21.05.30 Bayesian Optimization란?

노트북:	[Machine learning]		
만든 날짜:	2021-05-30 오후 5:55	**수정한 날짜:**	2021-05-30 오후 9:58
작성자:	황인범		
URL:	https://wooono.tistory.com/102		

베이지안 최적화란?

베이지안 최적화는 알려져 있지 않은 목적함수를 최대화(혹은 최소화)로 하는 최적해를 찾는 기법으로 지금까지 확보된 데이터화 평가지표의 숨겨진 관계를 모델링하는 Surrogate model과 Surrogate Model를 활용해 다음 탐색지점을 결정하는 Acquisition function으로 구성되어 있습니다. 기본적으로 Surrogate Model은 가우시안 프로세스를 거쳐 만들어지며, Acquisition function은 maximize expected improvement하며 학습이 진행됩니다.

*"**Bayesian Optimization**의 핵심은 **사전 정보를 최적값 탐색에 반영하는 것**이다!"*

Bayesian Optimization에서 사전 정보를 바탕으로 탐색하기 위해선 다음과 같은 정보가 필요하다.

1.
어떻게 모델 내에서 사전 정보를 학습하고 자동적으로 업데이트할까?

- 정답: **Surrogate Model**
- **기존 입력값(x1, f(x1)), (x2, f(x2)), ... , (xt, f(xt))들을 바탕**으로, **미지의 목적 함수 f의 형태에 대한 확률적인 추정**을 하는 모델

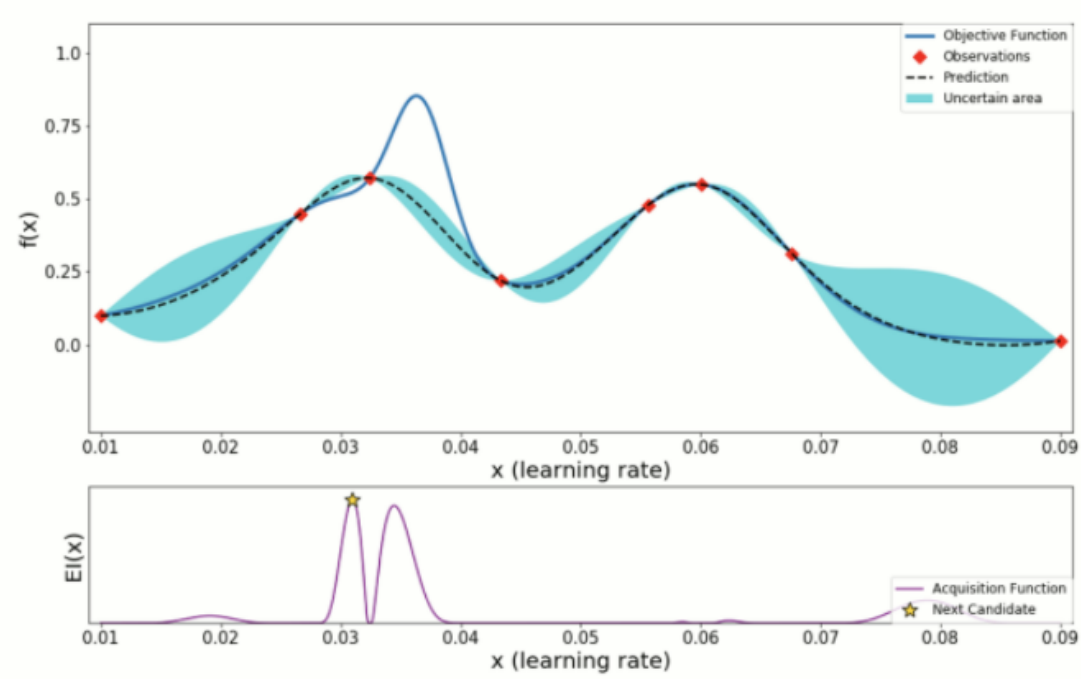
2.
수집한 사전 정보를 바탕으로 어떤 기준으로 다음 탐색값을 찾을까?

- 정답: **Acquisition Function**
-

Surrogate Model이 목적 함수에 대해 확률적으로 추정한 결과를 바탕으로, 바로 **다음 번에 탐색할 입력값 후보를 추천**해 주는 함수

- **최적값일 가능성이 높은 값 = Surrogate Model에서 함수값이 큰 값**
- **아직 Surrogate Model에서의 추정이 불확실한 값 = Surrogate Model에서 표준편차가 큰 값**

Bayesian Optimization 수행 과정



수행 과정 그림

위의 **파란색 선**은 우리가 찾으려고 하는 목적함수(unknown black box function)를 나타내고,

검정색 점선은 지금까지 관측한 데이터를 바탕으로 우리가 **예측**한 estimated function을 의미한다.

검정색 점선 주변에 있는 **파란 영역**은, 목적함수 f(x)가 존재할만한 confidence bound(function의 variance)를 의미한다.

밑에 있는 **EI(x)**는 위에서 언급한 **Acquisition function**을 의미하며 **다음 입력값 후보를 추천**해준다.

Acquisition function 값이 컸던 지점의 function 값을 관측하고 **estimation을 update**한다.

계속 update를 진행하면 **estimation**과 **실제 fuction**이 **흡사**해진다.

관측한 지점 중 **best point**을 **argmin f(x)**로 선택한다.

자세한 수행 과정

입력값, 목적 함수 및 그 외 설정값들을 정의합니다.

- 입력값 x : 여러가지 hyperparameter
- 목적 함수 f(x) : 설정한 입력값을 적용하여 학습한 딥러닝 모델의 검증 데이터셋에 대한 성능 결과 수치(e.g. 정확도)
- 입력값 의 탐색 대상 구간 : (a,b)
- 맨 처음에 조사할 입력값-함숫값 점들의 갯수 : n
- 맨 마지막 차례까지 조사할 입력값-함숫값 점들의 최대 갯수 : N

설정한 탐색 대상 구간 (a,b) 내에서 처음 n 개의 입력값들을 랜덤하게 샘플링하여 선택합니다.

선택한 n 개의 입력값 x1,x2,...,xn 을 각각 학습률 값으로 설정하여 딥러닝 모델을 학습한 뒤, 검증 데이터셋을 사용하여 학습이 완료된 모델의 성능 결과 수치를 계산합니다. 이들을 각각 함숫값 f(x1),f(x2),...,f(xn) 으로 간주합니다.

입력값-함숫값 점들의 모음 (x1,f(x1)),(x2,f(x2)),...,(xn,f(xn)) 에 대하여 Surrogate Model로 확률적 추정을 수행합니다.

조사된 입력값-함숫값 점들이 총 N 개에 도달할 때까지, 아래의 과정을 t=n,n+1,...,N−1 에 대하여 반복적으로 수행합니다.

- 기존 입력값-함숫값 점들의 모음 (x1,f(x1)),(x2,f(x2)),...,(xt,f(xt)) 에 대한 Surrogate Model의 확률적 추정 결과를 바탕으로, 입력값 구간 (a,b) 내에서의 EI의 값을 계산하고, 그 값이 가장 큰 점을 다음 입력값 후보 xt+1 로 선정합니다.
- 다음 입력값 후보 xt+1 를 학습률 값으로 설정하여 딥러닝 모델을 학습한 뒤, 검증 데이터셋을 사용하여 학습이 완료된 모델의 성능 결과 수치를 계산하고 이를 f(xt+1) 값으로 간주합니다.

새로운 점 (xt+1,f(xt+1)) 을 기존 입력값-함숫값 점들의 모음에 추가하고, 갱신된 점들의 모음에 대하여 Surrogate Model로 확률적 추정을 다시 수행합니다.

총 N 개의 입력값-함숫값 점들에 대하여 확률적으로 추정된 목적 함수 결과물을 바탕으로, 평균 함수 μ(x)을 최대로 만드는 최적해를 최종 선택합니다. 추후 해당값을 학습률로 사용하여 딥러닝 모델을 학습하면, 일반화 성능이 극대화된 모델을 얻을 수 있습니다.

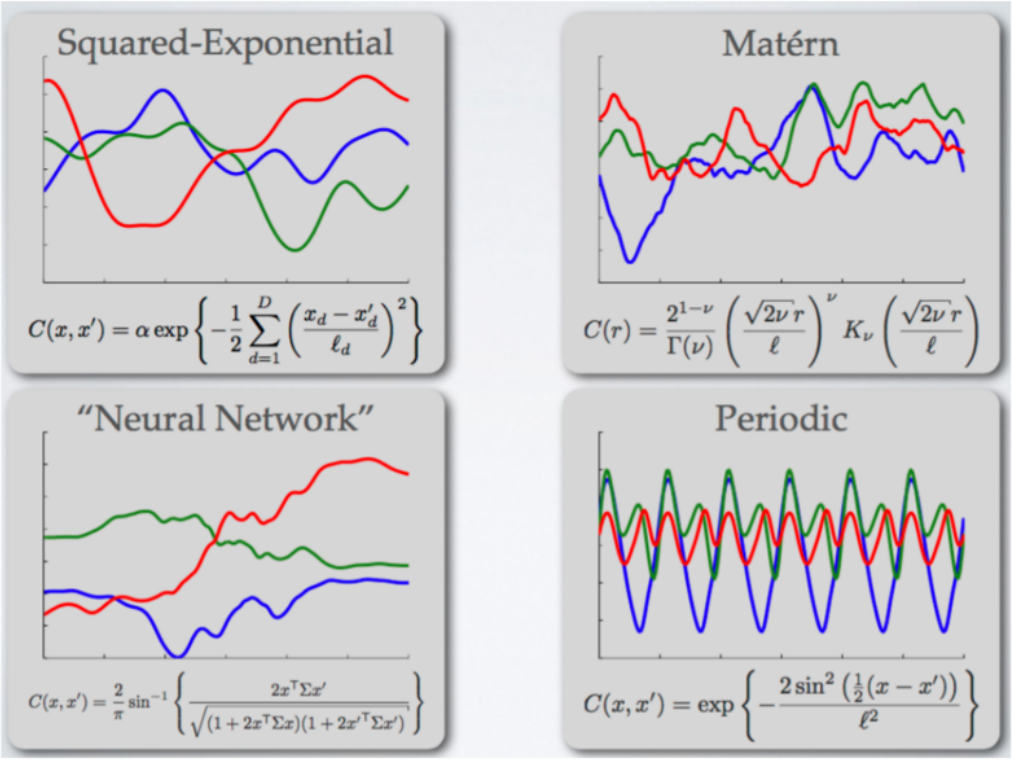
Surrogate Model

Surrogate Model로 여러가지 모델이 사용될 수 있는데,
가장 많이 사용되는 확률 모델은 Gaussian Process (GP) 이다.
Gaussian Process에 대해서 설명하자면 끝도 없는데,
직관적으로 이해하자면 일종의 함수이고,
각 입력에 대해서 normal distribution random variable을 출력하는 함수라고 생각하면 된다.
(radom variable은 각 입력에 대해 real value를 출력하는 함수라는 사실을 인지해야 한다.)
수학적인 증명은 다 떼고, 여기서 핵심 개념은 이거다.

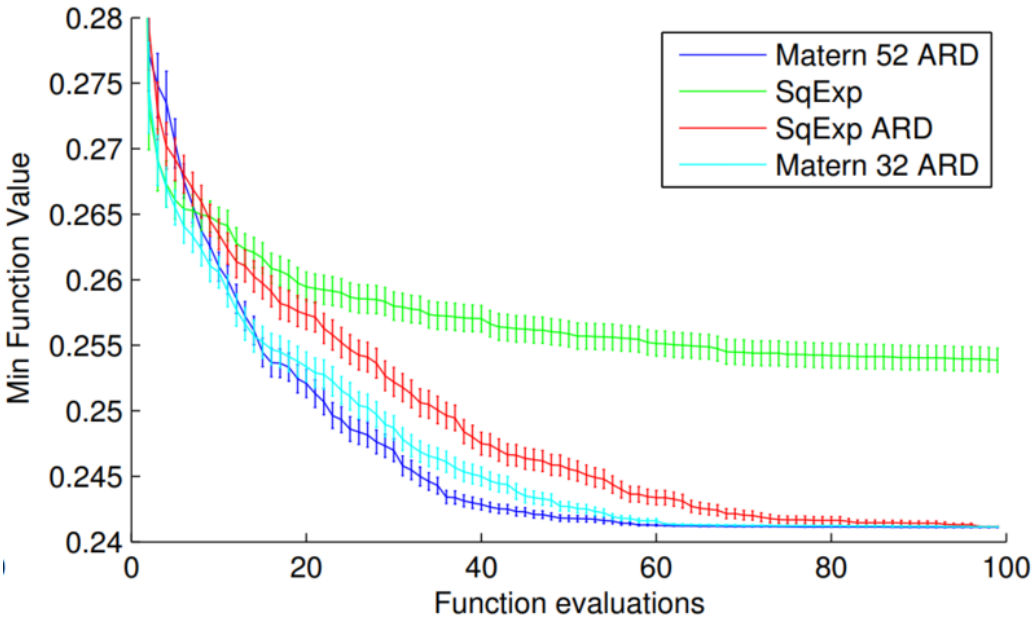
$$f(x) \sim \mathcal{GP}(\mu(x), k(x, x')).$$

Gaussian Process는 mean function과, covariance function으로 정의가 되는데,
이 두가지가 정해지고 나면, 임의의 모든 점에 대한 확률분포를 얻어낼 수 있다.
논문에 따르면 mean function은 별로 중요하지 않고 covariance function이 중요하다고 한다.
여기서 covariance function은 kernel function을 통해 얻어내게 되는데,
kernel function은 주어진 GP sample 들이 어떤 relationship을 가지는지 정의하는 함수이다.
따라서 k(x, x')는 두 점에 대해 정의되게 되는데, 일반적으로 점 사이의 거리가 가까우면 relationship이 크고

멀면 relationship이 작을 것이라는 가정을 한다고 한다.

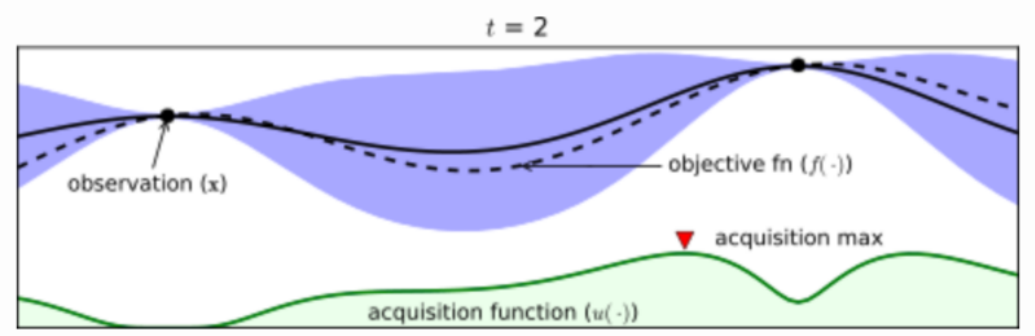


일반적으로 많이 쓰이는 kernel function은 Squeared-Exponential 이라는 함수지만 이는 너무 smooth하다는 단점이 있고,
한 논문(J Snoek et al. 2012)에 따르면 matern 5/2 를 사용했을 때 가장 좋다고 한다.
5/2란 matern 함수의 각 parameter에 5와 2를 사용한 것을 뜻한다.



Acquisition Function

Surrogate Model이 목적함수에 대해 확률적으로 추정한 결과를 바탕으로 다음번에 조사할 만한 입력값을 추천해주는 함수이다.
이 때 추천되는 입력값은 '최적 입력값을 찾는데 있어 가장 유용할 만한' 것이다.
'가장 유용할 만한'의 의미는 무엇일까?



이전의 그림으로 돌아와서, 현재까지 조사된 (x, f(x)) 점은 두개이고,
오른쪽 점에서 최대값이 나타났다.
이 데이터를 토대로 '오른쪽 점 근처에 최적값이 존재할 것이다' 라고 예측하는 것은 어느 정도 그럴싸해보인다.
이를 exploitation 이라 한다.
한편으로는 표준편차가 가장 큰 점, 즉 '불확실성이 가장 높은 점 근처에 최적값이 존재할 것이다.'라고 예측하는 것도 그럴싸해보인다.
이를 exploration 이라 한다.
그런데 이미 조사된 최대값 근처의 point에서는 표준편차가 작게 나타나고,
멀어질 수록 표준편차가 크게 나타나므로
exploitation과 exploration는 trade-off의 관계에 있다고 할 수 있다.
따라서 좋은 Acquisition Function은 이러한 exploitation과 exploration의 상대적 강도를 적절히 고려하여, 이를 균형있게 반영할 수 있어야 한다.
이를 위한 함수가 여러가지 있는데, 가장 널리쓰이는 EI에 대해 알아볼 것이다.

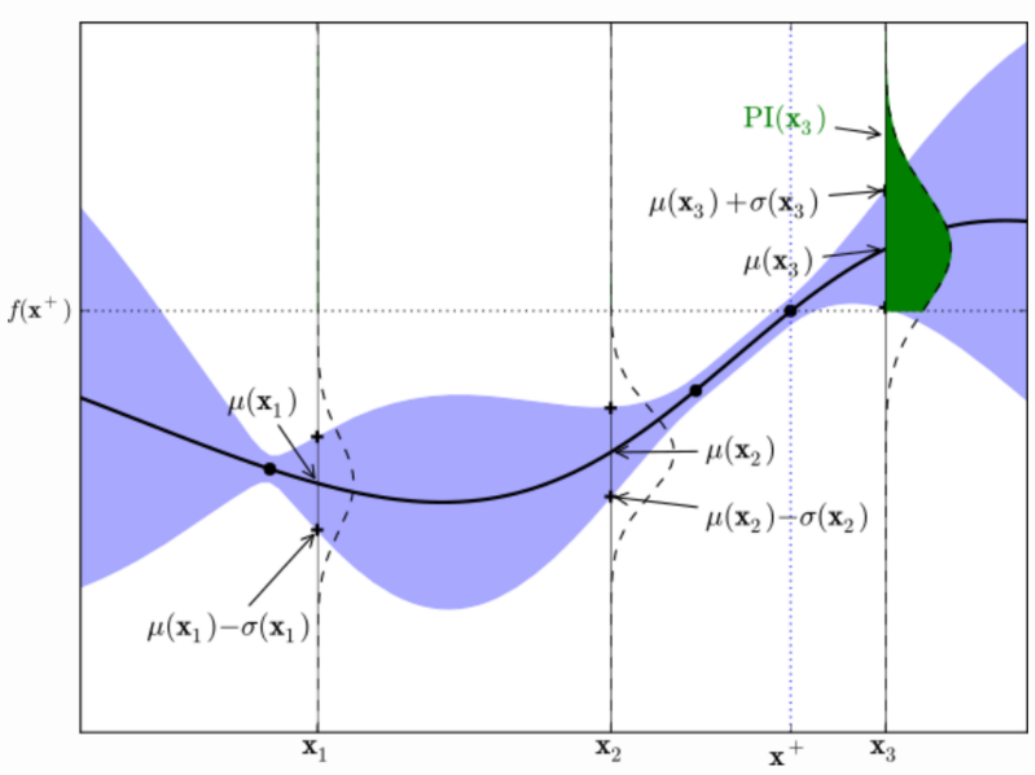
Probability of Improvement (PI)

PI는 어떤 후보 입력값이 기존의 최대값보다 클 확률을 계산한다.

$$\alpha_{\mathrm{PI}}(\mathbf{x};\mathcal{D}_n) := \mathbb{P}[v > \tau] = \Phi\left(\frac{\mu_n(\mathbf{x}) - \tau}{\sigma_n(\mathbf{x})}\right)$$

Expected Improvement(EI)

EI 함수는 exploitation과 exploration를 내재적으로 잘 고려하도록 설계된 함수로서, Acquisition Function으로 가장 많이 사용된다.



그림을 보면, 3가지 점들이 조사되었고, 이 중 최대값을 가지는 점은 x+ 이다.
EI는 x1, x2, x3 에 대해 조사를 수행하는데,
1. 각 후보의 함수값이 현재까지 조사된 최대값 보다 클 확률이 얼마나 되는가
2. 그러한 확률이 존재한다면, 얼마나 더 클 것인가
를 종합적으로 고려하게 된다.
1번은 Probability of Improvement라 하는데, 그림에서 표시된 녹색 CDF영역이 된다.
2번으로는 f(x3) - f(x+) 의 기댓값을 반영하게 된다.
복잡한 유도과정을 거쳐서 수식으로 나타내면 다음과 같다고 한다.

$$
\begin{aligned}
EI(x) &= \mathbb{E}[\max(f(x) - f(x^+), 0)] \\
&= \begin{cases} (\mu(\boldsymbol{x}) - f(\boldsymbol{x}^+) - \xi)\Phi(Z) + \sigma(\boldsymbol{x})\phi(Z) & \text{if } \sigma(\boldsymbol{x}) > 0 \\ 0 & \text{if } \sigma(\boldsymbol{x}) = 0 \end{cases}
\end{aligned}
$$

$$
Z = \begin{cases} \frac{\mu(\boldsymbol{x}) - f(\boldsymbol{x}^+) - \xi}{\sigma(\boldsymbol{x})} & \text{if } \sigma(\boldsymbol{x}) > 0 \\ 0 & \text{if } \sigma(\boldsymbol{x}) = 0 \end{cases}
$$

솔직히 수식이 어려워서 잘 모르겠고
그냥 직관적으로만 살펴보면,
후보 x에 대한 EI는 분산이 0일 경우에는 곧 이미 조사된 점이라는 소리이므로 기대되는 개선확률은 0이되고
그렇지 않을 경우 두가지의 term으로 구성되는데
첫번째 term의 경우 x에 대한 평균값이 기존 최대값보다 얼마나 더 클 것인가를 고려하므로 exploitation 에 해당하고
두번째 term의 경우 x에 대한 분산을 고려하고 있으므로 exploration에 해당한다.

ξ는 이 둘사이의 강도를 조절하는 parameter로서, 값이 커질수록 exploration에, 값이 작아질수록 exploitation에 비중을 두게 된다.
(일반적으로 0.01)

결론

Bayesian Optimization은 실험결과를 반영해가면서 효율적으로 hyperpameter를 찾을 수 있는 알고리즘이다.
이는 Surrogate Model과 Acquisition Function으로 구성된다.

최적화 하는데 있어서 최소화하려는 함수 $f(x)$를 **목적함수(objective function)**, **비용함수(cost function)**, **손실함수(loss function) 오차함수(error function)** 등으로 부른다.

Surrogate Model은 목적함수의 형태를 추정하기 위한 모델이며, Gaussian Process를 주로 사용한다.
Acquisition Function은 목적함수를 잘 추정하기 위한 입력값을 추천해주는 함수이며, Expected Improvement를 주로 사용한다.