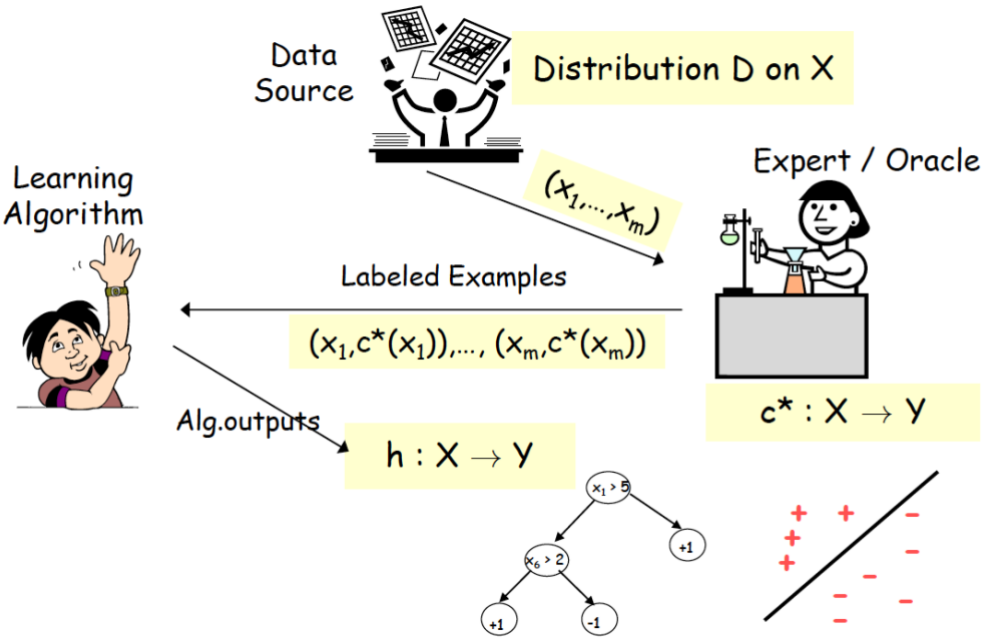
5. PAC LEARNING IN LARGE INPUT SPACES

PAC(Probably Approximately Correct) learning은 이론적으로 모델의 성능을 측정하는 방법 중 하나로, 비록 practical하게 모델의 성능 측정에 쓰기는 힘들지만 개념적으로 왜 특정 모델이 더 좋은지, 언제 모델의 성능이 좋고 언제 나쁜지에 대해 설명하는 것은 가능하다. "높은 확률로 (Probably)" 주어진 모델이 "작은 error를 가진다 (Approximately Correct)"와 같은 분석을 할 수 있다.



매커니즘을 간단하게 설명해 보면 위 그림과 같다.

앞의 section에서 상대적으로 작은 instance $|X|$ 경우에 대해 알아보았다면 이제는 instance $|X|$의 크기가 큰 경우를 살펴볼 것이다. 이 논문에서 보여주고자 하는 것은 distribution D, 즉, train data D에 대한 조건부 독립 가정이 만족될 때 concept class (각각의 view에서 추론한 함수)가 standard PAC model의 random classification noise로부터 학습이 가능하다면 처음의 weak predictor는 boosted될 수 있다.

그러면 구체적으로 distribution D와 target function f_1, f_2(concept class의 원소)가 조건부 독립 가정을 만족하는 조건에 대해서 알아볼 것이다. 확률이 0이 아닌 $(\hat{x}_1, \hat{x}_2) \in X$에 대해서

$$Pr_{(x_1,x_2)\in X}[x_1 = \hat{x}_1 | x_2 = \hat{x}_2] = Pr_{(x_1,x_2)\in X}[x_1 = \hat{x}_1 | f_2(x_2) = f_2(\hat{x}_2)]$$

를 만족하면서 비슷하게

$$Pr_{(x_1,x_2)\in X}[x_2 = \hat{x}_2 | x_1 = \hat{x}_1] = Pr_{(x_1,x_2)\in X}[x_2 = \hat{x}_2 | f_1(x_1) = f_1(\hat{x}_1)]$$

를 만족해야 한다,

정리하자면 x_1과 x_2는 label에 대해서 조건부 독립이라는 말이다. 조금 더 쉽게 예시로 설명하자면 P라는 페이지의 단어들은 P라는 페이지의 하이퍼링크와 서로 독립이다. 이 말은 만약 P페이지를 구축한 사람과 하이퍼링크를 만든 사람이 다른 사람이라면 그럴듯하다. 하지만 아래의 Theorem 1을 보면 왜 이것이 생각만큼 그럴듯하지 않은지 알 수 있다.
첫 번째 이론을 살펴보기에 앞서 function f의 "weakly-useful predictor" h를 정의해야 한다. weakly-useful predictor는 instance space의 instance 하나로 predictor를 만드는 것이고 그래서 약한데 유용한 predictor라고 하는 것이 아닌가 생각한다.(뇌피셜...) 그리고 ϵ는 $\epsilon > 1/poly(n)$으로 정의한다.
예를 들어 어떤 웹 페이지의 "handout"이라는 단어가 그 웹 페이지가 강의 홈페이지라고 예측하는 weakly useful predictor가 될 수 있다고 가정할 때 다음과 같은 조건들을 만족해야 한다.

1. $Pr_D[h(x) = 1] \geq \epsilon$, 그리고
2. $Pr_D[f(x) = 1 | h(x) = 1] \geq Pr_D[f(x) = 1] + \epsilon$

1번 식은 "handout"이라는 단어가 그 웹페이지에서 무시할 수 없는 부분이라는 조건이고, 2번 식은 "handout"이라는 단어가 그 웹 페이지에 나타났을 때 그 웹 페이지가 강의 홈페이지일 확률이 "handout"이라는 단어가 나타나지 않았을 때의 확률보다 높다는 조건이다. weakly useful predictor에 조금 더 설명하자면 f가 unbiased 함수라고 할 때, 즉, 1일 확률과 0일 확률이 각각 1/2일 때 $Pr_D\big(h(x) = f(x)\big) \geq \frac{1}{2} + 1/poly(n)$이 성립한다. 즉, h를 이용하면 모두 0으로 찍거나 모두 1로 찍는 것보다 좀 나아진다. 그렇다면 f가 unbiased하지 않을 때는 h를 이용하면 그냥 0 또는 1로 아무거나 찍는 것보다 현저하게 나은 결과가 나온다.

Theorem 1은 다음과 같다.
만약 C_2(두 번째 view에서 나온 목적 함수들의 collection)가 classification noise로 PAC 모델에 학습가능하고, 조건부 독립 가정이 성립한다면 weakly useful predictor인 $h(x_1)$이 주어졌을 때 (C_1, C_2)가 unlabeled data만 있는 상황에서 Co-training model 학습이 가능하다.

이 정리를 증명하기에 앞서 standard classification noise model의 변수를 정의하고 가는 것이 편하다. positive example의 noise와 negative example의 noise는 다를 수 있다. 따라서 (α, β)를 noise rate set이라 하고 α를 진짜 positive example이 잘못 labeling된 확률이라 하고 β를 진짜 negative example이 잘못 labeling된 확률이라고 하면 $\alpha = \beta$일 필요는 없다. 학습 알고리즘의 목표는 여전히 non-noisy data에 대해서 target function과

비슷한 hypothesis를 찾는 것이다. 이러한 경우 다음과 같은 부명제가 있다.

Lemma 1은 다음과 같다.
concept class C(목적 함수들의 집합)가 standard classification noise model에서 학습이 가능하다면 C는 $\alpha+\beta<1$일 때 (α,β) classification noise로 학습이 가능하다. running time은 $1/(1-\alpha-\beta)$이고, $\hat{p}=\min[Pr_D(f(x)=1), Pr_D(f(x)=0)]$일 때 $1/\hat{p}$로 표현된다.

Lemma1에서 보듯이 각각의 noise rate는 1/2보다 작아야 한다. 따라서 $\alpha+\beta<1$를 가정한다. 본격적으로 Theorem 1을 증명하기에 앞서 $f(x)$를 목적 함수, $p=Pr_D[f(x)=1]$ 는 랜덤하게 추출한 D 에서 positive의 확률, $q=Pr_D(f(x)=1|h(x_1)=1)$, $c=Pr_D(h(x_1)=1)$이라고 정의한다.

$$Pr_D[h(x_1)=1|f(x)=1]$$

$$=\frac{Pr_D[f(x)=1|h(x)=1]Pr_D[h(x_1)=1]}{Pr_D[f(x)=1]}$$

$$=\frac{qc}{p}$$

이 식은 실제 class가 1일 때 가설이 1로 예측할 확률을 베이지안 확률 정의로 표현한 것이고, 앞서 정의한 p,q,c로 나타내었다.

$$Pr_D[h(x_1)=1|f(x)=0]=\frac{(1-q)c}{1-q}$$

위의 식은 실제 class가 0일 때 x_1으로 예측한 가설이 1로 잘못 예측할 확률을 나타내며, 앞서 정의한 p,q,c로 나타내었다.
조건부 독립 가정에 의해 random example $x=(x_1,x_2)$에 대해 $h(x_1)$ (첫 번째 view의 instance x_1 로 만든 predictor)는 x_2 와 독립이다. "handout" 예시에서 말한 것처럼 "handout"이라는 단어로 예측한 웹 페이지의 class는 그 웹 페이지의 하이퍼링크, 하이퍼링크의 라벨과는 독립이다. 그렇기 때문에 만약 $h(x_1)$ 을 x_2 의 잘못 labeling된 label로 사용한다면, 이것은 (α,β) -classification noise와 같아진다. 여기서 $\alpha=1-qc/p$이고, $\beta=(1-q)c/(1-q)$이다. 이것을 이용하면

$$\alpha+\beta=1-\frac{qc}{p}+\frac{(1-q)c}{1-q}=1-c(\frac{q-p}{p(1-p)})$$

식을 이렇게 쓸 수 있다. 앞에서 언급했던, weakly-useful predictor h를 정의하기 위해 했던 가정들을 p,q,c로 나타내면 $Pr_D[h(x)=1]\geq\epsilon$이므로 $c=Pr_D(h(x_1)=1)$니까 $c\geq\epsilon$로

표현할 수 있다. 마찬가지로 $Pr_D[f(x)=1|h(x)=1] \geq Pr_D[f(x)=1]+\epsilon$ 이므로 이항하면 $q-p \geq \epsilon$로 나타낼 수 있다. 그러니까 $c=\epsilon$, $q-p=\epsilon$라고 하면 $\alpha+\beta$의 값은 기껏해야 $1-c^2/(p(1-p))$ 가 되며, 결국 기껏해야 $1-4\epsilon^2$ 이다.(f가 unbiased 하여$p=1/2$ 라고 생각하면) 따라서 $\alpha+\beta<1$이라는 Lemma1에 적용하여 조건부 독립 가정이 성립한다면 weakly useful predictor인 $h(x_1)$ 이 주어졌을 때 (C_1, C_2) 가 unlabeled data만 있는 상황에서 Co-training model 학습이 가능하다는 Theorem1을 확인할 수 있다.

5.1 RELAXING THE ASSUMPTIONS

가정을 완화하는 것인데 이제껏 엄격한 우리의 가정에서 목적 함수 (f_1, f_2)에 대해 $f_1(x_1) \neq f_2(x_2)$인 (x_1, x_2)는 있을 수 없었다. 이제는 조건부 독립 가정은 유지하되, 이 가정이 상당히 약화될 수 있다고 가정하며, 마찬가지로 weakly-useful predictor를 unlabeled data로 boost할 수 있게 허용한다. 그러니까 약화된 가정에서는 $f_1(x_1) \neq f_2(x_2)$ 과 같은 경우도 존재할 수 있다는 것이다. 따라서 각각의 경우를 다음과 같이 정리한다.

$$p_{11} = Pr_D[f_1(x_1)=1, f_2(x_2)=1]$$
$$p_{10} = Pr_D[f_1(x_1)=1, f_2(x_2)=0]$$
$$p_{01} = Pr_D[f_1(x_1)=0, f_2(x_2)=1]$$
$$p_{00} = Pr_D[f_1(x_1)=0, f_2(x_2)=0]$$

엄격한 가정 하에서는 $p_{10}=p_{01}=0$으로 가정했었다. 그리고 p_{11}와 p_{00} 둘 다 0에 가깝지 않다는 가정을 함축하고 있었다. 이제는 다음과 같이 가정을 완화하도록 한다.

$$p_{11}p_{00} > p_{10}p_{01} + \delta$$

여기서 $\delta = 1/poly(n)$이다. 우리는 앞서 계속 언급했던 조건부 독립 가정을 계속 유지할 것이고, 따라서 p_{10} 은 random하게 뽑은 x_1 이 positive이면서, 이와 독립적으로 random하게 뽑은 x_2가 negative인 확률이다.

시나리오를 완전하게 이해하기 위해서는 labeling process에 대한 설명이 필요하다. (x_1, x_2)가 $f_1(x_1)=1, f_2(x_2)=0$일 때 positive로 labeling되었다면 그 확률은 어떻게 될까? 이 문제는 우리가 f_1 의 weakly-useful predictor h를 얻기 위해 labeled data로부터 어떻게든 충분한 정보를 얻었으며, 앞으로는 unlabeled data에만 신경 쓴다는 가정을 통해 처리될 수 있다. 특히, 다음과 같은 정리를 얻을 수 있다.

Theorem 2는 다음과 같다.

$h(x_1)$을 다음과 같이 hypothesis로 정의한다.

$\alpha = Pr_D[h(x_1)=0|f_1(x_1)=1]$

그리고 $\beta = Pr_D[h(x_1)=1|f_1(x_1)=0]$이면

$Pr_D[h(x_1)=0|f_2(x_2)=1] + Pr_D[h(x_1)=1|f_2(x_2)=0]$

$$= 1 - \frac{(1-\alpha-\beta)(p_{11}p_{00} - p_{01}p_{10})}{(p_{11}+p_{01})(p_{10}+p_{00})}$$

즉, h가 f_1에 대해 usable한 (α, β) classification noise를 만든다면 이 h는 f_2에 대해서도 usable한 $(\acute{\alpha}, \acute{\beta})$ classification noise를 만들 수 있을 것이다. 위의 정리를 풀어서 계산해보면 다음과 같다.

$$Pr_D[h(x_1) = 0 | f_2(x_2) = 1] + Pr_D[h(x_1) = 1 | f_2(x_2) = 0]$$

$$= \frac{p_{11}\alpha + p_{01}(1-\beta)}{p_{11}+p_{01}} + \frac{p_{10}(1-\alpha) + p_{00}\beta}{p_{10}+p_{00}}$$

$$= 1 - \frac{p_{11}(1-\alpha) + p_{01}\beta}{p_{11}+p_{01}} + \frac{p_{10}(1-\alpha) + p_{00}\beta}{p_{10}+p_{00}}$$

$$= 1 - \frac{(1-\alpha-\beta)(p_{11}p_{00} - p_{01}p_{10})}{(p_{11}+p_{01})(p_{10}+p_{00})}$$

$\frac{p_{11}\alpha+p_{01}(1-\beta)}{p_{11}+p_{01}}$ 에서는 $p_{11}+p_{01}$이 $f_2(x_2) = 1$일 확률이고 p_{11}이 $f_1(x_1) = 1 \& f_2(x_2) = 1$인 확률(서로 독립이니까 곱으로 표현 가능) $\alpha = Pr_D[h(x_1) = 0 | f_1(x_1) = 1]$이므로 베이지안 확률을 통해 계산 하면 $\frac{p_{11}\alpha}{p_{11}+p_{01}}$ 이 부분이 $h(x_1) = 0$이면서 $f_1(x_1) = 1$인 확률이 되고, $\frac{p_{01}(1-\beta)}{p_{11}+p_{01}}$ 이 부분은 $h(x_1) = 0$이면서 $f_1(x_1) = 0$인 확률이 되어 둘을 더하면 $h(x_1) = 0$의 확률이 되고, $Pr_D[h(x_1) = 0 | f_2(x_2) = 1]$를 α와 β, p 등으로 나타내고 있다는 것을 알 수 있다. 마찬가지로 뒤에는 이렇게 식을 계속 변형해 나가는데 결국 $Pr_D[h(x_1) = 0 | f_2(x_2) = 1] + Pr_D[h(x_1) = 1 | f_2(x_2) = 0]$ 이 확률을 $1 - \frac{(1-\alpha-\beta)(p_{11}p_{00}-p_{01}p_{10})}{(p_{11}+p_{01})(p_{10}+p_{00})}$ 이렇게 표현가능하고, $\frac{(1-\alpha-\beta)(p_{11}p_{00}-p_{01}p_{10})}{(p_{11}+p_{01})(p_{10}+p_{00})}$ 이 부분이 $h(x_1)$과 $f_2(x_2)$의 labeling이 같을 확률이라고 할 수 있다. $1-\alpha-\beta$는 앞서 언급했던 $\alpha+\beta<1$ 가정에 의해 $1-\alpha-\beta > 0$이 된다 $p_{11}p_{00} - p_{01}p_{10}$ 이 부분은 앞의 완화된 가정 $p_{11}p_{00} > p_{10}p_{01} + \delta$ 을 통해 $p_{11}p_{00} - p_{01}p_{10} > \delta$로 생각할 수 있고, 따라서 이 확률이 그리 작지 않다고 말할 수 있다. 즉, f_1에 대해 usable한 (α, β) classification noise를 만든 $h(x_1)$이 f_2에 대해서도 usable한 $(\acute{\alpha}, \acute{\beta})$ classification noise를 만들 수 있다고 정리할 수 있다.

PAC-learning을 시작할 때 언급했던 내용인데 x_1과 x_2는 label에 대해서 조건부 독립이라는 말이 그럴듯하지 않을 수 있다는 것을 Theorem1으로 보여줄 수 있다고 했는데 그 예시로 설명하면 웹 페이지에 있는 단어로 만든 weakly usable predictor $h(x_1)$이 주어졌을 때, C_2, 그러니까 그 홈페이지의 하이퍼링크로 만든 목적 함수가 classification noise로 PAC model에 학습이 가능하기 때문에 x_1가 x_2의 label과 완전히 독립은 아닐 수 있다는 말이 아닐까 한다...