Can Bayesian statistics help with hyperparameter tuning?

Bayesian Analysis
A.Y. 2023-2024

Davide Matta
Jing Wang

HYPERPARAMETER TUNING

- It influences the algorithm.
- The hyperparameters are set before the learning process.
- It refers to the selection of the best set of hyperparameters.
- It helps to find the model with lowest error or highest accuracy.
- Frequentist vs Bayesian based approaches

FREQUENTIST APPROACHES

GRID SEARCH

- It explores all possible combinations of the hyperparameters within the specified grid to find the best combination.
- It is used when the search space is relatively small and the hyperparameters are equally important.

RANDOM SEARCH

- It samples hyperparameters randomly.
- It is used when the search space is large or when certain hyperparameters are more influential than others.

Grid and Random search
<u>do not</u> pay attention to the past results.

BAYESIAN OPTIMIZATION

- **It keeps track of past evaluation results.**

- **The following surrogate function serves as a simplified approximation of the true objective function, which leads to find the highest accuracy :**

p(score|hyperparameters)

BAYESIAN OPTIMIZATION - SMBO

- Sequential model based global optimization methods are a formalization of Bayesian optimization.

- It involves the building a surrogate model to approximate the objective function and use it to guide the search for optimal hyperparameters.

BAYESIAN OPTIMIZATION - TPE

- **Tree structured Parzen Estimator is a specific type of surrogate model used in SMBO .**
- **It builds the model by applying directly the Bayes rule:**

$$P(y|x) = \frac{P(x|y) \times P(y)}{P(x)} \longrightarrow p(x|y) = \begin{cases} l(x) & \text{if } y < y^* \\ g(x) & \text{if } y \geq y^* \end{cases}$$

- **Expected Improvement equation:**

$$EI_{y^*}(x) = \frac{\gamma y^* l(x) - l(x) \int_{y^*}^{R} p(y)\, dy}{\gamma l(x) + (1-\gamma) g(x)} \propto \left(\gamma + \frac{g(x)}{l(x)}(1-\gamma)\right)^{-1}$$

ML ALGORITHMS

We choose a heterogeneous set of algorithms:

1. XGBoost
2. Ridge Regression
3. Neural Networks

XGBOOST

Table 1: Summary of Hyperparameter Optimization Results - XGBoost

Method	Learn. Rate	Max Depth	N Estim.	Accuracy	Time
Grid Search	0.1	9	500	96.45%	7 min 47s
Random Search (20)	0.1	15	500	96.60%	1 min 18s
Bay Opt (50)	0.064	9	437	96.55%	5 min 12s
Bay Opt (20)	0.029	9	443	96.40%	2 min 10s

- **Despite they select different parameters, there is no significant difference in accuracy between the models**
- **Computation times vary**

RIDGE REGRESSION

Table 2: Summary of Hyperparameter Optimization Results - Ridge Regression

Method	Alpha	Accuracy	Time
Grid Search	0.01	85.05	267 ms
Bayesian Optimization (100)	46.265	85.10	5.58 s

- No need to apply Random Search, because we only have one parameter
- The two methods find different alphas

NEURAL NETWORKS

Table 3: Summary of Hyperparameter Optimization Results - Neural Networks

Method	Hidden Dim	Lear Rate	Batch Size	Accuracy	Time
Random Search	224	0.00119	32	96.65	2min 21s
Bayesian Optimization	220	0.0048	128	96.00	2min 28s

- **Grid search is excluded because it is not commonly used in this context**
- **Similar performances and computation time from the two methods**

CONCLUSIONS

- The three optimization methods had similar performances with all the three algorithms

- The computation time also depends on hardware availability and on the efficiency of the libraries that implent the optimization methods

- Bayesian approach may be preferable when:
  - there is a large number of hyperparameters
  - the ML algorithm is particularly quick

THANKS FOR YOUR ATTENTION!