Learning From Data: 통계 학습 이론 정리

jongman@gmail.com

January 24, 2015

Abstract

이 노트는 통계 학습 이론에 관한 학부 수준의 좋은 교과서인 Learning From Data(Abu-Mostafa 외)에서 통계 학습 이론 관련된 내용만을 발췌 요약한 것이다.

1 학습은 가능한가?

1.1 학습은 불가능하다!

기계 학습 알고리즘이 임의의 자료가 주어질 때 그 자료로부터 무언가를 배울 수 있을까? In-sample 자료를 살펴보면 그 밖의 자료에 대해서도 무언가를 알 수 있다고 우리가 수학적으로 보장할 수 있을까? 어떤 가정도 없이 자료로부터 무언가를 배우는 것은 불가능하다는 것을 다음 주장을 통해 알 수 있다.

크기 3의 진리값 벡터를 입력으로 갖는 이진 분류(binary classification) 문제가 있다고 하자. 따라서 8가지 가능한 입력이 있고, 존재 가능한 정답 함수는 2^8개가 있다. 만약 8가지 가능한 입력 중 5개가 학습용으로 주어졌다고 하자. 그러면 가능한 정답 함수 2^8개 중 2^3개는 이 5개의 예제 입력에 대한 정답을 모두 만족할 것이다. 과연 이 중 무엇이 정답에 더 가까울지 골라낼 수 있을까? 주어진 자료 하에서는 이 함수들은 구분할 수 없다. 따라서 우리는 아무것도 배울 수 없다!

과연 현실이 이렇게 우울할까? Out-of-sample 자료에 대해서는 우리는 어떠한 보장도 할 수 없을까?

1.2 확률적 접근

이와 같은 문제를 해결하기 위해 통계 학습 이론이 사용하는 중요한 가정은 in-sample 입력은 가능한 모든 입력 중에서 임의로 선택되었다는 것이다. 이와 같은 가정을 하면 우리는 이제 우리가 보지 못한 자료에 관해서도 알 수 있게 된다.

간단한 예제를 들어 보자. 커다란 상자에 엄청나게 많은 구슬이 들어 있는데, 이중 일부는 붉은색, 나머지는 초록색이라고 하자. 전체 구슬 중 붉은 색의 비율은 $\mu \times 100\%$ 이다. 이 상자에서 임의로 구슬을 N개 뽑았는데, 그 중 붉은 구슬의 비율은 $\nu \times 100\%$였다. 이 때 ν를 보면 μ에 대해 알 수 있는 것이 있을까? 물론이다. 만약 1000개의 구슬을 뽑았는데 붉은 구슬이 그 중 990개였다고 하자. 우리는 당연히 상자에는 붉은 구슬이 훨씬 많이 들어 있을 것이라고 예상할 수 있다. 이 직관을 어떻게 수학적으로 나타낼 수 있을까? 확률론의 힘을 빌리면 가능하다. 우리가 뽑은 구슬은 모두 임의로 뽑았기 때문에, 이것을 베르누이 확률 변수로 나타내자. 그러면 변수들의 관찰 평균값(ν)이 기대치(μ)에서 벗어날 확률을 계산하는 정리인 Hoeffding 부등식을 사용할 수 있는데, 다음과 같은 형태를 갖는다:

$$\mathrm{P}\left[|\nu - \mu| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}$$

여기서 ϵ는 우리가 정하는 값으로, ν와 μ 사이의 오류를 나타낸다. 이 오류가 정해지면, Hoeffding 부등식을 이용해 우리가 확인한 ν가 참값인 μ에서 ϵ 이상 벗어날 확률을 ϵ과 N의 함수로 나타낼 수 있게 된다. 따라서 샘플의 크기가 커지면, 부등식의 우변이 줄어들기 때문에 ν가 μ에 점점 가까워지게 됨을 알 수 있다.

이 예와 학습 문제 사이에는 어떤 관련이 있을까? 우리가 어떤 형태의 가설 h를 세웠다고 하자. 이 가설의 형태는 중요하지 않다. 이 가설을 가능한 모든 데이터에 대해 적용하면, out-of-sample 오류율 $E_{out}(h)$를 얻게 될 것이다. 물론 모든 데이터를 얻는 것은 불가능하므로, 이 데이터 중 임의로 뽑은 in-sample 데이터에 대해 가설을 적용해 보고 오류율 $E_{in}(h)$를 계산해 보자. 샘플의 크기가 늘어나면 늘어날 수록 $E_{in}(h)$가 실제 오류율 $E_{out}(h)$에 가까워질 것이라는 신뢰를 가질 수 있게 된다. 따라서 이를 이용해, 우리가 아직 보지 못한 자료에 관해서 다음과 같이 말할 수 있게 된다:

> 우리가 정한 이 가설 h의 오류율이 이 범위를 벗어날 확률은 얼마이다.

여기에서 가설은 우리가 알고 있는 것(in-sample 오류율)을 우리가 모르는 것(out-sample 오류율)으로 전달하는 역할을 한다. 우리는 목적 함수에 대해 안다고 말할 수 없으며, 우리가 가진 가설의 오류율에 대해서만 이야기할 수 있다.

마지막으로:

- 교과서의 논의는 분류에 대해서만 다뤘지만, 이와 같은 형태의 증명은 회귀 분석과 같은 다른 지도 학습(supervised learning)에도 적용될 수 있다(고 한다).
- Hoeffding 부등식은, 당연하게도, in-sample 데이터가 임의로 선택되었을 때만 적용된다. 모든 통계 학습 이론은 이 가정에 기반을 두고 있다.

1.3 다수의 가설과 합계 상한(Union Bound)

모든 통계 학습 문제는 여러 개의 가설 $\{h_i\}$중에서 가장 그럴 듯한 것을 선택하는 과정이다. (애초에 가설이 하나밖에 없다면, 학습이 아니라 검증에 불과할 것이다.) 이 때 어떤 가설을 고르는 것이 좋을까? 당연히 in-sample 오류율이 가장 낮은 가설을 택하는 것이 좋을 것이다. 이 가설을 g라고 부르자. 그러면 g의 성능, 다시 말해 $E_{out}(g)$에 대해 무슨 말을 할 수 있을까? Hoeffiding 부등식이 여기에도 적용될까? 안타깝게도, 그렇지 않다.

여기서 문제는 g는 우리가 in-sample 데이터를 본 후에야 결정된다는 것이다. 물론 h들의 집합을 알고 있으니, 이 중의 하나가 g가 된다는 것은 알지만, 이 중에 무엇일지 알 수가 없다. 따라서 $E_{out}(g)$의 분포는 $E_{out}(h)$의 분포와 달라지게 된다. 이 문제에 대한 적절한 설명을 하기 어려운데, 간단한 예만 들도록 하겠다. 어느 나라의 고3 학생들의 키를 모두 모아 보면 정규 분포를 따른다고 하자. 임의의 학생을 하나 고르면, 이 학생의 키는 해당 정규 분포를 따르게 된다. 그런데 각 학교마다 가장 큰 학생들을 모아 놓고, 이 중에서 임의의 학생을 뽑으면 어떨까? 원래의 정규 분포와는 당연히 다를 것이다.

이런 이유로, Hoeffding 부등식을 바로 적용하면 오류율을 엄청나게 낮춰잡는 결과를 가져오게 된다. 그러면 어떻게 할까? 이제는 확률 $\mathrm{P}\left[|E_{in}(g) - E_{out}(g)| > \epsilon\right]$에 대해 어떤 보장도 할 수 없을까? 무식한 방법 하나는 최종 가설 g는 항상 우리의 가설 집합 $\{h_i\}$중에 있다는 것을 이용하는 것이다. 따라서 만약 g의 오류율이 ϵ을 넘는다면, $\{h_i\}$중 최소한 하나는 ϵ을 넘는 오류율을 가져야 할 것이다. 사건 A가 B의 충분조건일 때, $\mathrm{P}(A) \leq \mathrm{P}(B)$ 이므로, 다음과 같은 상한을 얻을 수 있다.

$$\begin{aligned}
\mathrm{P}\left[|E_{in}(g) - E_{out}(g)| > \epsilon\right] &\leq \mathrm{P}\left[\cup_{i=1\sim M} |E_{in}(h_i) - E_{out}(h_i)| > \epsilon\right] \\
&\leq \sum_{i=1}^{M} \mathrm{P}\left[|E_{in}(h_i) - E_{out}(h_i)| > \epsilon\right] \\
&\leq 2Me^{-2\epsilon^2 N}
\end{aligned}$$

이 상한을 합계 상한(union bound)라고 부른다. 이 수식에는 이제 가능한 가설의 수인 M이 포함되어 있다는 것에 주목해 보자. 상한이 나온 것은 좋은데, 대부분의 기계 학습 알고리즘에는 무한한 수의 가설이 있다. 실수 인자(parameter) 하나만 있다고 하더라도 무한한 수의 가설을 갖게 되기 때문이다. 이 경우, 합계 상한은 무한대가 되기 때문에 아무런 의미가 없다. 뒤에서 우리는 VC-차원(VC-dimension)이라 부르는 값을 소개하는데, 이를 이용하면 유한한 확률 상한을 구할 수 있다.

1.4 학습 문제의 두 측면

기계 학습의 궁극적 목표는 out-of-sample 오류율 0을 달성하는 것이다. 그러나 Hoeffding 부등식은 단지 in-sample 오류율 E_{in}과 out-sample 오류율을 연결해 줄 뿐이다. 따라서 out-of-sample 오류율을 낮추는 일을 다음과 같이 두 단계로 나눌 수 있다.

1. E_{in}을 충분히 작게 만들기
2. E_{out}과 E_{in}을 충분히 가깝게 만들기

이렇게 문제를 쪼개 보면 많은 통찰을 얻을 수 있다. 그 한 예로 가설 집합의 복잡도가 갖는 트레이드오프에 대해 이해할 수 있게 된다. 아주 복잡하고 강력하며, 크기 M도 큰 가설 집합을 사용하면 E_{in}을 원하는 만큼 줄일 수 있을 것이다. 그러나 가설 집합의 크기가 크므로, 결과적으로 E_{out}이 E_{in}에 가까울 확률을 떨어뜨린다. 또 다른 예는 복잡한 목적 함수이다. 목적 함수는 문제 2를 어렵게 하진 않지만, 문제 1을 풀기 어렵게 만든다. 문제 1을 풀기 위해서 더 복잡한 가설 집합을 사용하면, 결과적으로 문제 2를 풀기 어려워지게 된다.

2 훈련과 테스트

이 챕터에서는 가설의 일반화(generalization) 속성, 즉 훈련 데이터에서 배운 것을 어떻게 테스트 데이터에 적용할 수 있는가를 나타낼 수 있는 여러 도구들에 대해 소개하고, 직관적인 이해를 돕는다.

2.1 합계 상한 고치기

최종 가설 g에 Hoeffding 부등식을 적용할 수 없었던 것은, 우리가 데이터를 본 후에야 g가 정해지기 때문이었다. 이 문제를 해결하기 위해, 합계 상한은 사건 $\mathcal{B}_i$를 i번 가설의 out-sample 오류율이 예상에서 크게 벗어나는 사건 $|E_{in}(h_i) - E_{out}(h_i) > \epsilon|$으로 둔 뒤, 다음과 같은 상한을 이용한다.

$$\mathrm{P}[\mathcal{B}_1 \vee \cdots \vee \mathcal{B}_M] \leq \sum_i \mathrm{P}[\mathcal{B}_i]$$

이 상한의 문제는 무엇일까? $\mathcal{B}_i$ 들의 사건 공간이 대개 겹친다면 우변의 상한이 좌변에 비해 너무 커져 버리고, 제대로 된 상한으로써의 의미를 잃게 된다. 그런데 이것은 많은 현실 세계의 문제에 대해서도 성립한다. 선형 회귀를 생각해 보면, 계수가 아주아주 조금 다른 수많은 다른 가설들이 있지만, 실질적으로 이들 간의 차이는 거의 없다.

2.2 증가 함수 (The Growth Function)

2.2.1 직관적 이해

교과서에서 다루는 분석은 오직 이진 분류 문제만을 다루고 있다. 강의에서는 이와 같은 방식의 접근으로 회귀 분석도 다룰 수 있다고 언급하는데, 논증 과정이 쓸데없이 복잡해 다루지 않는다고 한다.

우리는 이 절에서 증가 함수를 정의하는데, 이것은 합계 상한에서 가설의 수 M을 대체하게 될 함수로 가설 크기의 속성에 따라 변화한다. 그 정의는 약간 기계적이지만, 직관적으로는 위에서 말한 대로 서로 다른 가설들이라도 실제로 분류 결과가 다르지 않다면 의미가 없다는 점을 이용한다. 간단한 예를 들어 설명하자. 실수 집합 위의 점들을 분류하려 하는데, 우리의 가설들은 역치 a에 대해 다음 형태라고 하자:

$$h_a(x) = \begin{cases} 1 & \text{if } x \geq a \\ -1 & \text{otherwise} \end{cases}$$

가능한 역치는 무한히 많으므로, 우리가 사용 가능한 가설도 무한히 많다. 그러나 임의의 훈련 데이터가 주어질 때, 이 중 엄청나게 많은 수는 서로 다를 것이 없다. 훈련 데이터의 x가 $\{1, 4, 6, 8\}$라 하면, $a = 5$와 $a = 5.00001$ 은 분명 다르지만 결과적으로 각 입력들을 똑같이 분류하기 때문에, 훈련 데이터 입장에서는 서로 다를 것이 없다.

증가 함수는 서로 유의미하게 다른 가설들의 수를 센다. 따라서 직관적으로는 우리가 가설의 수 M을 증가 함수 $m_{\mathcal{H}}(N)$로 바꾸는 것이 아주 말 되는 전략으로 보인다.

2.2.2 기술적 정의

특정 형태의 가설 집합이 있다고 하자. 크기가 N인 훈련 데이터를 만들면, 각 점을 이진 분류할 때 가능한 결과값은 2^N가지가 있다. 훈련 데이터는 임의로 만들 수 있다고 가정할 때, 2^N개의 결과값 중 이 가설 집합으로 최대 몇 개를 표현할 수 있을까? 증가 함수는 이 값을 나타낸다. 기술적으로는 다음과 같이 쓸 수 있을 것이다.

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1, \cdots, \mathbf{x}_N} |\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N)|$$

이 때 $\mathcal{H}()$는 주어진 훈련 데이터 $\{\mathbf{x}_i\}$에 대해, $\mathcal{H}$ 의 모든 가설들을 이용해 만들 수 있는 결과값들의 수를 나타낸다. 위에서 설명한 1차원 분류 문제와 가설 집합을 가져오면, $m_{\mathcal{H}}(N) = N + 1$이 될 것이다.

2.2.3 Practicality of Calculating The Growth Function

Different hypothesis sets call for different growth functions, and it might not be practical to calculate them each time. Therefore we resort to calculating the upper bound of growth function. It is going to make a looser bound than the actual growth function, but then we can make general arguments.

This is where the notion of shattering comes handy. A hypothesis set $\mathcal{H}$ can shatter a particular set of size N, if it can generate all of the 2^N possible dichotomies for the given points.

We also say $N = k$ is a breaking point for the hypothesis set $\mathcal{H}$ when it cannot shatter any arrangement of unique N points. For example, a single perceptron in a 2D space can shatter some sets of 3 points, but cannot shatter any of the possible 4 points sets. Thus, $N = 4$ is a breaking point for the simple perceptron.

2.2.4 Bounding the Growth Function

Why is k important? It is used in calculating the upper bound of $m_{\mathcal{H}}(N)$ for arbitrary N. Sweet deal! Figuring this out involves introducing a new quantity $B(N, k)$ which can serve as the upper bound of $m_{\mathcal{H}}(N)$, and finding a recurrence on B and solving it. The end result:

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

which is a k-th order polynomial of N. Polynomial!!!

2.3 The VC Dimension

So this break point basically defines the generalization property of a given hypothesis set, so we decided to give it a more impressive-sounding name: the Vapnik-Chervonenkis Dimension d_{VC} of a hypothesis set is defined as the largest N that $\mathcal{H}$ can shatter any input of such size: so it is the break point - 1. Then, the growth function is a d_{VC}-th order polynomial of N:

$$m_{\mathcal{H}}(N) = O\left(N^{d_{VC}}\right)$$

So how do we use the VC dimension in the generalization bound? What we wanted to do was just replace M with $m_{\mathcal{H}}(N)$, so we can get bounds of the form

$$E_{out} \leq E_{in} + \sqrt{\frac{1}{2N} \ln \frac{2m_{\mathcal{H}}(N)}{\delta}}$$

This bound doesn't exactly hold, but it is true in spirit. What happens if $m_{\mathcal{H}}$ is finite and is a polynomial of N? That will be multiplied by a nontrivially large number $(1/\delta)$, but it will go through a log; so hopefully $2N$ will dominate that $O(k \lg N)$ term when N is large, and we get a decent generalization bound.

The actual bound is described in the following theorem:

VC Generalization Bound For any tolerance $\delta > 0$,

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(N)}{\delta}}$$

with probability $\geq 1 - \delta$.

What does this bound say? For a hypothesis set with a finite VC dimension, E_{out} will eventually converge to E_{in} as N increases, even the cardinality of the set is infinite!

2.3.1 Practical Considerations

The VC dimension does give a bound, but it is a quite loose bound. For example, take a very simple model with growth function $m_{\mathcal{H}}(N) = N + 1$. The VC dimension here is just 1, and what is the probability such that E_{out} will be within 0.1 of E_{in} given 100 training examples? Solving the following

$$\sqrt{\frac{8}{100} \ln \frac{4 \cdot 101}{\delta}} = 0.1$$

for δ gives 356.52. How is that useful for a probability estimate? Not very much. :-p So this bound is not practically useful in general, but the text covers some useful rules of thumb:

- The analysis is equally loose for different types of hypothesis. It is (experimentally) observed that hypothesis sets with smaller VC dimension generalize better than ones with larger dimension, so we can use it as a first cut method to compare hypothesis sets.
- The text introduces another interesting rule of thumb. If your hypothesis set has d_{VC}, you better have at least $10 \times d_{VC}$. This is a rather small number, I guess.

2.3.2 Significance of Test Set

Of course, we have better ways to assess E_{out}. Validation errors are one. Also we can set aside a separate test set; since test set is not used in training, the effective hypothesis size is 1. So the simple Hoeffding bound can be applied, which will be a reasonably tight bound.

2.4 Bias-Variance Tradeoff

Another way to formalize E_{out} is called the Bias and Variance tradeoff. The text formulates this around a classification problem. The setup is as follows.

We believe the training data set, $\mathcal{D}$ is a random variable drawn randomly from the population. Therefore, the final hypothesis g resulted from training is also a random variable, and should be annotated by $\mathcal{D}$: let's call it $g^{(\mathcal{D})}$. So, what is the expected out-of-sample error

$$E_{out}\left(g^{(\mathcal{D})}\right) = \mathbb{E}_{\mathbf{x}}\left[\left(g^{(\mathcal{D})}\left(\mathbf{x}\right) - f\left(\mathbf{x}\right)\right)^2\right]$$

with respect to $\mathcal{D}$?

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[E_{out}\left(g^{(\mathcal{D})}\right)\right] &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{x}}\left[\left(g^{(\mathcal{D})}\left(\mathbf{x}\right) - f\left(\mathbf{x}\right)\right)^2\right]\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}\left(\mathbf{x}\right) - f\left(\mathbf{x}\right)\right)^2\right]\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}\left(\mathbf{x}\right)\right]^2 - 2\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}\left(\mathbf{x}\right)\right] f\left(\mathbf{x}\right) + f\left(\mathbf{x}\right)^2\right]
\end{aligned}$$

Note the term $\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}\left(\mathbf{x}\right)\right]$; this is the result from the average hypothesis. It is equivalent to generating lots of training sets, learning from individual test sets, and averaging the responses from different models. Let's call this hypothesis $\bar{g}\left(x\right)$. Then, we can rewrite above using $\bar{g}$ as follows:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[E_{out}\left(g^{(\mathcal{D})}\right)\right] &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}\left(\mathbf{x}\right)\right]^2 - 2\bar{g}\left(\mathbf{x}\right) f\left(\mathbf{x}\right) + f\left(\mathbf{x}\right)^2\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}\left(\mathbf{x}\right)\right]^2 - \bar{g}\left(\mathbf{x}\right) + \bar{g}\left(\mathbf{x}\right) - 2\bar{g}\left(\mathbf{x}\right) f\left(\mathbf{x}\right) + f\left(\mathbf{x}\right)^2\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\underbrace{\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}\left(\mathbf{x}\right) - \bar{g}\left(\mathbf{x}\right)\right]^2}_{\text{variance}} + \underbrace{\left(\bar{g}\left(\mathbf{x}\right) - f\left(\mathbf{x}\right)\right)^2}_{\text{bias}}\right]
\end{aligned}$$

which shows the bias-variance decomposition.

2.4.1 Interpretation of Bias-Variance

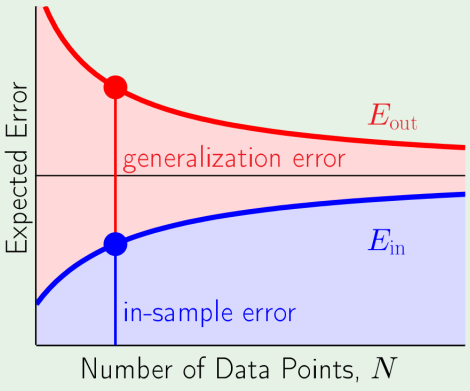
What are the possible interpretations?

- Bias represents the squared error between the "average" hypothesis and the target function. So it represents how much we are going to be off from the target, in general. In some sense, they measure the flexibility of the learning model to fit the target function.
- Variance represents how much a given $g^{(\mathcal{D})}$ is going to be off from the "average" hypothesis. It represents how our final hypothesis is vulnerable to changes in $\mathcal{D}$. In some sense, they measure the stability of the learning model.

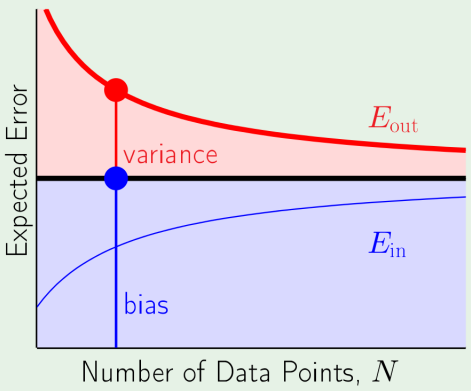
A simple model of course has a high bias; but it will have lower bias. A complex model will have a higher variance, but lower variance. As we increase the size of the sample, variance will slowly decrease and eventually bias will dominate E_{out}.

2.5 Comparison of VC Analysis and Bias-Variance Analysis

Following figure, stolen from the lecture slides, illustrates beautifully how the two analysis splits E_{out}.



VC analysis



bias-variance