Vincent
GUIBLAIN

MI OIVM / TP2 ATDN

Optimisation Bayésienne et Modèles Bayésiens à Noyau
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.

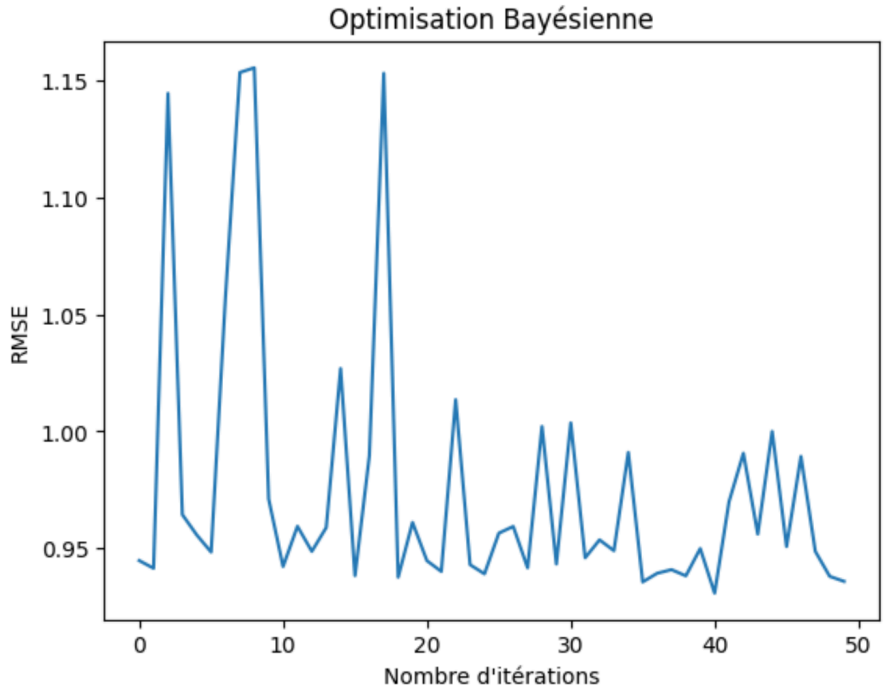
Partie 1 : Optimisation Bayésienne (10 points)

Fondements théoriques
1. (1 pt) Expliquez le principe de l'optimisation bayésienne.
Décrivez comment elle permet de gérer les fonctions coûteuses à évaluer.
On utilise l'optimisation bayésienne lorsqu'on estime que la fonction est difficile à estimer. On utilise donc un modèle probabiliste avec un processus gaussien.

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 sont utilisés car ils permettent de modéliser des relations non linéaires.

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 permet d'identifier où évaluer la fonction coûteuse en s'appuyant sur le modèle probabiliste. L'Expected Improvement mesure l'espérance du gain potentiel par rapport au meilleur résultat connu jusqu'à présent. La fonction UCB combine directement la moyenne et l'incertitude.
Ces deux fonctions permettent de trouver le prochain point à tester en trouvant le bon compromis entre exploration et exploitation (découvrir de nouvelles zones intéressantes ou approfondir dans celles qu'on a 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.

```
RMSE optimisation bayésienne: 0.9307094420295965
RMSE Grid Search: 0.9739080629455165
RMSE Random Search: 0.975434970212989
```

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.

A partir d' environ 20 itérations on observe une réduction important du RMSE. Au delà de 20 itérations nous n'observons pas de changement important.

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'il est efficace pour les fonctions coûteuses à évaluer. Il utilise également un modèle probabiliste ce qui en fait un bon atout et il est adapté aux petits jeux de données.

En revanche, l'optimisation bayésienne est très couteuse 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.

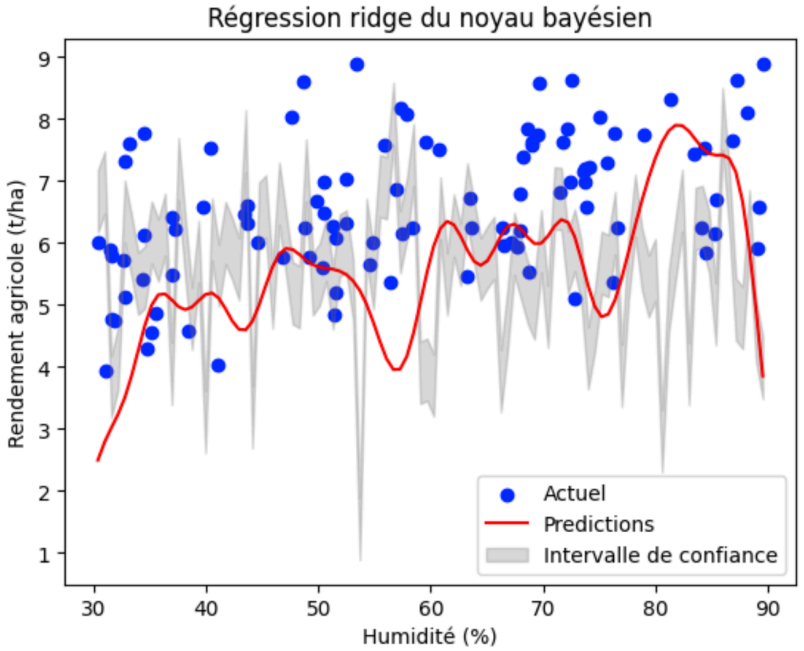
Comment met-on à jour les croyances avec de nouvelles données ?

L'inférence bayésienne est une méthode probabiliste qui s'appuie sur le théorème de Bayes et permet de mettre à nos jours nos prédictions. On obtient une distribution à posteriori à partir d'une distribution à priori et de la vraisemblance. On réactualise la distribution du paramètre inconnu dès qu'on observe de nouvelles données.

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 de grande dimension. Le processus gaussien repose entièrement sur une fonction de noyau. Le noyau permet de capturer des relations complexes et intègre naturellement l'incertitude dans les prédictions.

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.
Une distribution a priori représente ce que l'on croit savoir avant d'observer les données. Elle encode notre connaissance initiale (ou hypothèses). Une distribution a posteriori est le résultat de la mise à jour bayésienne : c'est notre nouvelle croyance après avoir observé les données.
Par exemple, pour la prédiction agricole on peut estimer à priori que le rendement est en moyenne de 7 t/ha avec une incertitude de +/- 1t. On modélise alors par une loi normale $N(7, 1^2)$. Puis, après observation des données on obtient un modèle plus centré et moins dispersé on le met à jour et on a une distribution qui suit $N(7.5 , 0{,}4^2)$.

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.*



On remarque une tendance globale: plus il y a d'humidité et meilleur est le rendement agricole. Les prédictions restent dans l'ensemble inférieures aux valeurs actuelles mais suivent la tendance globale. Les intervalles de confiance sont de forme sinusoïdale et restent au centre du tableau sans suivre la tendance globale.
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.*

```
Précision de la classification du noyau bayésien: 0.31
Précision du SVM classique: 0.31
```

Les précisions obtenues restent très faibles mais égales entre les modèles.

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.