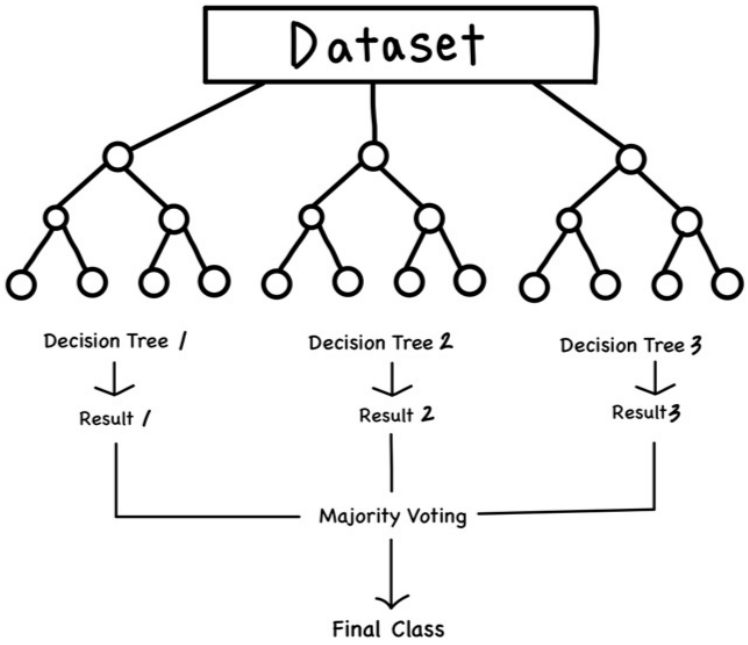


AI가 훈련 데이터를 한번 통과한 후 레이블에 의해 데이터를 분할하는 규칙을 만들고, 만들어진 트리를사용해 새로운 검증 또는 테스트 데이터셋을 예측하는 방법입니다.

최대한 균일한 데이터 세트를 구성할 수 있도록 분할 하는 것이 중요하며 분류 및 회귀 모델을 구축하는데 사용됩니다. 단일 의사 결정 트리는 쉬운 시각화로 가독성이 높으며, 변수의 정규화가 필요없고 X Y의 인과관계와 종속변수간 영향력 파악이 쉽다는 장점이 존재합니다.

그러나 알고리즘이 심플한 만큼 예측력이 떨어지고, 과대적합이 발생하여 일반화 성능이 저해됩니다. 즉 새 데이터가 적용될 경우 낮은 예측력이 나오며 이를 개선 시키기 위한 알고리즘은 랜덤 포레스트 입니다.

02)랜덤포레스트

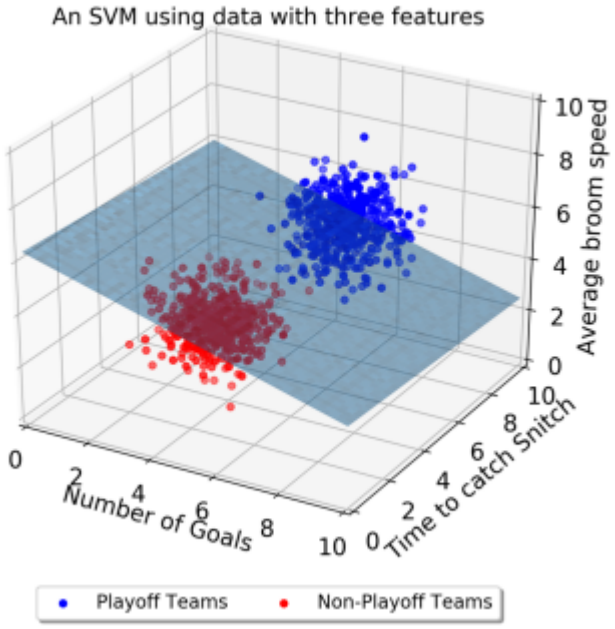


랜덤 포레스트 알고리즘은 단일 의사 결정 트리의 분류보다 정확도를 개선시키기 위해, 여러 개의 나무(단일 의사 결정 트리 모델)를 생성한 후 각각의 예측을 조합하여 결론을 내는 구조입니다.

다수의 나무들로부터 분류를 집계하기 때문에 과대적합이 나타나는 나무의 영향력을 줄일 수 있습니다. (예측 모델의 일반화 성능이 향상됩니다.) 또한 단일 의사 결정 트리에 비해 과대적합이 잘 되지 않습니다.

그러나 트리를 많이 생성하다 보니 개별 트리 분석이 어렵고 트리 분리가 복잡해집니다. 또한 차원이 크고 희소한 데이터에 대해서는 성능이 미흡합니다.

3.SVM



데이터를 분리하기 위해 트리를 만드는 대신 알고리즘을 사용해 데이터를 분리하는 공간에 초평면 (hyperplane)을 만들어 레이블에 의해 데이터를 분리하는 방법입니다.

[✋잠깐, '초평면'이란?✋

어떤 N차원 공간에서 한차원 낮은 N-1차원의 subspace를 말합니다.

3차원에서는 면이 초평면이며, 2차원에서는 선이 초평면입니다.]

딥러닝 이전 뛰어난 성능으로 많은 주목을 받았던 서포트 벡터 머신으로 분류되지 않은 새로운 점이 나타나면 경계의 어느 쪽에 속하는지 확인해서 분류 과제를 수행합니다. 결정경계는 클래스간 경계가 균일하면서 클래스 내에서 거리가 먼 경우 선택합니다.

오류 데이터의 영향이 적고, 과적합 되는 경우가 적다는 장점이 있지만 최적의 모델을 찾기 위해서 커널과 모델에서 다양한 테스트가 필요하고, 학습 속도가 느리며 해석이 어렵기도 합니다.

4.지도 학습의 이상 탐지 기법

지도 학습은 비지도 학습과 다르게 이상 탐지 기법에서 실무에 적용되기 어려운 기법입니다.😢

훈련 데이터의 모든 개체에 라벨링이 되어 있을시 쓰는 방법인데, 데이터는 정상에 비해 이상의

비율이 적은 불균형한 상태에 있으며 정확한 분류가 어렵습니다.

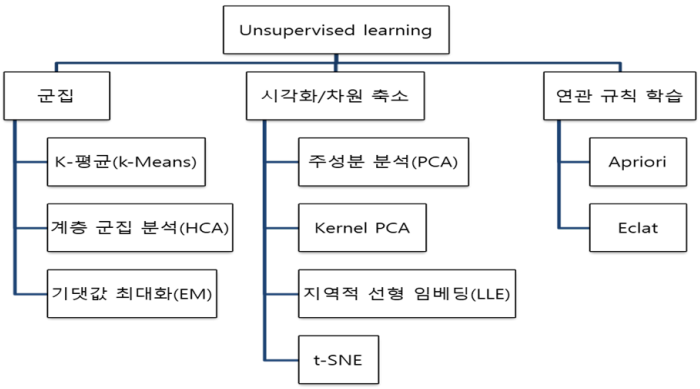
또한 정황상 이상한 것이 맞다고 판단하여도 판별할 관측데이터가 없는 경우 결과적으로 알아낼 수 없습니다. **따라서 지도 학습은 실무에서 적용하기 힘든 기법입니다.**

▼ 비지도 학습

1. 비지도 학습의 정의

   비지도 학습은 정답 라벨이 없는 데이터를 비슷한 특징끼리 군집화하여 새로운 데이터에 대한 결과를 예측하는 것을 말합니다. 정답 라벨이 없기 때문에 엄격하게 정의된 레이블이 없으므로 더 흥미로운 패턴을 발견할 수도 있는 장점이 있습니다.

   그럼 비지도 학습에 대한 모델링 기법은 어떤 방법들이 있는지 공부해 볼까요?🙂

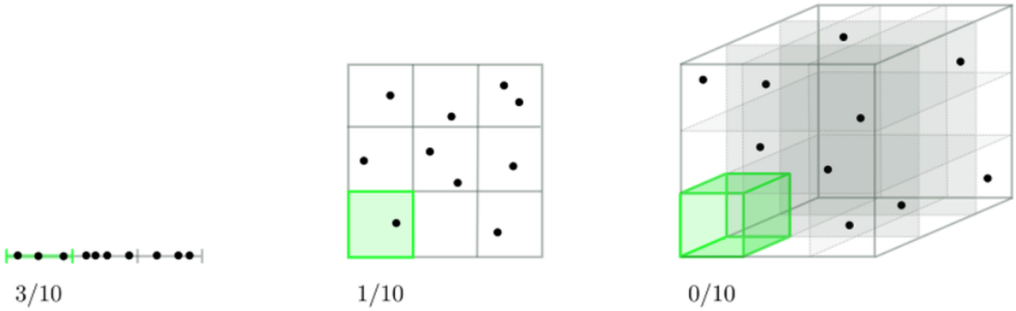


2. 차원의 저주

   모델링에 대해서 공부하기 전, 우리는 차원의 저주를 공부해 볼 필요가 있어요!

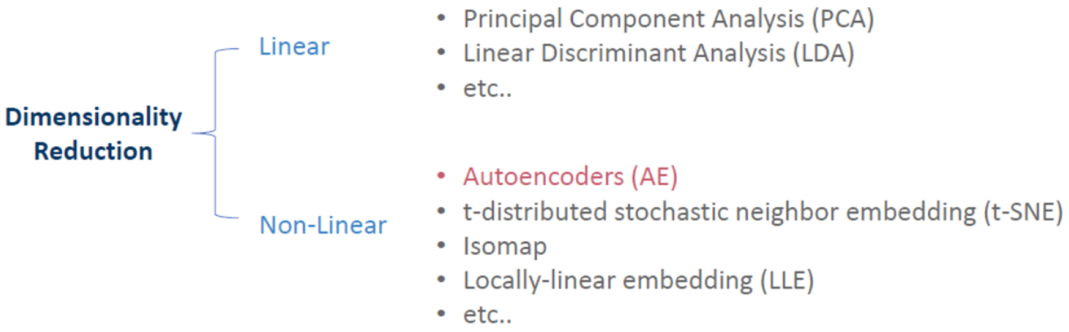
   아래의 그림을 보시면 feature의 수가 많아지게 되면 dimension의 수도 증가하기 때문에 공간의 부피가 기하급수적으로 증가하는 것을 볼 수가 있어요. 그렇게 되면 빈 공간이 많이 생기기 때문에 모델이 학습을 하는데 불완정해질 수도 있고, overfitting의 위험도 있어요. 그럼 어떻게 이 문제를 해결할 수 있을까요? 🧐

우리는 차원 축소를 통해 이 문제를 해결하려고 합니다.



3. 차원 축소

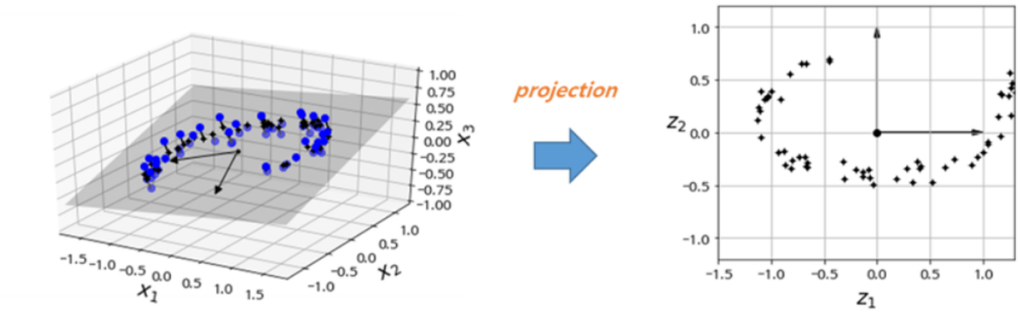
차원 축소에서도 선형과 비선형으로 구분되는데 다음 그림을 보면 여러 가지의 모델링 방법들이 있는 것을 확인할 수 있어요. :)



그럼 선형과 비선형은 어떻게 정의되고, 차원 축소와 관련된 모델링 방법들에 대한 이론에 대해서 간단히 설명해 보도록 할게요!

3-1. 차원 축소 (선형)

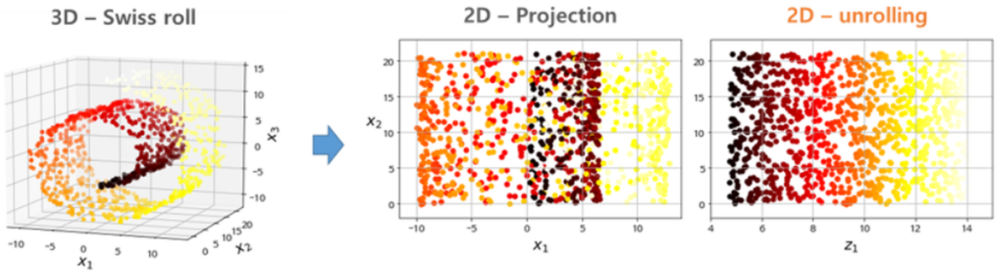
아래 왼쪽 그림을 보시면 3차원 공간에 존재하는 데이터들이 어떤 평면상(부분공간)에 투영되어서 오른쪽 그림과 같이 2차원 공간으로 축소되는 모습을 선형 차원 축소라고 합니다.



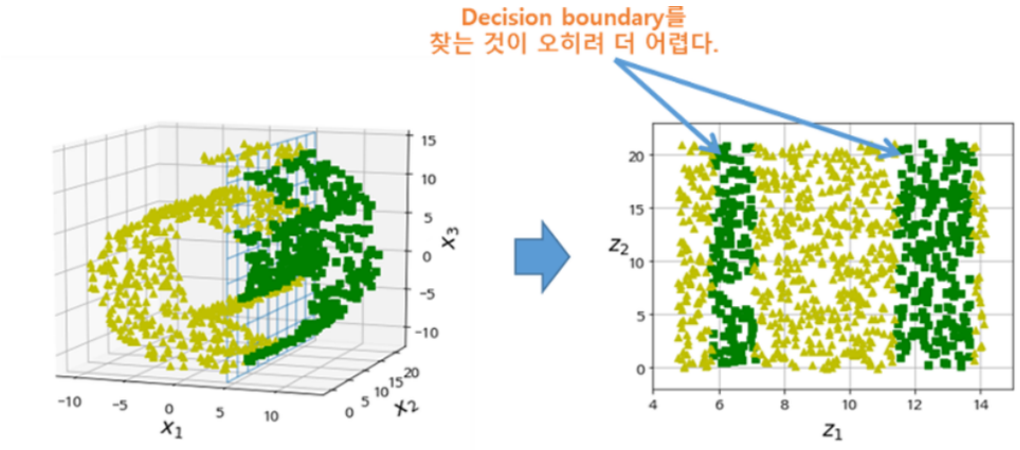
3-2. 차원 축소 (비선형)

그럼 비선형은 어떻게 차원 축소를 할까요?

비선형 차원 축소는 매니폴드 학습(Manifold Learning)으로 불리기도 하는데, 아래 왼쪽 그림을 보시는 것처럼 말려져 있는 데이터를 하나의 평면으로 본다고 가정하는 거예요! 그래서 오른쪽 그림을 보시면 선형과 비선형으로 차원 축소한 모습을 나타냈는데, 비선형으로 차원 축소한 모습이 더 잘 분류한 것을 알 수 있어요 🙂



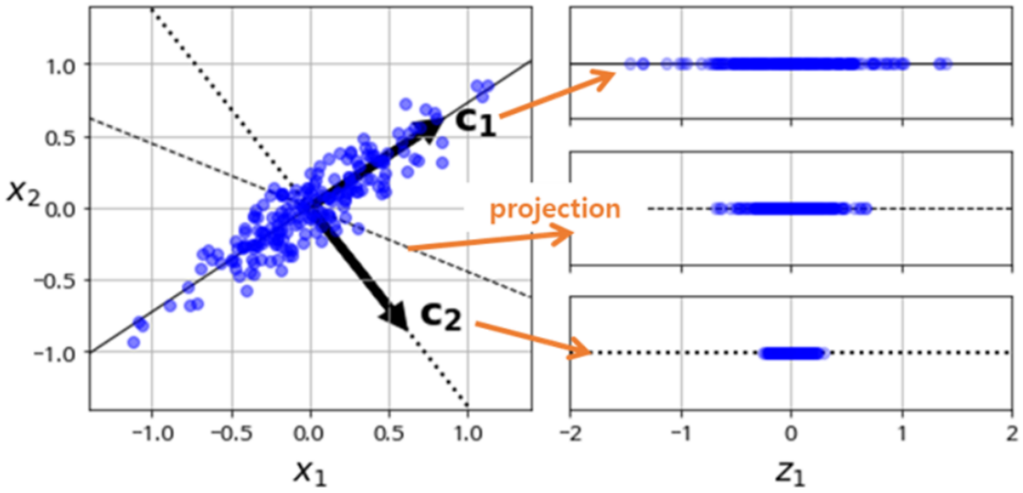
하지만 비선형 차원 축소도 단점은 분명히 존재하는데 고차원에서 Decision boundary를 찾는 것이 더 편리할 수 있기 때문에, 데이터를 확인하면서 선형으로 축소할 것인지, 비선형으로 축소할 것인지 결정해야 돼요 :)



4. 비지도 학습 모델 (차원 축소)

4-1. Principal Component Analysis (PCA, 선형)

데이터의 변동성을 설명할 때 전체 feature 중 어떤 feature가 가장 중요한지 파악하는 것이 중요한데, 이때 PCA 알고리즘을 이용할 수 있어요! PCA 알고리즘이란 최대한 다양성을 유지하면서 데이터의 저차원 표현을 찾는 방법이에요. 아래 그림을 보시는 것처럼 x1과 x2의 관계 중에서 저차원으로 이동시킬 때, 데이터를 잘 보존하기 위해서 분산을 최대로 하는 c1 축을 찾는 것이 PCA의 특징이에요.



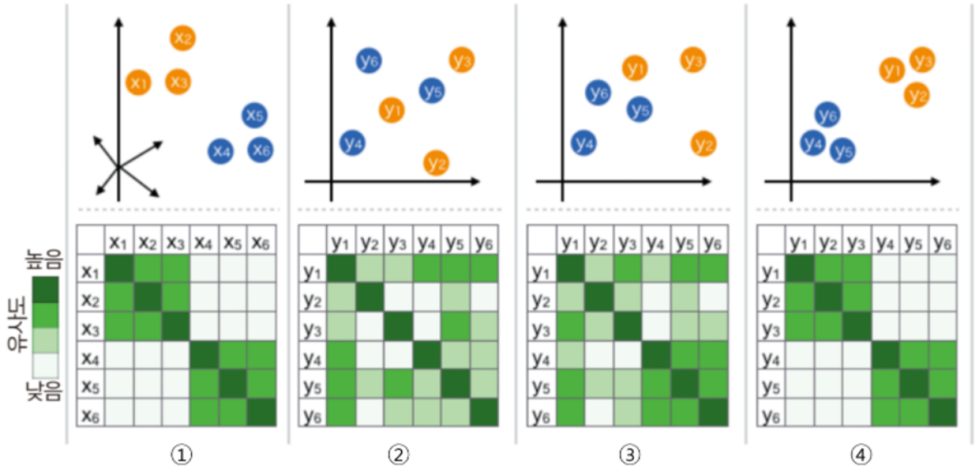
기존 PCA 방법 외에도 여러 가지로 변형된 Incremental PCA, Kernel PCA, Sparse PCA등이 있답니다!

4-2. Singular Value Decomposition (SVD, 선형)

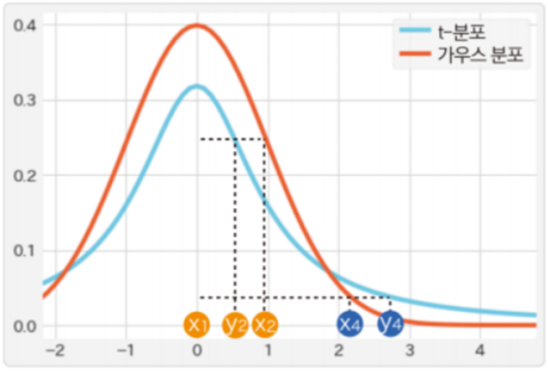
원래 행렬의 차원을 적은 차원으로 줄여 더 작은 차원의 행렬에서 일부 벡터의 선형 결합을 사용해 원래 행렬을 다시 만들 수 있도록 합니다.

4-3. t-distributed stochastic neighbor embedding (t-SNE, 비선형)

t-SNE의 기본적인 원리는 고차원의 데이터간 거리를 저차원으로 축소하였을 때,똑같이 유지하는 것입니다. 따라서 원 공간에서 가까운 점들도 고려하고, 좀 더 멀리 있는 점들의 위치도 고려하는 방법입니다. 다음 그림을 보시면 이해하기 쉬울 거예요. 🙂



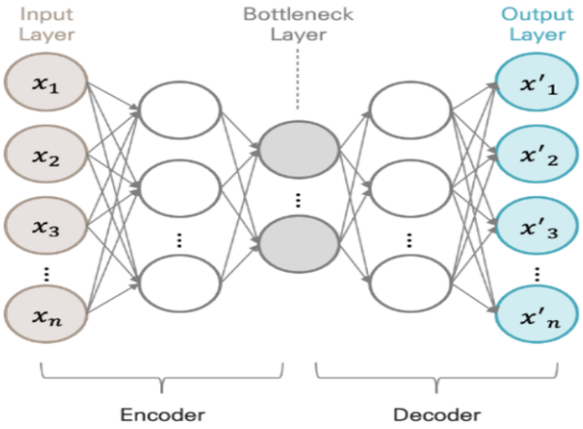
1번 그림을 보시면 3차원의 데이터가 존재하는데 이 데이터를 2차원으로 축소해서 최종적으로 4번의 그림처럼 각 데이터의 거리를 유지시키는 방법이에요. 1번 그림에서 각 포인트마다 기준을 정하고, 정규분포를 이용하여 유사도를 시각화해서 나타냅니다. 2번 그림에서는 원래의 데이터들을 랜덤으로 2차원 공간에 배치시켜서 t 분포의 특성(꼬리 부분이 두꺼움)을 이용하여 나타냅니다. 이 과정을 반복해서 최종적으로 4번 그림과 같이 저차원의 데이터로 이동시켜서 거리를 유지해 주도록 합니다. 아래 그림처럼 t 분포를 이용하여 고차원에서 유사도(y축)가 높은 관계면 데이터 포인트를 더 가까이 배치시키고, 낮은 관계면 더 멀리 배치할 수 있도록 합니다.



4-4. Autoencoder (비선형)

오토인코더는 비지도 학습 기반의 대표적인 방법입니다. 입력 샘플을 인코더를 통해 저차원으로 압축 시키고, 디코더 과정을 거쳐서 다시 원래의 차원으로 복원하는 방법이에요. 오터인코더의 특징 중 하나는 디코더를 통해 원래의 차원으로

복원하는 방법인데, 이 방법을 통해서 입력 샘플과 복원 샘플의 복원 오차 (reconstruction error)를 산출할 수 있어요. 😃



5. 비지도 학습 모델 (클러스터링)

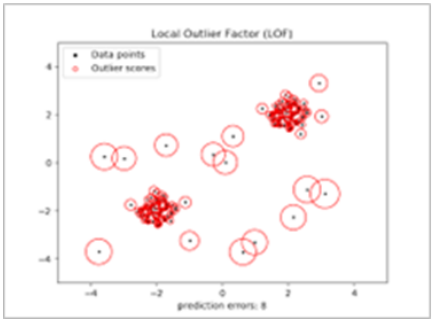
5-1. LOF (Local Outlier Factor)

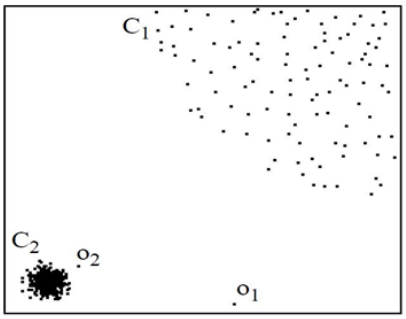
local 이상값을 찾기 위해 설계된 밀도 기반 방법입니다.

1) 각 데이터 포인트에 대해 NN이 계산

2) 계산된 이웃을 사용하여 로컬 밀도 계산 (LRD)

3) 데이터 포인트의 LRD와 이전에 계산된 NN의 LRD 비교하여 LOF 계산해서 다른 개체보다 밀도가 낮게 특정 되는 데이터를 이상치로 판단합니다.



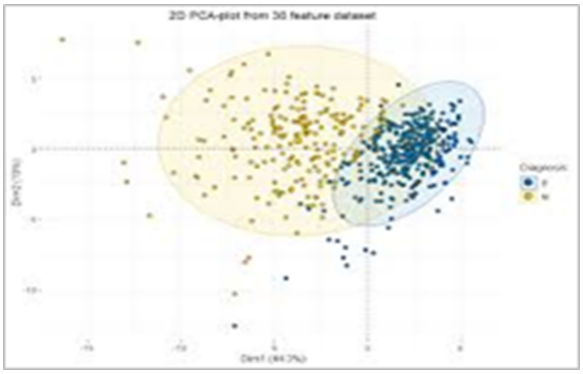


5-2. K-means

데이터를 특성에 따라 k개의 그룹으로 클러스터링하는 알고리즘입니다. 클러스터의 중심에서 멀리 떨어진 데이터는 비정상 데이터로 판단합니다.

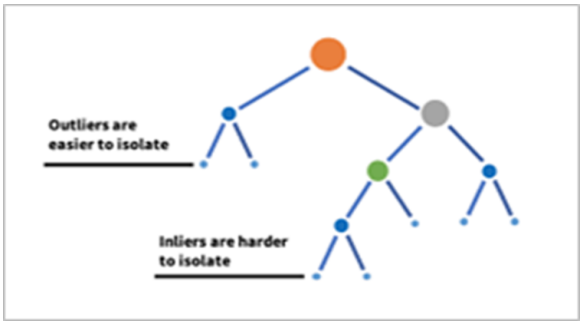
5-3. rPCA (Robust Principal Component Analysis)

차운을 축소하고 복원을 하는 과정을 통해 비정상 sample을 검출하는 방법입니다. 데이터를 더 낮은 차원의 부분공간으로 보내면 그 공간에서는 정상과 이상이 구분된다고 가정합니다.



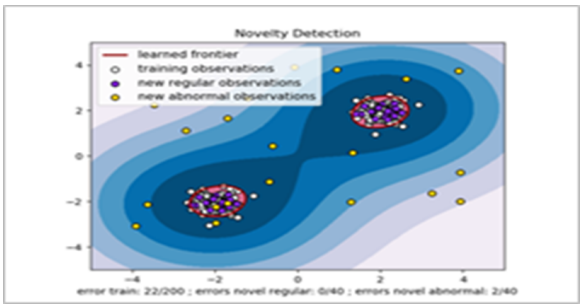
5-4. Isolation Forest

비정상 데이터는 의사결정나무의 루트에서 가까운 깊이에 고립 될 것이라고 가정합니다. 따라서 leaf 노드 까지의 거리를 outlier score로 정의하고, 평균 거리가 짧을 수록 outlier일 가능성이 크다고 판단합니다.



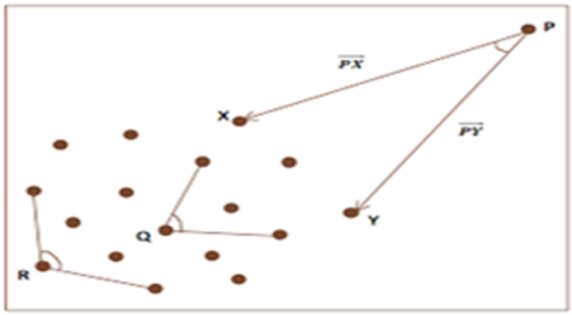
5-5. One Class SVM

정상 데이터를 원점으로부터 최대한 멀리 위치하게 만드는 초평면(Hyperplane)을 찾고, 새로운 데이터가 원점과 초평면 사이에 위치하면 이상치 데이터로 판별하는 알고리즘입니다.



5-6. Angle Based Outlier detection (ABOD)

각도 기반 이상치 감지 알고리즘입니다. 데이터 포인트와 다른 포인트 사이의 각도 분산을 이상 점수로 이용됩니다.



6. 비지도 학습의 이상탐지

비지도 학습은 레이블을 사용할 수 없어 AI 에이전트의 작업이 명확히 정의되지 않습니다. 그러나 엄격하게 정의된 작업이 없는 만큼 더 흥미로운 패턴을 발견할 수 있으며 제대로 활용히 아주 강력한 솔루션이 됩니다. 일반적으로 신용카드 사기, 통신 금융 사기 등 다양한 사기 탐지에 사용되며 악의적이고 드문 이벤트를 식별하는 데도 사용됩니다. 😃

▼ 참고 자료

- 차원 축소 및 PCA

  https://excelsior-cjh.tistory.com/167?category=918734

- t-SNE

  https://gaussian37.github.io/ml-concept-t_sne/#

- 클러스터링

  https://medium.com/analytics-vidhya/algorithm-selection-for-anomaly-detection-ef193fd0d6d1

- 이상탐지 활용 사례

  https://kh-kim.github.io/blog/2019/12/12/Deep-Anomaly-Detection.html

  https://leedakyeong.tistory.com/entry/Anomaly-Detection-by-Auto-Encoder

- knn

  https://bskyvision.com/563

- svm

  https://hleecaster.com/ml-svm-concept/