Optimisation Bayésienne et Modèles Bayésiens à Noyau

Ml OIVM / TP2 ATDN

Vincent GUIBLAIN

Objectifs pédagogiques

Les objectifs pédagogiques de ce travail sont les suivants :

— Comprendre les principes de l'optimisation bayésienne
— Maîtriser les modèles bayésiens à noyau
— Mettre en œuvre ces techniques sur des problèmes concrets
— Analyser les avantages par rapport aux méthodes classiques

Génération des données

Les données seront simulées afin de garantir un environnement d'apprentissage optimal. Un fichier CSV sera fourni contenant des valeurs modifiées pour éviter toute fuite de données sensibles. Les données incluent des caractéristiques pour la prédiction de rendement agricole basé sur divers facteurs environnementaux.

Partie 1 : Optimisation Bayésienne (10 points)

Fondements théoriques

1. (1 pt) **Expliquez le principe de l'optimisation bayésienne.**
   L'optimisation bayésienne est utilisée lorsqu'on estime que la fonction est difficile à estimer. On utilise donc un modèle probabiliste avec un processus gaussien pour gérer les fonctions coûteuses à évaluer.
2. (1 pt) **Définissez et expliquez les processus gaussiens. Pourquoi sont-ils utilisés pour modéliser la fonction objective ?**
   Un processus gaussien est une distribution de probabilité pour chaque fonction. Ils permettent de modéliser des relations non linéaires de manière souple.
3. (1 pt) **Décrivez les principales fonctions d'acquisition (Expected Improvement, Upper Confidence Bound, etc.). Expliquez leur rôle dans le compromis exploration/exploitation.**
   La fonction d'acquisition identifie où évaluer la fonction coûteuse en se basant sur le modèle probabiliste.
   L'Expected Improvement mesure l'espérance du gain potentiel par rapport au meilleur résultat connu.

La fonction UCB combine directement la moyenne et l'incertitude. Ces fonctions permettent de trouver un bon compromis entre exploration (nouvelles zones) et exploitation (zones prometteuses déjà explorées).

Implémentation et applications

4. (2 pts) Implémentez une optimisation bayésienne pour maximiser la production agricole en fonction de l'humidité et de la température. Visualisez les étapes du processus.

5. (2 pts) Utilisez l'optimisation bayésienne pour ajuster les hyperparamètres d'un modèle de régression (ex : Random Forest) sur les données agricoles fournies. Comparez les résultats avec Grid Search et Random Search.
On remarque ici que les résultats obtenus sont très similaires avec les deux méthodes.

6. (2 pts) Visualisez le processus d'optimisation (courbe de convergence, choix des points). Commentez la manière dont le modèle explore l'espace de recherche.
À partir d'environ 20 itérations, on observe une réduction importante du RMSE. Au-delà de 20 itérations, nous n'observons pas de changement significatif, indiquant que le modèle a probablement atteint un optimum local ou global.

7. (1 pt) Analysez les avantages et limites de l'optimisation bayésienne face aux méthodes classiques.
Le principal avantage de l'optimisation bayésienne est qu'elle est efficace pour les fonctions coûteuses à évaluer. Elle utilise un modèle probabiliste, ce qui constitue un atout majeur, et elle est bien adaptée aux petits jeux de données.
En revanche, elle peut être très coûteuse en termes de calcul et ne convient pas aux fonctions bruitées et/ou discontinues.

Partie 2 : Modèles Bayésiens à Noyau (10 points)

Fondements théoriques

8. (1 pt) **Expliquez le concept d'inférence bayésienne. Comment met-on à jour les croyances avec de nouvelles données ?**
L'inférence bayésienne s'appuie sur le théorème de Bayes et permet de mettre à jour nos prédictions. On obtient une distribution à posteriori à partir d'une distribution a priori et de la vraisemblance.

9. (1 pt) **Décrivez la théorie des méthodes à noyau et leur lien avec les processus gaussiens. Pourquoi utiliser un noyau dans un modèle bayésien ?**
Les méthodes à noyau permettent de modéliser des relations complexes. Le processus gaussien utilise une fonction de noyau pour intégrer l'incertitude et les relations non linéaires.

10. (1 pt) **Qu'est-ce qu'une distribution a priori et une distribution a posteriori ? Donnez un exemple appliqué à la prédiction de rendement agricole.**
La distribution a priori reflète notre connaissance initiale. Exemple : on suppose que le rendement est distribué selon une loi normale $N(7, 1^2)$. Après observation des données, la distribution devient $N(7.5, 0.42^2)$.

Implémentation et applications

11. (2 pts) Implémentez une régression bayésienne à noyau sur les données agricoles fournies.
Visualisez les prédictions et les intervalles de confiance.

12. (2 pts) Réalisez une classification bayésienne à noyau pour prédire le type de sol (argileux, sableux, limoneux) en fonction des données climatiques.
Comparez les résultats avec un SVM classique.

13. (1 pt) Analysez l'incertitude dans les prédictions.
Commentez les zones où le modèle est moins confiant.

14. (1 pt) Testez différents noyaux (linéaire, RBF, polynomial).
Quelle est la différence entre eux et quel impact ont-ils sur la précision du modèle ?

15. (1 pt) Discutez de l'influence des choix de noyau et de la distribution a priori sur les résultats.

Instructions

— Vous devez commenter chaque bloc de code pour expliquer vos choix.

— Les résultats doivent être présentés sous forme de graphiques et d'analyses.

— Veillez à ce que votre code soit propre et bien organisé.

— Un rapport synthétisant vos observations sera demandé en complément du code.