Bayesian Optimization for Hyperparameter Optimization

Success Stories

Bernd Bischl Frank Hutter Lars Kotthoff
Marius Lindauer Joaquin Vanschoren

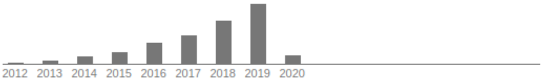
Spearmint [Snoek et al. 2012]

- First successful open source Bayesian optimization implementation
- Implements standard Bayesian optimization with MCMC integration of the acquisition function, asynchronous parallelism, input warping and constraints
- Startup based on Spearmint got acquired by Twitter in 2015
- Still heavily used and cited and available at `https://github.com/HIPS/spearmint`:

JasperSnoek / **spearmint** — Watch 101 | Star 1.3k | Fork 331

HIPS / **Spearmint** — Watch 84 | Star 1.3k | Fork 290

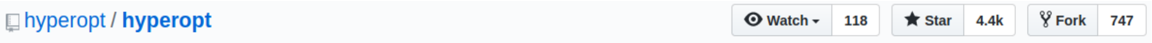
Cited by 3073



Practical bayesian optimization of machine learning algorithms
J Snoek, H Larochelle, RP Adams - Advances in neural information processing systems, 2012
Cited by 3073 Related articles All 26 versions

Hyperopt [Bergstra et al. 2011, Bergstra et al., 2013, Bergstra et al., 2013, Bergstra et al., 2015]

- Hyperopt is another successful open source Bayesian optimization package
- Implements the TPE algorithm and supports asynchronous parallel evaluations
- Maintained since 2013
- Available at `https://github.com/hyperopt/hyperopt`



SMAC [Hutter et al. 2011]

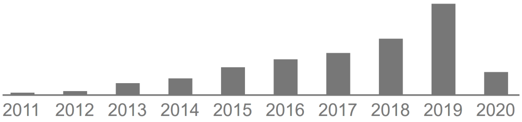
- Standard BO tool based on random forests (RFs), reflecting the strengths of RFs in terms of scalability & flexibility:
  - High dimensionality (low effective dimensionality)
  - Computational efficiency ($\rightarrow$ low overhead)
  - Supports continuous/categorical/conditional parameters
  - Supports non-standard noise (non-Gaussian, heteroscedastic)
  - Usability off the shelf (robustness towards model's own hyperparameters)

SMAC [Hutter et al. 2011]

- Standard BO tool based on random forests (RFs), reflecting the strengths of RFs in terms of scalability & flexibility:
  - High dimensionality (low effective dimensionality)
  - Computational efficiency ($\rightarrow$ low overhead)
  - Supports continuous/categorical/conditional parameters
  - Supports non-standard noise (non-Gaussian, heteroscedastic)
  - Usability off the shelf (robustness towards model's own hyperparameters)
- SMAC also handles a more general problem: $\arg\min_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} \sum_{i=1}^{N} c(\boldsymbol{\lambda}, i)$

SMAC [Hutter et al. 2011]

- Standard BO tool based on random forests (RFs), reflecting the strengths of RFs in terms of scalability & flexibility:
  - High dimensionality (low effective dimensionality)
  - Computational efficiency ($\rightarrow$ low overhead)
  - Supports continuous/categorical/conditional parameters
  - Supports non-standard noise (non-Gaussian, heteroscedastic)
  - Usability off the shelf (robustness towards model's own hyperparameters)
- SMAC also handles a more general problem: $\arg\min_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} \sum_{i=1}^{N} c(\boldsymbol{\lambda}, i)$
- Maintained since 2011, now available in version 3: `https://github.com/automl/SMAC3`



Sequential model-based optimization for general algorithm configuration
F Hutter, HH Hoos, K Leyton-Brown - International conference on learning and intelligent …, 2011

Tuning AlphaGo [Chen et al. 2018]

- “During the development of AlphaGo, its many hyperparameters were tuned with Bayesian optimization multiple times.”
- “This automatic tuning process resulted in substantial improvements in playing strength. For example, prior to the match with Lee Sedol, we tuned the latest AlphaGo agent and this improved its win-rate from 50% to 66.5% in self-play games. This tuned version was deployed in the final match.
- Of course, since we tuned AlphaGo many times during its development cycle, the compounded contribution was even higher than this percentage.

Company usage

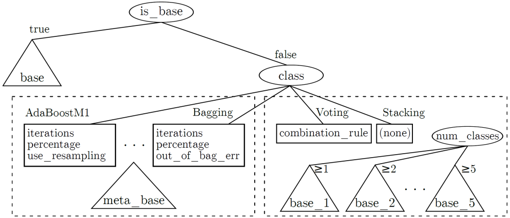
- SIGOPT: startup offering Bayesian optimization as a service
- Facebook provides an open source Bayesian optimization package [BoTorch]
- Amazon provides an open source Bayesian optimization package [EmuKit]
- Uber tunes algorithms for *Uber Pool*, *UberX* and *Uber Eats* [source]
- Many more, but less openly

Auto-WEKA [Thornton et al, 2013, Kotthoff et al, 2017, Kotthoff et al. 2019]

- First general AutoML system, carrying out **C**ombined **A**lgorithm **S**election and **H**yperparameter optimization (CASH), jointly optimizing
  - Choice of algorithm (out of 26 classifiers)
  - The algorithm's hyperparameters (up to 10)
  - Choice of preprocessing method and its hyperparameters
  - Choice of ensemble & meta methods

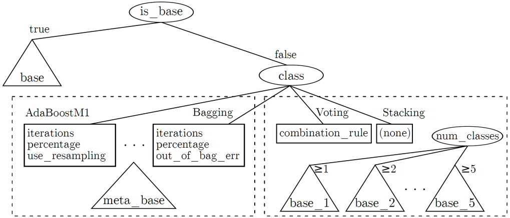
Auto-WEKA [Thornton et al, 2013, Kotthoff et al, 2017, Kotthoff et al. 2019]

- First general AutoML system, carrying out **C**ombined **A**lgorithm **S**election and **H**yperparameter optimization (CASH), jointly optimizing
  - Choice of algorithm (out of 26 classifiers)
  - The algorithm's hyperparameters (up to 10)
  - Choice of preprocessing method and its hyperparameters
  - Choice of ensemble & meta methods
- Parameterized WEKA [Frank et al, 2016]: 768 hyperparameters, 4 leves of conditionality



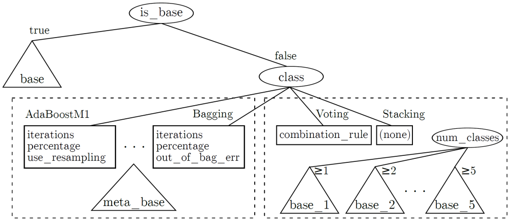
Auto-WEKA [Thornton et al, 2013, Kotthoff et al, 2017, Kotthoff et al. 2019]

- First general AutoML system, carrying out **C**ombined **A**lgorithm **S**election and **H**yperparameter optimization (CASH), jointly optimizing
  - Choice of algorithm (out of 26 classifiers)
  - The algorithm's hyperparameters (up to 10)
  - Choice of preprocessing method and its hyperparameters
  - Choice of ensemble & meta methods
- Parameterized WEKA [Frank et al, 2016]: 768 hyperparameters, 4 leves of conditionality
- Optimized 10-fold cross-validation via SMAC [Hutter et al, 2011]



Auto-WEKA [Thornton et al, 2013, Kotthoff et al, 2017, Kotthoff et al. 2019]

- First general AutoML system, carrying out **C**ombined **A**lgorithm **S**election and **H**yperparameter optimization (CASH), jointly optimizing
  - Choice of algorithm (out of 26 classifiers)
  - The algorithm's hyperparameters (up to 10)
  - Choice of preprocessing method and its hyperparameters
  - Choice of ensemble & meta methods
- Parameterized WEKA [Frank et al, 2016]: 768 hyperparameters, 4 leves of conditionality
- Optimized 10-fold cross-validation via SMAC [Hutter et al, 2011]



- Results:
  - Better than an oracle of the 26 base classifiers with default hyperparameters
  - 100× faster than grid search over base classifiers, and still better in 14/21 cases
  - Better than the only other applicable method TPE in 19/21 cases
- Impact for practitioners: Auto-WEKA plugin was downloaded tens of thousands of times

Questions to Answer for Yourself / Discuss with Friends

- Repetition. List several success stories of Bayesian optimization
- Repetition. List several prominent tools for Bayesian optimization
- Discussion. Recall the algorithm selection problem; how does CASH relate to this (after all, it also has “algorithm selection” as part of its name)? (Hint: they are quite different.)