Summary 14: Practical Bayesian Optimization of Machine Learning Algorithms

Susan Cherry

Next, the paper gives details of Bayesian optimization with Gaussian process priors. Bayesian optimization constructs a probabilistic model for $f(x)$ and uses this model to make make decisions about where in X to evaluate the function next while integrating out the uncertainty. This procedure allows us to find the minimum of non-convex functions with relatively few evaluations. The two major choices when performing Bayesian optimization are which prior to select (Gaussian Process in this paper) and the choice of an acquisition function, which is used to construct a utility function from the model posterior. The paper next gives an brief overview of Gaussian Processes. Then in discusses acquisition functions, which determine what point in X should be evaluated next via a proxy optimization. This paper focuses specifically on Expected Improvement (EI) criterion, which maximizes the expected improvement over the best.

Then the paper discuses several limitations that previously prevented Bayesian optimization from being widely used. First, it is unclear what choice of covariance function and hyperparameters are appropriate. This paper proposes a covariance function that results in sample functions that are twice differentiable do not require the smoothness of the squared exponential that is the traditional choice. Another issue with Bayesian optimization is that problems may vary significantly in duration. This paper proposes optimization with the expected improvement per section, which prefers points that are both likely to be good and likely to be evaluated quickly. Finally, optimization methods should take advantage of multi-core parallelism. The paper proposes a sequential strategy that takes advantage of the inference properties of the GP to compute Monte Carlo estimates of the acquisition function under different possible results.

Next the paper moves to empirical analysis. The paper's methods of expected improvement while marginalizing GP hyperparameters are refereed to as GP EI MCMC, GP EI Opt, GP EI per Second and parallelized GP EI MCMC. First, the Branin-Hoo and Logistic Regression approaches are analyzed. The paper's method significantly outperforms the TPA. Next, the LDA is analyzed. GP EI MCMC is the most efficient in terms of function evaluations but the parallelized GP EI MCMC finds the best parameters much faster. In addition, the parallelized GP EI MCMC finds a much better minimum value than found in a grid search, as used by Hoffman. Next is motif finding with structured support vector machines. In this case, Bayesian optimization strategies are again more efficient than grid search. Finally, convoluted networks are tested on CIFAR-10. Neural networks are notorious for their need of careful tuning, so this is an important test. The best hyperparameters were found by GP EI MCMC, which is far better than other methods. Notably, this algorithm also performed better than a human expert at selective hyperparameters.

In summary, the paper presents methods for performing Bayesian optimization for hyperparameter selection on machine learning algorithms. It introduces Bayesian treatment for EI and algorithms for dealing with variable time regimes and for running in parallel. The effectiveness of these methods are tested empirically and Bayesian optimization is found to find better hyperparameters faster than the traditional approaches.