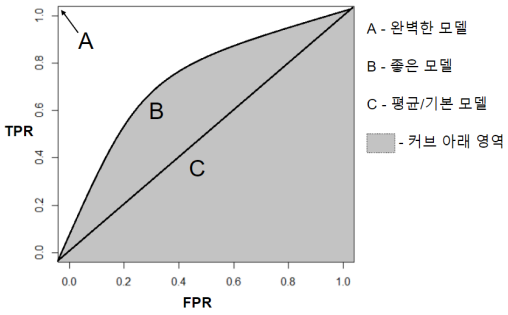
# AI+X:머신러닝 Assignment #4

ICT융합학부 2021093581 임동희

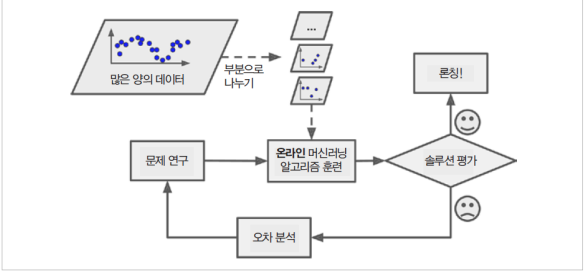
먼저, Precision (정밀도)란 TP / (FP+TP)로, 예측을 양성으로 한 대상중에 예측과 실제값이 양성으로 일치한 데이터의 비율입니다. 양성 예측 성능을 더욱 정밀하게 측정하기위한 평가지표로 양성예측도라고도 불립니다. 다음으로 Recall (재현율)이란 TP /(FN+TP)로, 실제값이 양성인 대상 중에 예측과 실제 값이 양성으로 일치한 데이터의 비율. 민감도라고도 불립니다. 정밀도와 재현율 지표의 중요성은 머신러닝이 적용되는 비즈니스의 특성에 따라 상이합니다. 우선, 재현율이 더 중요한 경우는 양성데이터를 음성으로 잘못 판단하면 영향이 발생하는 경우로 FN값을 줄이는 것을 목적으로합니다. 예시로는 암환자 예측이 있습니다. 암을 발견하지 못해서 치료시기를 놓치는경우가 있어서는 안되기 때문에 재현율이 더욱 중요합니다. 또, 금융사기 적발모델이 있습니다. 왜냐하면 금융사기를 확인하지 못하면 피해자 확인과 범죄자 적발이 어려워집니다. 또한, 금융거래를 정상거래로 처리하면 회사에 미치는 손해가 크기 때문에 더욱 중요합니다. 이외에도, 코로나환자판별등의 예시가 있습니다. 반대로, 민감도는 재현율이 더 중요합니다. FP값을 최대한 줄이는 것이 재현율을 사용하는 주요목적으로, 스팸메일 여부 적발모델이 대표적인 예시입니다. 왜냐하면 중요한 메일을 스팸처리 하지 않는 것이 더 중요합니다. 또, 스팸전화 차단어플의 알고리즘 경우도 동일합니다다. 업무전화나 합격전화와같은 중요한 전화를 놓치는경우를 없애는 것이 더욱 중요합니다다. 따라서 상황에 맞는 지표를 알맞게 조정하는 것이 중요합니다.



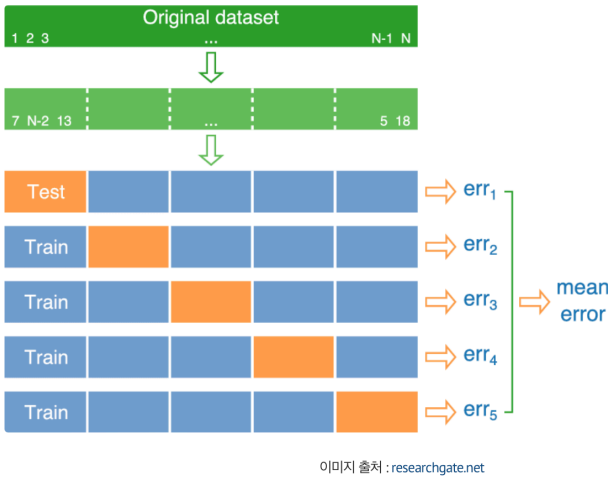

AUC커브는 Area Under the ROC curve로 ROC곡선 아래의 영역을 의미합니다. ROC는 확률 곡선으로 앞서말했던 내용을 판단하는 이진분류기입니다. AUC가 높으면 높을 수록 모든 임계값에서 분류모델의 성능이 높다는 것을 보여주는 것입니다. 그럴수록 그래프가 사각형에 가깝게 그려지게 됩니다. 그러면 곡선아래의 영역이 넓어 지게됩니다. 따라서, AUC가 높다는 것은 클래스를 구별하는 모델의 성능이 우수하다는 것을 의미하게 되는 것입니다. 따라서 머신러닝 성능판단에 많이 사용됩니다.

다음으로, 센서에서 연속적으로 수집되는 형태의 데이터에 머신러닝을 적용하는경우에  센서가 받는 데이터가 연속적으로 계속 저장되기 때문에 학습단계가 느리고 비용이 많이 들 수 있다는 단점이 존재합니다. 그래서 이를 해결할 수 있는 방법은 온라인 학습입니다. 온라인 학습에서는 데이터를 순차적으로 한개식 혹은 미니배치라 부르는 작은 묶음 단위로 주입하여 시스템을 훈련시키게 됩니다. 온라인학습은 연속적으로 데이터를 받고 빠른 변화에 스스로 적응해야하는 시스템에 적합합니다. 온라인 학습시스템은 새로운 데이터 샘플을 학습하면 학습이 끝난 데이터는 더이상 필요하지 않으므로 버리면 되기 때문에 많은 공간을 절약할 수 있다는 장점 또하 존재합니다.

그림 1-14 온라인 학습을 사용한 대량의 데이터 처리



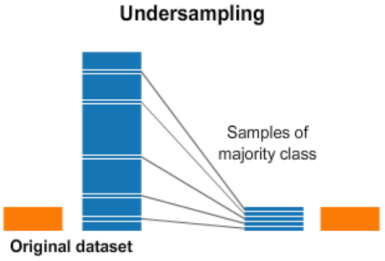
그러나 나쁜 데이터가 주입되면 시스템 성능이 점진적으로 감소할 수 있기 때문에 시스템을 면밀히 모니터링하고, 센서등의 성능을 확인하면 즉각 학습을 중지시키거나 이전 운영상태로 되돌려야합니다.

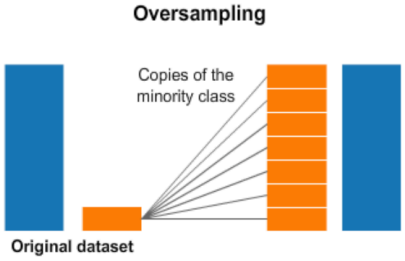


이미지 출처 : researchgate.net

그리고 다음은 검증을 사용하는 이유와 방법중 한가지에 대해 이야기해보고자합니다. 머신러닝은 학습데이터로 모델을 생성한 뒤 중간과정 없이 테스트 데이터만으로 모델을 평가한다면 과적합이나 과소적합이 발생할수 있습니다. 하지만 모델이 우리가 가지고있는 모든 데이터를 봤기 때문에 이모델을 개선하거나 다시 테스트할 수 있는 방도가 존재하지 않습니다. 과소적합이나 과적합이 일어나지않았다하더라도 이 모델이 최선인지 더 나은 모델은 없는지 알 수 있는 방법이 존재하지않습니다. 따라서 이를 확인할 수 있는 검증(Validation)과정이 필요합니다. 검증 과정중 한가지 예시는 교차검증의 대표적인 방식인 K-fold 교차검증(K-fold cross validation)이 있습니다. 데이터를 K개로 나눈 뒤 차례대로 하나씩을 검증 데이터셋으로 활용하여 K번 검증을 진행하는 방식. 학습과정마다 학습데이터에 손실이 존재하기는 하지만 전체 데이터를 다 볼 수 있고, 검증횟수를 늘릴 수 있다는 장점 또한 존재합니다.

다음으로는 클래스 불균형과 이의 해소 방법에 대해 이야기해보고자 합니다. 현실 데이터에는 클래스 불균형 문제가 자주있습니다. 어떤 데이터에서 각 클래스가 갖고 있는 데이터의 양에 차이가 큰 경우, 클래스 불균형이 있다고 말합니다. 클래스균형은 소수의 클래스에 특별히 더 큰관심이 있는 경우에 필요합니다. 이럴 때 성능 평가에 정확도만 사용하게 되면  클래스의 무게 차이가 매우 크기 때문에, 정확도에도 큰 차이가 존재하게 됩니다. 이렇게 소수의 클래스에 비해 다수의 클래스가 훨씬 더 큰 정확도를 가지기 때문에 리스크가 큰 문제의 경우, 소수의 클래스도 고려하기 위해서 클래스 균형을 맞추는 것이 필요합니다. 이를 위한 방법에는 Weight balancing이 있다. 학습데이터에서 각 loss를 계산할 때 특정 클래스에 대해서는 더 큰 loss를 계산해주는 방법이다. 더 큰 정확도가 필요하기 때문에 더 큰 loss를 취해주는 것입니다. 또 다른 방법으로는 클래스의 비율에 대해 가중치를 두는 방법이 있습니다. 예를 들어 두 개의 클래스 비율이 1:9라면 가중치를9:1로 줌으로써 전체 클래스의 loss에 동일하게 기여하도록하는 방법입니다.




또 다른 방법에는 oversampling과 undersampling도 있습니다. 이는 현저히 양이 많은 데이터인 그림의 경우, 파란색데이터에 맞춰주는 것인데, 양을 줄여버린 파란

색 데이터들은 양이 많았던 원본 데이터의 대표성을 잘 지니고 있어야합니다. oversampling은 양이적은 데이터를 양이많은 데이터에 맞춰주는 방법으로 양이 적은 주황색데이터를 늘릴 때 데이터를 복사하는 개념이기 때문에 양이 늘어난 주황색데이터는양이 적었던 원본데이터와 성질이 동일합니다.

마지막으로, 피쳐란 머신러닝 알고리즘에 넣는 데이터들의 다른이름으로, 피쳐 중요도란 어느 데이터가 확률 값 계산에 중요하게 작용을 했는지의 정도에 대해 나타내는지표입니다. 머신러닝은 예측력이 좋아지는 대신 굉장히 복잡해 졌기 때문에 어떤 변수가 예측에 있어 중요한 변수인지를 파악하기 어렵기 때문에 모형의 예측력과 관련해 개별 변수의 영향력을 측정하는 도구가 필요해졌기 때문에 피처 중요도를 구하게 됩니다.  피처 중요도를 구하는 방식에는 Correlation, 상관분석이 존재합니다. 상관분석은 확률론과 통계학에서 두 변수간에 어떤 선형적 관계를 갖고 있는지를 분석하는 방법입니다. 두 변수는 서로 독립적인 관계로부터 서로 상관된 관계일 수 있으며, 이때 두 변수간의 관계의 강도를 상관관계라고 합니다. Correlation은 구현이 쉽고 직관적입니다. 뿐만 아니라 설명 변수 단위에 영향을 받지 않고, 특별한 예측 모형 학습 과정이 필요하지 않다는 장점이 존재합니다. 하지만 여러 설명변수에 대한 영향력이 존재하는 경우에는 실제로 중요한 변수가 관계없는 변수보다 오히려 중요도가 낮게 나올 수 있고, 범주형 반응 변수에는 적용이 어려울 수 있다는 단점이 존재합니다.