Optimisation Bayésienne pratique des algorithmes d'apprentissage machine

Résumé : Dans cet article, de bons entraînements pratiques ont été identifiés pour l'optimisation Bayésienne d'algorithmes d'apprentissage machine. Il en ressort qu'un traitement Bayésien complet du noyau (kernel) sous-jacent du processus Gaussien GP est préféré à l'approche basée sur l'optimisation des hyperparamètres du GP. La deuxième contribution est la description de nouveaux algorithmes pour prendre en considération le coût variable et inconnu d'expériences (durée) ou la disponibilité de cœurs (cores) multiples pour effectuer des expériences en parallèle. Ces nouveaux algorithmes apportent une amélioration sensible et nette par rapport aux procédures automatiques antérieures et peuvent atteindre ou même surpasser l'optimisation de l'expertise humaine pour beaucoup d'algorithmes qui incluent l'allocation latente de Dirichlet, les SVMs structurés et les réseaux neuraux convolutionnels.

1- Introduction

L'optimisation Bayésienne fonctionne typiquement en supposant que la fonction inconnue a été échantillonnée selon un GP et en maintenant une distribution postérieure pour cette fonction comme des observations déjà faites ou, dans notre cas, comme des résultats issus de tests d'algorithmes d'apprentissage avec différents hyperparamètres observés. Pour choisir les hyperparamètres du prochain test, on peut optimiser l'amélioration attendue (EI) sur le meilleur résultat courant ou la borne supérieure de confidence du processus Gaussien (UCB).

2- L'optimisation Bayésienne avec des a priori du processus Gaussien :

Le GP est une distribution a priori commode et puissante sur les fonctions étudiées et sur l'impact des fonctions de covariance. Il y a deux choix majeurs quant à l'exécution de l'optimisation Bayésienne : sélectionner un a priori sur fonctions qui exprime des suppositions antérieures autour de la fonction à être optimisée (choix des a priori du GP) et choisir une fonction d'acquisition, qui est utilisée pour construire une fonction d'utilité à partir du modèle postérieur, nous permettant de déterminer le prochain point à évaluer. Ces fonctions dépendent uniquement du modèle à travers les fonctions prédictives de sa moyenne et de sa variance. Parmi les stratégies intuitives, sous le GP, il y a le choix entre maximiser la probabilité d'amélioration sur la meilleure valeur courante, ou bien maximiser l'amélioration attendue (EI) sur le meilleur courant ou enfin l'idée d'exploiter les bornes inférieures de confidence (supérieures, en cas de maximisation) du GP (GP-UCB) pour construire des fonctions d'acquisition qui minimisent le risque sur le cours de leur optimisation intégrant un paramètre de réglage pour équilibrer l'exploitation contre l'exploration. Le comportement du critère EI est meilleur que celui de la probabilité d'amélioration et son algorithme s'exécute bien dans les problèmes de minimisation, une comparaison directe est donnée entre approches basées EI et GP-UCB.

3- Considérations pratiques pour l'optimisation Bayésienne des hyperparamètres :

Il y a plusieurs limitations pour optimiser les hyperparamètres dans les problèmes d'apprentissage machine. En premier, pour les problèmes pratiques c'est vague de ce qu'un choix approprié est pour la fonction de covariance et ses hyperparamètres associés. En second, comme l'évaluation de la fonction elle-même peut impliquer une procédure d'optimisation gourmande en temps, les problèmes peuvent varier considérablement dans la durée. Troisièmement, les algorithmes d'optimisation devraient profiter du parallélisme du multi-cores. Dans cet article, des solutions ont été proposées à chacune de ces problématiques.

- Les fonctions de covariance et le traitement des hyperparamètres des covariances : En particulier, la détermination de la pertinence automatique (ARD) du noyau à exponentielle carrée est souvent un choix par défaut pour une régression GP ; cependant, pour une utilisation pratique réaliste, on préfère utiliser le kernel de l'ARD Matérn 5/2. L'article montre comment l'amélioration attendue intégrée change la fonction d'acquisition en l'améliorant grâce au traitement complet Bayésien pour expliquer l'incertitude dans les hyperparamètres (avec comparaison de MCMC).

- Modélisation des coûts : L'objectif d'optimisation Bayésienne est de trouver aussi rapidement que possible un bon réglage des hyperparamètres. Grâce au suivi de l'amélioration attendue EI, on essaie d'améliorer la prochaine évaluation de la fonction. D'un point de vue pratique, on s'intéresse cependant à la rapidité à travers l'optimisation de l'amélioration attendue par seconde.

- L'acquisition de Monte Carlo pour la parallélisation de l'optimisation Bayésienne : Une stratégie séquentielle est proposée tirant profit des propriétés déductibles du GP pour calculer les estimés selon la procédure Monte Carlo de la fonction d'acquisition fonctionnant sous différents résultats possibles des évaluations de la fonction considérée. On trouve que la procédure adoptée de l'estimation de Monte Carlo est efficace dans l'entraînement.

4- Analyses empiriques

Application de la méthode d'amélioration attendue EI en marginalisant les hyperparamètres GP, l'optimisation des hyperparamètres GP, EI par seconde et N fois GP EI MCMC parallélisé. Les résultats montrent la progression de la meilleure valeur courante $\mathbf{x}_{\text{best}}$ sur le nombre d'évaluations de la fonction ou sur le temps (durée) moyenné sur de multiples

déroulements de chaque algorithme. Si non spécifié autrement, par une optimisation guidée $\mathbf{x}_{\text{next}}$ qui détermine quel point prochain dans le sous ensemble considéré, devrait être évalué et la recherche est à base du gradient avec des redémarrages multiples (Le code utilisé est publiquement disponible à *http://www.cs.toronto.edu/.jasper/software.html*).

-Branin-Hoo et régression logistique : La fonction Branin-Hoo est une référence pour les techniques d'optimisation Bayésienne et l'Algorithme de Parzen à Arbres (TPA) pour la classification en régression logistique sur les dataset connues de MNIST (à 4 hyperparamètres) : surclassement de GP EI par rapport à TPA en nombre d'évaluations de la fonction et GP EI par seconde sur GP EI MCMC en minutes (durée).

- LDA en ligne: L'Allocation latente de Dirichlet (LDA) Plusieurs stratégies d'optimisation sur la même grille ont été testées sur le problème LDA. Clairement vouloir intégrer sur les hyperparamètres est supérieur à celui d'utiliser un point d'évaluation dans ce cas. Le GP EI MCMC est le plus effectif quant à l'évaluation de la fonction, on relève aussi que le GP EI MCMC parallélisé abouti considérablement à de meilleurs paramètres en un moins de temps.

-A la recherche de motifs dans les SVM structurées : Une comparaison de plusieurs stratégies pour optimiser les hyperparamètres de modèles de l'entropie minimale à marge maximale (Max-marge Min-entropie (M3E) sur le motif de la protéine ADN, lesquels incluent des SVM Structurées Latentes où présence de variables cachées dépendantes explicitement du problème considéré. Les hyperparamètres, tel que le terme de régularisation, C, de SVMs structuré reste un défi car ils prennent du temps de la procédure de recherche de la grille et donc sont coûteux. Il a été observé que les stratégies d'optimisation Bayésienne sont considérablement plus effectives que la recherche de la grille. Dans ce cas, GP EI MCMC est supérieur à GP EI par seconde quant aux évaluations de la fonction mais GP EI par seconde trouve de meilleurs paramètres plus rapides que GP EI MCMC. 3x GP EI par seconde, est le moins effectif quant aux évaluations de la fonction mais permet de découvrir de meilleurs paramètres plus rapidement que tous les autres algorithmes.

Réseaux Convolutionnels sur CIFAR-10 : Les **r**éseaux neuraux et les méthodes de deeplearning (apprentissage en profondeur) exigent notoirement le réglage prudent de nombreux hyperparamètres (tels ceux de régularisation), et qui sont prohibitifs de point de vue temps de calcul. Dans cette analyse empirique, neuf hyperparamètres sont réglés sur les dataset du CIFAR-10 qui utilisent le code disponible à: *http://code.google.com/p/cuda-convnet/*. Les meilleurs hyperparamètres trouvés par le GP EI MCMC approchent une erreur sur l'ensemble de l'épreuve de 14.98% et sont mieux que celle de l'expert et de l'état de l'art sur 3% sur le compétitif CIFAR -10.

5 Conclusion

Cet article présente des méthodes pour exécuter l'optimisation Bayésienne pour la sélection des hyperparamètres des algorithmes d'apprentissage machine et l'introduction du traitement Bayésien complet pour EI, et des algorithmes pour l'étude des régimes à temps variable et des expériences courantes en traitement parallèle. L'optimisation Bayésienne résultante trouve de meilleurs hyperparamètres plus significatifs quant à la rapidité que les approches utilisées par d'autres auteurs et dépasse même l'expertise humaine ou l'état de l'art à pouvoir sélectionner les hyperparamètres sur de compétitifs datasets. EI et UCB ont montré leur efficacité dans le nombre d'évaluations requises de fonction pour trouver l'optimum global de beaucoup de fonctions multimodales à boîte noire. Cependant, pour les algorithmes d'apprentissage machine, chaque évaluation de fonction peut exiger une quantité variable de temps (durée), et les tests sur l'apprentissage machine sont souvent effectués en parallèle, sur des cores multiples des machines. Dans les deux situations, l'approche séquentielle standard d'optimisation GP peut-être sous-optimale. Récemment, plusieurs stratégies ont été explorées pour optimiser les hyperparamètres d'algorithmes d'apprentissage machine en notant que les stratégies de recherche en grille (grid computing) sont inférieures à la recherche aléatoire, et que l'usage de l'optimisation Bayésienne pour les GP est conseillé en optimisant les hyperparamètres d'une covariance à exponentielle carrée.

Quant au lien avec la problématique de la sélection de modèles, il ressort que dans le cadre de la modélisation statistique des données, l'approche générale de l'optimisation Bayésienne apporte un éclairage nouveau quant à la modélisation algorithmique via l'apprentissage machine, souplesse, efficacité et même rapidité. Il s'avère que dans les méthodes de modélisation algorithmique (machine learning), on a besoin d'utilitaires dérivés des statistiques pour l'évaluation , le réglage et l'optimisation des hyperparamètres des modèles (exemple de termes de régularisation pour les RNs et SVMs, les meilleurs trouvés sont dus à GP EI MCMC (modélisation stochastique des données paramètres)). Il en ressort qu'il faut en vue de trouver de bonnes solutions, de s'intéresser d'une part davantage aux données avant de s'y mettre dans la modélisation (d'où l'intérêt d'algorithmes d'apprentissage machine), et d'autre part, valider le modèle en s'appuyant sur des indices corrects issus de la modélisation des données. En bref, il faut tenir compte des apports complémentaires des deux types de modélisation statistique. En somme l'optimisation d'algorithmes d'apprentissage machine passe par Mr. Thomas Bayes, un statisticien.