FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Bc. Jakub Arnold

Bayesian Optimization of Hyperparameters Using Gaussian Processes

Institute of Formal and Applied Linguistics

Supervisor of the master thesis: RNDr. Milan Straka, Ph.D.
Study programme: Computer Science
Study branch: Artificial Intelligence

Prague 2019

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date .... signature of the author

Dedication.

Title: Bayesian Optimization of Hyperparameters Using Gaussian Processes

Author: Bc. Jakub Arnold

Institute: Institute of Formal and Applied Linguistics

Supervisor: RNDr. Milan Straka, Ph.D., Institute of Formal and Applied Linguistics

Abstract: Abstract.

Keywords: gaussian process bayesian optimization global optimization neural network

Contents

Introduction

1. Introduction

intro (2 stranky) - co jous hyperparam, ze to model neumi nastavit, atd. - proc chceme delat black box, ...

1.1 Our Contributions

- co jsme udelali

2. Bayesian Optimization overview

Consider the problem of optimizing an arbitrary continuous function $f : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X} \subset \mathbb{R}^d, d \in \mathbb{N}$. We call f the *objective function* and treat it as a black box, making no assumption on its analytical form, or on our ability to compute its derivatives. Our goal is to find the global minimum $\mathbf{x}_{\text{opt}}$ over the set $\mathcal{X}$, that is

$$\mathbf{x}_{\text{opt}} = \arg\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}).$$

We also assume that the evaluation of f is expensive, as the goal of Bayesian optimization is to find the optimum as quickly as possible. Consider the case when evaluating f means performing a computation that is not only time consuming, but for example also costs a lot of money. We might only have a fixed budget which puts a hard limit on the number of evaluations we can perform.

If the function can be evaluated cheaply, other global optimization approaches such as simulated annealing or evolution strategies could potentially yield better results (TODO ref).

Bayesian optimization techniques are some of the most efficient approaches in terms of the number of function evaluations required. Much of the efficiency stems from the ability to incorporate prior belief about the problem and to trade of exploration and exploitation of the search space. [Nando 2012] It is called Bayesian because it combines the prior knowledge $p(f)$ about the function together with the data in the form of the likelihood $p(x|f)$ to formulate a posterior distribution on the set of possible functions $p(f|x)$. We will use the posterior distribution to figure out which point should be evaluated next to give a likely improvement over the currently obtained maximum.

2.1 Acquisition Functions

2.2 Related work

2.3 Hyperparam vs. Architecture Search

2.4 Diskretni hyperparam vs onehot vs ...

3. Gaussian Processes

- uvod, proc to delame - ucbnicove - kernely existujou

Definition 1. *A random variable* X *has a* **univariate Gaussian distribution** *, written as* $\mathrm{X} \sim \mathcal{N}(\mu, \sigma^2)$*, when its density is*

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}.$$

The parameters μ *and* σ *are its* mean *and* standard deviation*.*

Definition 2. *We say* X *has a* **degenerate Gaussian distribution** *when* $\mathrm{X} \sim \mathcal{N}(\mu, 0)$*.*

Definition 3. *A random variable* $\boldsymbol{X} \in \mathbb{R}^n$ *has a* **multivariate Gaussian distribution** *if any linear combination of its components is a univariate Gaussian, i.e.* $\boldsymbol{a}^T\mathrm{X} = \sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{X}_i$ *is a Gaussian for all* $\boldsymbol{a} \in \mathbb{R}^n$*. We then write* $\boldsymbol{X} \sim \mathcal{N}(\mu, \Sigma)$ *where* $\mathbb{E}[\boldsymbol{X}_i] = \mu_i$ *and* $cov(\boldsymbol{X}_i, \boldsymbol{X}_j) = \Sigma_{ij}$*.*

Remark. The parameters μ and Σ uniquely determine the distribution $\mathcal{N}(\mu, \Sigma)$.

Definition 4. *A random variable* $\boldsymbol{X} \sim \mathcal{N}(\mu, \Sigma)$ *has a* **degenerate multivariate Gaussian distribution** *if* $\det \Sigma = \mathbf{0}$*.*

Remark. Given a random variable $\boldsymbol{X} \sim \mathcal{N}(\mu, \Sigma)$, random variables $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are independent with distributions $\boldsymbol{X}_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ if and only if $\mu = (\mu_1, \dots, \mu_n)$ and $\Sigma = diag(\sigma_1^2, \dots, \sigma_n^2)$.

Theorem 1. *If a random variable* $\boldsymbol{X} \in \mathbb{R}^n$ *is a multivariate Gaussian, then* $\mathrm{X}_i, \mathrm{X}_j$ *are independent if and only if* $cov(\mathrm{X}_i, \mathrm{X}_j) = 0$*. Note that his is not true for any random variable, as it is a special property of the multivariate Gaussian.*

Proof. TODO □

4. Bayesian Optimization in depth

- bopt alg - jaky kernely v bayes opt., co pouzivame, proc (ref)

5. Software

- co umime - vizualizace - jak se to pousti, runnery, serializace

6. Experiments

- toy tasky - porovnani existujici fuj fce na optimalizaci - srovnani acq/kernel, random search (ze to neco dela)

- maly ulohy

- velka uloha

- interpretace vysledku

- parser - tokenizer/segmentace - speech recognition - opennmt lemmatizace - $reinforce_{w}ith_{b}aseline$

—

Let $\mathcal{D}_n = \{(\mathbf{x}_i, y_i), i \in 1 : n\}$ denote a set of n samples (evaluations) of the function f, that is $y_i = f(\mathbf{x}_i)$. Our goal is to pick the next $\mathbf{x}_{n+1}$ to maximize our chance of finding the optimum quickly.

Consider the set of all continuous functions $f \in \mathcal{F}$ with a prior distribution $p(f)$. Conditioning on our samples gives us a posterior distribution over possible functions $p(f \mid \mathcal{D})$.

—

Conclusion

Bibliography

List of Figures

List of Tables

List of Abbreviations

A. Attachments

A.1 First Attachment

Index