A theory of learning from different domains

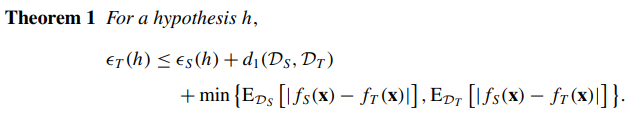
저자는 두 가지 질문에 대한 대답을 논문에서 제시한다.

1. 어떠한 조건 아래에서 target domain에 대한 source domain에서 training된 classifier가 잘 잘동하는가?

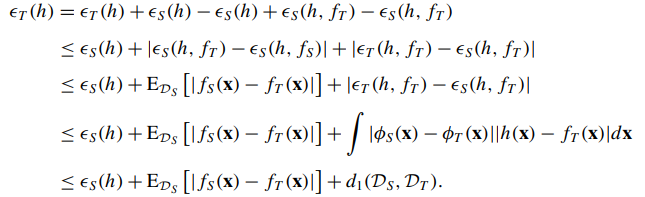
2. source data와 target data를 어떻게 적절하게 이용해야 target data에 대한 error를 줄일 수 있는가?

우선 첫 번째 질문에 대한 대답을 살펴보자.

저자는 target domain에 대한 classifier의 performance를 다음과 같이 bound 할 수 있다고 한다.

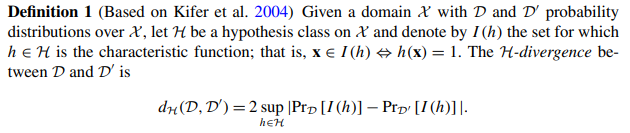


식을 먼저 이해하자면 target domain에 대한 error는 source domain에 대한 error와 두 분포 간의 거리 그리고 각 도메인에서 labeling function의 합으로 제한된다. 다른 말로는 source domain에서도 잘 작동하고 domain간의 거리도 가까워야 하고(가까우면 잘 작동할 것이라는 전제이다.) 또한 domain과 상관없이 labeling function의 차이가 적어야 한다는 것이다. 이에 대한 증명은 다음과 같다.



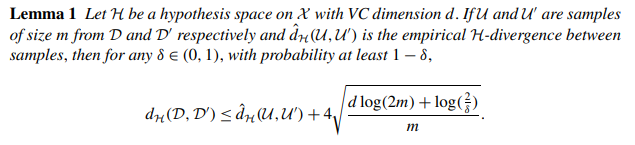
증명을 제외하고 Theorem 1의 각 부분을 살펴보면 우변의 첫 번째 term은 source domain 에서의 error이다. 그리고 마지막 term은 source domain 또는 target domain에서 labeling function의 차이이다. 하지만 중간에 두 번째 term이 문제이다. 왜냐하면 는 finite sample from specific domain에 대하여 정확한 계산이 불가능하고 매우 general하게 target error를 제안한다. 이는 실제 error보다 더 큰 값을 제시할 수 있는 상황이 발생하게끔 만든다.

의 정의는 다음과 같다.

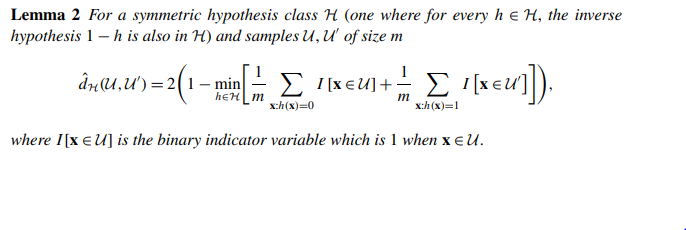
는 기존의 의 문제점을 해결하였다. 첫 번째로 유한한 VC dimension을 가진 class상에서 는 finite sample만으로도 estimation이 가능하다. 두 번째로 어떠한 class 의 는 보다 무조건 작다.

식을 정성적으로 살펴보면 다음과 같다. 라는 hypothesis class안에 존재하는 어떤 classifier 는 sample data가 들어오면 sample data가 source data에서 왔으면 1을 return하고 target domain에서 오면 0을 return하는 classifier라고 하자. 만약 가 어느 쪽에서 data가 입력이 되든 둘 다 1을 return하게 되는 경우 두 가 0이 된다. 가장 이상적인 상황이다. 그런데 둘 다 0을 return하는 경우는 이 또한 두 Domain을 구별하지 못하는 것이므로 이 상황도 이상적인 상황이 된다.

[잠깐 VC dimension에 대한 이야기를 하고 넘어가겠다. 수식적인 증명 및 설명은 완벽히 이해가 되지 않았기에 정성적인 이야기를 하자면 다음과 같다. 가 얼마나 복잡한지 측정하는 measurement가 VC dimension이다. 는 classifier의 집합이라고 하면 VC()는 가 나눌 수 있는 training point의 개수로 정의가 가능하다. 이러한 VC dimension은 classifier의 형태를 정의하는 순간 고정이 된다. Training에 의한 parameter가 변하는 것은 집합 안에서 적절한 classifier를 sampling하는 것이므로 VC dimension에 대한 차원의 변화는 생기지 않는다. 이러한 VC dimension의 Upper bound는 model의 Node와 Edge의 개수에 비례한다.]



이제 의 Lemma에 대해서 설명을 하자면 에서 sampling한 data subset 의 에 결국 converge 한다는 것을 나타낸다. 이는 우리가 Labeled domain과 Unlabeled domain사이에서 error를 일시적으로 구할 수 있게 해준다. 그런데…. divergence값을 마구 줄여버리게 되면 model의 complexity 가 끝도 없이 증가한다. 그러면 divergence의 upper bound가 증가하는 모양이므로 좋지 않다. 결국 목표는 를 줄이고 sampled divergence의 값을 줄여야 하는 것이다.



Lemma2를 살펴보면 이게 뭔 소리냐? 이럴 수 있다. 수식적인 증명은 그다지 필요없는 것 같고 이 Lemma를 통해 를 적절하게 approximation할 수 있다고 한다.