Résumé: Practical Bayesian Optimization of Machine Learning Algorithms

Auteur : Jasper Snock, Hugo Larochelle, Ryan P. Adams

Les algorithmes de Machine Learning comportent des paramètres et des hyper-paramètres dont le choix demande beaucoup d'habileté, d'expérience et de temps. Pour rendre moins laborieuse cette tâche, il est rapidement né l'envie de rendre ce choix automatique. L'optimisation Bayésienne, qui modélise les performances d'un algorithme comme un phénomène gaussien, peut répondre à ce problème. Dans cet article, il est question de voir le choix des paramètres comme un problème d'optimisation classique et d'utiliser les méthodes connues de la théorie bayésienne pour y répondre. Les auteurs préconisent un traitement complet avec la théorie bayésienne pour optimiser le coût économique de l'algorithme et permettre une parallélisation des calculs. L'article décrit d'abord la théorie bayésienne utilisée dans notre cas avant de s'attarder sur des considérations pratiques pour choisir les hyper-paramètres du modèle. Enfin, ces méthodes bayésiennes sont testées sur différents algorithmes de machine learning : Logistic Regression, LDA et SVM.

1. Théorie bayésienne avec un processus gaussien

Pour trouver le minimum d'une fonction f, les approches classiques s'intéressent au gradient et à la Hessienne au point donné pour trouver le point suivant à évaluer. L'approche bayésienne quant à elle s'intéresse à l'ensemble des informations de f au point x et trouve alors des solutions à des problèmes non-convexes assez complexes. Pour ce faire, il faut choisir la fonction « prior » qui modélisera la fonction f à optimiser, un processus gaussien de paramètre théta dans notre cas, et la fonction d'acquisition nous permettant de construire notre fonction d'utilité et donc de savoir le nouveau point à évaluer.

On modélise chaque observation (x_n,y_n) par $y_n \sim N(f(x_n),v)$. Le but est de considérer la fonction d'acquisition « a » et de trouver son argmax. Ce point sera alors choisi pour calculer la prochaine itération. Il existe plusieurs fonctions « a » possibles : PI : Probability of Improvemet, EI : Expected Improvement, GP Upper Confidence Bound. L'article se concentre ici sur le EI qui n'a pas besoin de paramètre supplémentaire.

2. Considérations pratiques pour choisir les hyper-paramètres

L'approche bayésienne nécessite de calculer la fonction de covariante qui capture beaucoup l'information. Il n'est toutefois pas clair quel kernel il faut utiliser. L'article préconise de modifier lègerement le kernel passant d'ARD à ARD 5/2. Il est aussi nécessaire de choisir les paramètres théta du processus gaussien et ainsi que d'autres paramètres pour débuter notre analyse.

Pour un traitement totallement bayésien, l'astuce est de considérer la fonction a sur l'ensemble des points avec une formule intégrale :

$$\hat{a}(\mathbf{x}\,;\,\{\mathbf{x}_n,y_n\}) = \int a(\mathbf{x}\,;\,\{\mathbf{x}_n,y_n\},\theta)\,p(\theta\,|\,\{\mathbf{x}_n,y_n\}_{n=1}^{N})\,\mathrm{d}\theta,$$

Les figures 1 montrent l'amélioration en terme de convergence.

La méthode de Montecarlo rend possible l'estimation de la fonction d'acquisition et permet la parallélisation des calculs.

En choisissant seulement D+3 variables, l'optimisation bayésienne peut avoir lieu.

3. Etude de la méthode

Les différentes variantes du GP EI (gaussian process with Expected Improvement) sont testées sur 4 études.

1ere étude : Sur MNIST, un data d'images, on tente d'optimiser une régression logistique avec la technique de Branin-Hoo. L'approche bayésienne permet de trouver le minimum en deux fois moins d'itérations que la meilleure technique connue jusqu'ici.

2ieme étude : Pour optimiser une Online LDA, la démarche bayésienne et spécialement le GP EI MCMC est meilleur qu'un classique gridsearch.

3ieme étude : Pour un SVM, le résultat est aussi bien meilleur qu'un gridsearch.

4ieme étude : Sur un CNN, le GP EI MCMC surpasse encore tout le monde de 3 points.

Conclusion

L'article présente une approche complètement bayésienne avec la fonction d'Expected Improvement et permet un parallélisation des calculs. Avec la méthode GP EI MCMC, il est possible de trouver plus rapidement les hyper-paramètres. Dans la compétition CIFAR-10, ces méthodes ont battu de 3% les meilleures modèles connus.

Ces approches rendent le choix des hyper-paramètres beaucoup plus rapide que les méthodes « naîves » du grid search qui est une méthode exhaustive donc couteuse et lente. Surtout sur des modèles de Deep Learning qui ont de nombreux hyper-paramètres, comme ceux cités dans la partie 3, ces méthodes d'optimisation bayésiennes représentent une réelle percée.

Yacine Boukhelif Yahia et Gabriel Melki