1. 이상감지모델의 평가방법을 적어보고 의미를 간단하게 적어보시오.
2-5. 각 이상감지 알고리즘의 특징을 간결하게 적어보고 장단점을 다른 이상감지알고리즘과 비교하여 적어보시오.

1. Evaluation for anomaly detection model

ROC(Receiver operating characteristic) Curve를 주로 사용한다.
X축은 FP/N, Y축은 TP/P을 사용한다.
예를 들어, 화재 경보 관련 모델이라고 생각해보자. X축이 1에 가까워지는 경우는 FP와 N이 비슷해진다는 뜻이므로, 별거 아닌 것에도 화재 경보가 울린다고 해석할 수 있다. Y축이 1에 가까워지는 경우는 TP와 P가 비슷해지는 것이므로, 모델의 예측과 실제 값이 거의 일치한다고 할 수 있다. 따라서 X축이 늘어남에 따라 Y축이 급격하게 증가할수록 좋은 것이라고 할 수 있다. 그러므로 Curve 아래의 넓이(AUROC)가 클수록 좋다는 뜻이라고 할 수 있다.

2. Robust Covariance

①특징
- N개의 데이터 중 Covariance determinant를 최소로 만드는 pure한 h개를 뽑는 방식이다.
- p-variance일 때 h의 범위는 일반적으로 $[(N+p+1)/2] \leq h \leq N$ 의 값을 사용한다.
- 데이터 분포가 어떤 식으로 흩뿌려져있는지 이미 알고 있을 때 유용하다.

②장점
- 모든 데이터의 Covariance를 이용해서 모델을 만드는 방식들은 outlier에 큰 영향을 받을 수 있으나 Robust Covariance는 이를 보정할 수 있다.

③단점
- n개의 데이터 중 h개를 뽑는 경우의 수가 nCh 이므로 runtime이 오래 걸릴 수 있다.

3. One-Class SVM

①특징
- classification을 위한 supervised learning 방법 중 하나이다.
- p-variance일 때, (p-1)-hyperplane을 찾는 것이 목표이다.
- Hard margin 방식과 Soft Margin 방식이 있다..

②장점
- Sample들이 linearly separable 할 때 잘 동작한다.
- Dimension을 확장하면 non-linear한 Sample에서도 사용할 수 있다.

③단점
- Dimension 확장하여도 적합한 함수를 찾기 힘들다.
- Feature의 개수가 많아지면 Dimension이 커져서 사용하기 힘들다.

4. Local Outlier Factor

①특징
- 각각의 관측치가 데이터 안에서 얼마나 벗어나 있는가에 대한 정도(이상치 정도)를 나타낸다.
- 모든 데이터를 전체적으로 고려하는 것이 아니라, 해당 관측치의 Neighbor를 이용하여 Local하게 이상치 정도를 파악한다.

②장점
- 고려할 Neighbor의 개수인 Hyper-Parameter만 결정하면 된다.
- 밀집된 Cluster에서 조금만 떨어져 있어도 Outlier로 탐지해준다.

③단점
- 기준이 될 Point와 거리를 설정해야 하며, 차원이 늘어나면 이는 더욱 어려워진다.

5. Isolation Forest

①특징
- Random Tree를 기반으로 하였다.
- Bagging ensemble method를 사용한다.
- Dimension과 Cut Value를 Random하게 정한다.
- 하나의 점을 isolation시키는데 필요한 Split의 횟수(Path Length)가 클수록 Inlier에 가깝고 Split의 횟수가 작을수록 Outlier라고 본다.
- Path Length는 Sub-Sample의 크기에 영향을 받으므로 Average Path Length를 이용하여 Anomaly Score을 구하는 방식을 사용한다.

②장점
- 기존의 classification과 clustering 방식보다는 좋은 성능을 낸다.
- 현실적인 Dataset에는 불필요한 Dimension도 포함되어 있는 경우가 많은데 그 경우에도 여전히 잘 작동하는 편이다.

③단점
- Sample 데이터에 대한 Label이 있어야 한다.
- 설정해야 하는 parameter가 많은 편이다.