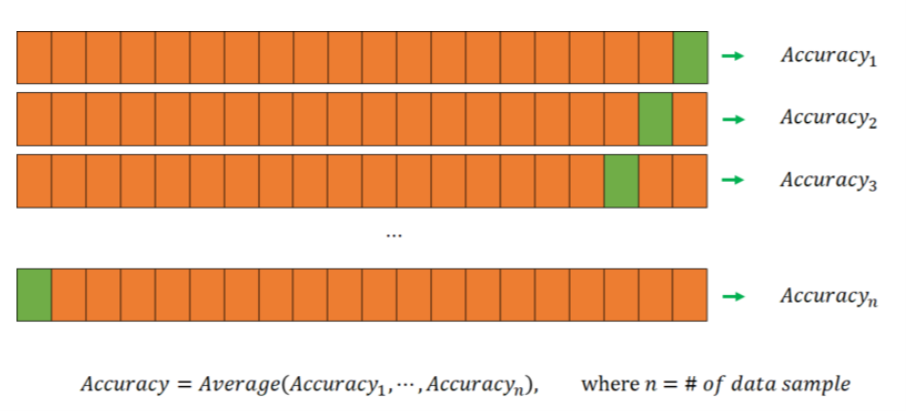
보통은 train set으로 모델을 훈련하고, test set으로 모델을 검증한다. 이때, 고정된 test set을 통해 모델의 성능을 검증하고 수정하는 과정을 반복하면 결국 이 모델은 test set에만 잘 동작하는 모델이 된다. 즉, test set에 overfitting하게 되는데 이를 해결하고자 하는 것이 cross validation이다. cross validation은 train set을 train set + validation set으로 분리한 뒤, validation set을 사용해 검증하는 방식이다.

cross validation은 모든 데이터셋을 훈련에 활용할 수 있다. 정확도를 높이며, 데이터 부족으로 인한 underfitting을 방지할 수 있다. 또한, 모든 데이터셋을 평가에 활용할 수 있다. 평가에 사용되는 데이터 편중을 막을 수 있고, 평과 결과에 따라 좀 더 일반화된 모델을 만들 수 있다. 그러나, iteration 횟수가 많기 때문에, 모델 훈련 및 평가 시간이 오래걸린다는 단점이 있다. 

K-Fold는 가장 일반적으로 사용되는 cross validation 방법이다. K는 폴드의 수를 나타내고 보통 5 또는 10을 사용한다. 과정은 다음과 같다. 전체 데이터셋을 training set과 test set으로 나눈다. Training set를 training set + validation set으로 사용하기 위해 k개의 폴드로 나눈다. 첫번째 폴드를 validation set으로 사용하고 나머지 폴드들을 training set으로 사용한다. 모델을 training한 뒤, 첫 번째 validation set으로 평가한다.차례대로 다음 폴드를 validation set으로 사용하며 3번을 반복한다. 총 k개의 성능 결과가 나오며, 이 k개의 평균을 해당 학습 모델의 성능이라고 한다.

Leve-one-out은 줄여서 LOOCV라고 불리우며, Leave-p-out에서 p=1일 때의 경우를 말한다. 폴드 하나에 샘플 하나만 들어있는 k-fold 검증으로 생각할 수 있다. Leave-p-out보다 계산 시간에 대한 부담은 줄어들고, 더 좋은 결과를 얻을 수 있기 때문에 선호된다. 검증에 사용되는 validation set의 개수가 적은 만큼 모델 훈련에 사용되는 데이터의 개수는 늘어난다. 모델 검증에 희생되는 데이터의 개수가 단 하나이기 때문에, 나머지 모든 데이터를 모델 훈련에 사용할 수 있다는 것이 장점이다.



교차검증의 결과로는 평균 정확도가 높고 각 Fold의 정확도가 일관되게 높은 경우, 모델이 잘 일반화되어 다른 데이터에 대해서도 좋은 예측을 수행할 것으로 기대할 수 있다. 평균 정확도가 낮거나 각 Fold의 정확도가 크게 다른 경우, 모델이 과적합된 것일 수 있다.

전처리 과정을 마친 데이터 CancerDataRm\_norm를 사용하여 LOOCV를 진행하였다. Data sample 개수인 487번을 반복하여 정확도를 계산하였다. 그 결과, 정확도의 평균이 1이 도출되었다.



데이터의 수가 많은 이유로, k-fold에서도 진행해본 결과 위와 같이 1이 나온 것을 볼 수 있다.



이에 오류가 발생했다고 본다. 첫 번째 추정 이유로는 데이터 전처리 과정에서 오류가 발생하여 모든 데이터가 동일한 값을 가질 수 있다. 이때, 모든 test data가 동일한 클래스로 예측되어 정확도가 1이 될 수 있다. 두 번째 추정 이유는 모델을 구현하는데 오류가 발생했을 수 있다. 세 번째 추정 이유는 LOOCV의 데이터를 분할하는 과정에서 데이터 분할이 올바르게 이루어지지 않아 모든 test data가 training data에 포함되는 경우가 발생할 수 있다.

오류 발생 원인을 알아보기 위해 전처리를 하지 않은 원데이터인 CancerData를 이용하여 cross validation을 진행하였다.



평균정확도가 0.9560633을 나왔다.

k-fold에서도 같은 결과를 보인다.



평균 정확도가 0.9562279로 비슷한 결과가 나왔고, 위의 오류 발생 원인에 기여를 한다.

Outlier를 제거한 데이터인 CancerDataRm를 이용하여 cross validation을 진행하였다. 487번을 반복하여 계산한 평균정확도는 0.9507187로 높은 정답률이 나왔다.



k-fold를 이용하여 계산해본 결과, 평균 정확도는0.9444728로 마찬가지로 높은 정답률이 나왔다.



이로 보아 오류의 원인은 첫 번째로 언급했던 데이터의 전처리로 인해 발생했다고 보는 것이 가장 옳다고 본다. Out;ier를 제거하는 과정은 유의미하다고 본다.