



# HPO 3부) AutoML, 앙상블 기반 최적화

[자동 머신 러닝 \(AutoML\)](#)

[AutoML 종류 및 특징](#)

[AutoML의 기술](#)

[결론](#)

[앙상블 기반 최적화 \(Ensemble-Based Optimization\)](#)

[유추 기반 최적화 \(Deductive Optimization\)](#)

## 자동 머신 러닝 (AutoML)

AutoML은 머신 러닝 모델을 학습하고, 배포하는 과정을 자동화하는 기술 혹은 도구를 말한다. 즉, 머신 러닝 모델을 구축하는 전체 과정을 자동화하는 기술이다. AutoML은 하이퍼파라미터 튜닝 뿐만 아니라, 머신러닝 프로세스인 전처리, 모델링, 후처리와 모델 선택, 앙상블 등의 과정도 자동화할 수 있다. 최적화의 경우, AutoML은 자동으로 최적의 하이퍼파라미터 조합을 찾아줌으로써, 머신 러닝 모델을 더 빠르고 쉽게 개발할 수 있도록 도와준다.

AutoML의 하이퍼파라미터 튜닝 과정은 다음과 같다.

1. 탐색 대상 하이퍼파라미터 공간 정의: 탐색할 하이퍼파라미터의 범위 지정
2. 초기 하이퍼파라미터 설정: 일부 초기 하이퍼파라미터를 임의로 선택
3. 모델 학습 및 평가: 초기 하이퍼파라미터를 사용하여 모델을 학습하고 검증 데이터에서 성능을 평가
4. 하이퍼파라미터 탐색: 다양한 하이퍼파라미터 값을 시도하여 최상의 성능을 얻는 하이퍼파라미터 조합을 찾는다. 이 과정에서 그리드 서치, 랜덤 서치, 베이지안 최적화 등 다양한 방법 사용 가능
5. 최적 하이퍼파라미터 선택: 검증 데이터에서 가장 높은 성능을 보인 하이퍼파라미터 조합을 선택
6. 모델 재학습 및 평가: 선택한 하이퍼파라미터 조합을 사용하여 모델을 재학습하고 테스트 데이터에서 성능 평가

AutoML 프레임워크는 이러한 과정을 자동으로 수행하며, 각 단계에서 여러 모델을 학습 및 평가하여 최상의 모델과 하이퍼파라미터 조합을 찾는다. 이러한 AutoML 프레임워크는 개발자가 하이퍼파라미터 튜닝에 대한 전문 지식이 없어도 높은 성능의 머신러닝 모델을 개발할 수 있도록 도와준다.

## AutoML 종류 및 특징

### 1) OSS (Open Source Software)

OSS 방식은 오픈소스 기반으로 라이브러리 형태로 제공되는 AutoML 기능을 뜻한다. 예를 들어 Scikit-learn 기반의 Auto-sklearn, TPOT, Keras 기반의 AutoKeras, Tensorflow 기반의 AdaNet 등이 있다. Scikit-Learn 기반의 AutoML 라이브러리는 구조화된 데이터에 보다 적합하고 기존 Scikit-learn의 전처리 방식을 자동화하여 적용 가능하다.

OSS 방식의 장단점은 다음과 같다.

- 장점: 기존 코드에 AutoML 라이브러리를 호출하여 사용할 수 있어 커스터마이징이 자유로움
- 단점: 높은 컴퓨팅 자원 활용이 가능해야 하며, 후퍼리 단계(모델 평가, 배포, 모니터링 등)에 대한 기능 지원이 아직 미흡

```
# Auto-sklearn classifier
import autosklearn.classification

clf = autosklearn.classification.AutoSklearnClassifier()
clf.fit(X_train, y_train)
results = clf.predict(X_test)
```

<https://github.com/automl/auto-sklearn>

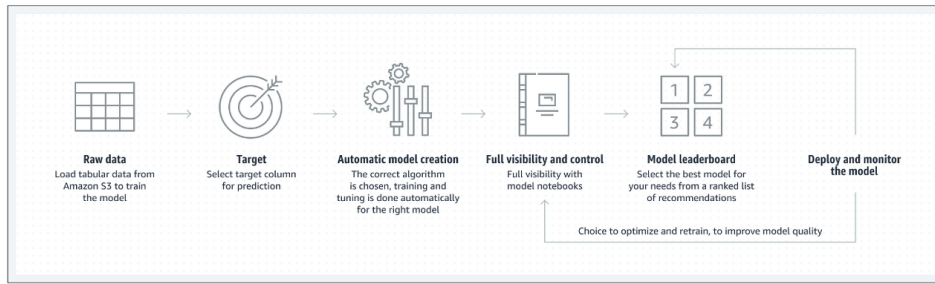
## 2) Cloud Provider Solution

주요 클라우드 서비스는 클라우드 환경에서 이용 가능한 AutoML 솔루션을 제공한다. 대표적으로 Google의 CloudML, 아마존의 Sagemaker Autopilot, Microsoft의 Azure ML이 있다. 이들은 코드 작성 없이 UI 방식과 Python의 API 방법을 둘 다 제공한다.

Cloud Provider Solution의 장단점은 다음과 같다.


- 장점:
  - 클라우드 환경에서 제공되기 때문에 필요한 만큼의 자원 할당이 가능하고, 자원 관리 서비스를 제공
  - 전처리부터 모델링, 결과 평가, 모델 배포까지 기능을 제공하여 전체 프로세스에 대한 구현 가능
- 단점: 사용자가 구현된 시스템 내부를 파악하기는 힘들 (기업 비밀)

예로 Amazon의 Sagemaker Autopilot에서 제공하는 세부 기능은 다음과 같다.



### AutoML - 자동화된 기계 학습 - Amazon Web Services

AutoML을 사용하여 완전한 제어와 가시성을 유지 관리하는 동시에 기계 학습 (ML) 모델을 자동으로 훈련 및 튜닝합니다.

 <https://aws.amazon.com/ko/sagemaker/autopilot/?sagemaker-data-wrangler-whats-new.sort-by=item.additionalFields.postDateTime&sagemaker-data-wrangler-whats-new.sort-order=desc>



### 3) Enterprise solution

AutoML은 서비스 제공을 목적으로 만들어진 DataRobot과 H2O와 같은 AutoML 플랫폼들이 존재한다. 해당 플랫폼은 AutoML에 특화된 기능과 각 프로세스에서 커스터마이징이 가능하도록 구현되어있다. 또한, AutoML 플랫폼 기업 제품들은 원하는 곳에 설치 가능하므로 민감한 데이터를 다루는 경우에는 적합한 방식이지만, 초기 라이선스 구매 비용이 큰 부담이 될 수 있다.

### AutoML의 기술

AutoML의 핵심 문제는 데이터가 주어졌을 때 가장 적절한 모델과 파라미터를 선택하는 CASH (Combined Algorithm Selection and Hyper-Parameter) 최적화 문제이다. CASH란 최적화 기계 학습 모델(algorithm) 및 해당 하이퍼 파라미터를 탐색하는 방법이다. 이 검색 공간은 대부분의 경우 비연속적이기 때문에, 높은 계산 비용을 가진다. 예를 들어, 그리드 서치의 경우, 모든 하이퍼 파라미터 조합을 전부 조사하는 방법은 비효율적이다.

Auto-WEKA라는 시스템을 개발한 논문에서는 데이터(D)가 주어졌을 때, 최적의 알고리즘 A와 람다를 선택하는 문제로 다음과 같이 표기했다.

$$A_{\lambda}^* = \arg \min_{A_{\lambda}^j \in A, \lambda \in \lambda^j} \frac{1}{k} \sum_{i=1}^k L(A_{\lambda}^j, D_{train}^{(i)}, D_{test}^{(i)})$$

CASH 알고리즘의 문제는 핵심 탐색 공간을 어떻게 설정하느냐에 있다. 실제로 모든 모델과 파라미터를 다 비교하는 것은 불가능하고, 어느 정도 경험치를 바탕으로 가능성이 없어보이는 모델을 탐색하지 않는 것이 핵심이다. 그리고 이것이 AutoML이 데이터 분석가를 완전히 대체할 수 없는 이유이다.

### 결론

정리하자면, AutoML은 쉽고 빠르게 모델 생성을 도와주는 기능으로 모델 개발 프로세스 효율을 향상시켜줄 수 있다. 기본적으로 하이퍼파라미터 튜닝 기능을 제공하기 때문에 알고리즘 별 하이퍼파라미

터에 대한 이해가 덜 요구된다. 하지만, 데이터 전처리 단계의 경우 인간의 개입이 필요한 부분이 많고, 배경지식이 여전히 중요하다는 사실은 변하지 않는다.

## 앙상블 기반 최적화 (Ensemble-Based Optimization)

앙상블 기반 최적화란 여러 하이퍼파라미터 튜닝 알고리즘을 결합하여 사용하는 방법이다. 예를 들어, 랜덤 서치와 그리드 서치를 결합하여 하이퍼파라미터 공간을 더 효과적으로 탐색할 수 있다. 이는 여러 개의 모델을 조합하여 최적화하기 때문에, 개별 모델의 성능을 향상시키고, 모델 간의 다양성을 증가시켜 전체적인 성능을 개선하는 데 도움을 준다.

앙상블 기반 최적화의 방법은 다음과 같다.

1. 다수결 투표 방식(Voting): 여러 모델의 예측 결과를 종합하여 최종 예측을 만드는 것으로, 분류 문제에서 다수결 투표 방식을 사용하면, 각 모델이 예측 한 클래스 중 가장 많이 투표를 받은 클래스가 최종 예측 결과가 된다.
2. 가중 평균 방식: 각 모델의 예측 결과에 가중치를 부여하여 종합한 값을 최종 예측 결과로 사용하며, 이때, 가중치는 각 모델의 성능이나 신뢰도 등을 고려하여 선정된다.

앙상블 기반 최적화는 하이퍼 파라미터 튜닝과 함께 사용되기도 한다. 앞선 방식과 동일하게, 여러 개의 모델을 학습하고, 각 모델의 하이퍼파라미터를 다르게 설정하여 최적의 하이퍼 파라미터 조합을 찾는데, 이러한 방법은 모델 간의 다양성을 증가시키고, 최적의 성능을 얻는 데 도움이 된다.

앙상블 기반 최적화는 머신 러닝 모델의 성능을 향상시키는 데 유용한 기술이지만, 다음과 같은 단점을 가지고 있다.

1. 여러 모델을 사용하므로 학습 및 예측 속도가 느려질 수 있다
2. 모델 간의 다양성을 유지하기 위해 여러 모델을 학습해야 하므로 모델의 개수가 증가할 수록 계산 비용이 증가

따라서, 앙상블 기반 최적화를 사용할 때는 이러한 단점을 고려해서 적절하게 선택해야 한다.

## 유추 기반 최적화 (Deductive Optimization)

유추 기반 최적화는 특정 문제에 대한 선호도를 유추하는 기술로, 이를 기반으로 하이퍼파라미터를 자동으로 조정하는 방법이며, 대규모 머신 러닝 시스템에서 특히 효과적이다. 유추 기반 최적화는 일반적으로 선호도 측정 기법과 검색 알고리즘을 포함한다. 이때, 선호도 측정 기법은 특정 하이퍼파라미터 조합에 대한 성능을 측정하고, 검색 알고리즘은 이러한 선호도 측정 결과를 사용하여 하이퍼파라미터 공간을 탐색하고 최적의 하이퍼파라미터 조합을 찾는다.

유추 기반 최적화의 핵심 아이디어는, 이전 실험에서 얻은 선호도를 바탕으로 다음 실험에서 시도할 하이퍼파라미터 조합을 선택하는 것입니다. 즉, 이전 실험에서 잘 작동한 하이퍼파라미터 조합은 다음 실험에서 더 자주 시도되는 것입니다.

이러한 유추 기반 최적화 기술은 기존 방법들과는 달리, 최적의 하이퍼파라미터 조합을 찾는 데 매우 적은 수의 실험을 필요로 합니다. 그러므로, 매우 복잡한 머신 러닝 모델의 경우에도 매우 효과적인 방

법 중 하나입니다.

장점:

빠른 속도: 유추 기반 최적화는 이전 실험에서 얻은 선호도를 활용하여 하이퍼파라미터 공간을 빠르게 탐색할 수 있습니다. 그러므로 하이퍼파라미터 최적화에 소요되는 시간을 크게 줄일 수 있습니다.

자동화: 유추 기반 최적화는 대부분 자동화됩니다. 이를 통해 사람이 수동으로 하이퍼파라미터를 선택하는 과정을 건너뛸 수 있습니다.

복잡한 문제 해결: 유추 기반 최적화는 기존의 최적화 방법들보다 더 복잡한 문제를 해결하는 데 효과적입니다.

단점:

오버피팅: 이전 실험에서 얻은 선호도를 기반으로 다음 실험에서 시도할 하이퍼파라미터 조합을 선택하는 것은, 이전 실험의 데이터에 대해 오버피팅될 가능성이 있습니다.

선호도 측정 오류: 선호도 측정 방법에 대한 오류로 인해 최적의 하이퍼파라미터 조합을 찾지 못할 수 있습니다.

선호도 측정 기법에 대한 선호도에 따른 한계: 유추 기반 최적화에서 사용되는 선호도 측정 방법은, 특정 문제에 대해 항상 최상의 결과를 보장하지 않습니다. 또한, 하이퍼파라미터 공간이 너무 크거나, 문제가 너무 복잡한 경우에는 유추 기반 최적화가 제한될 수 있습니다.

따라서, 유추 기반 최적화는 빠른 속도와 자동화 등의 이점이 있지만, 일부 한계점이 있을 수 있습니다.

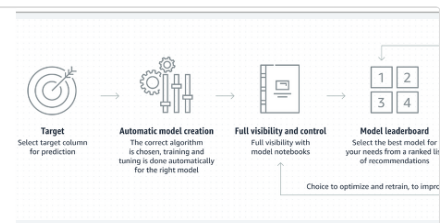
## ▼ 출처

### AutoML 이란? (종류 및 장단점)

이번 글에서는 머신러닝을 쉽고 빠르게 도입할 수 있도록 도와주는 AutoML (Automated Machine Learning) 에 대해 간단히 정리해보고자 한다.

AutoML 이란 무엇이고, AutoML 에는 어떤 종류가 있고, 종류별 특징 및 장

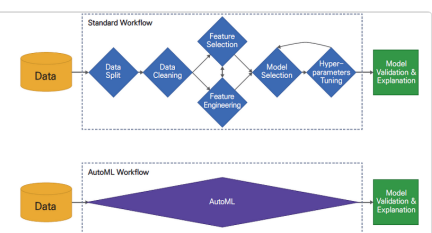
☞ <https://data-minggeul.tistory.com/12>



### [AutoML] AutoML이란 무엇인가?

이번에 회사 업무상 AutoML에 대해 정리해야 할 일이 있어, 온갖 자료를 다 끌어다가 정리해보고자 합니다. 특히, 제가 직접 AutoML 시스템을 개발해야 할 수도 있어서 상용화된 AutoML을 분석해보고, 구현을 위해 어떤 지식이

☞ <https://gils-lab.tistory.com/65>



### 한 줄의 코드로 자동학습! 머신러닝을 자동화하는 AutoML

3개의 요점 ✓ 테이블 데이터를 자동으로 학습하여 높은 성능을 발휘하는 AutoML 프레임워크 ✓ 기존 AutoML 프레임워크 모델이 하이퍼 매개변수의 선택을 중시하는 반면, 이 방법은 여러 레이어를 사용하여 모델의 앙상블

☞ <https://dooob.tistory.com/110>

