ATDN2 TP2

Remy XU M1 OIVM

31 Mars 2025

Fondements théoriques

Question 1

L'optimisation bayésienne est une technique pour trouver le meilleur point d'une fonction qui est chère à évaluer. On cherche à minimiser le nombre de fois où on doit évaluer cette fonction tout en obtenant le plus d'infos possible à chaque essai.

Comment ça marche :

- Modèle probabiliste : On utilise un modèle (souvent un processus gaussien) pour représenter la fonction qu'on veut optimiser. Ce modèle est mis à jour à chaque étape avec les nouvelles données.
- Fonction d'acquisition : C'est une sorte de guide qui nous dit où évaluer la fonction ensuite. Elle aide à équilibrer entre explorer de nouvelles zones et exploiter les zones déjà connues.
- Mise à jour : Après chaque évaluation, on met à jour le modèle pour qu'il soit plus précis.
- Répétition : On répète le processus jusqu'à ce qu'on soit satisfait du résultat ou qu'on atteigne un nombre max d'itérations.

Question 2

Un processus gaussien (GP) est un modèle qui nous aide à représenter des fonctions continues de manière probabiliste. Il est défini par une moyenne et une fonction de covariance.

Pourquoi on l'utilise :

- Flexibilité : Les GP peuvent modéliser plein de types de fonctions grâce à leur fonction de covariance.
- Incertitude : Ils nous donnent non seulement une estimation de la fonction, mais aussi une idée de l'incertitude autour de cette estimation.
- Mise à jour bayésienne : On peut facilement mettre à jour nos croyances sur la fonction à mesure qu'on obtient de nouvelles données.

Question 3

Expected Improvement (EI) :

- Idée : Mesure combien on s'attend à améliorer par rapport au meilleur point connu.
- Rôle : Aide à trouver des points où l'amélioration est probable.

Upper Confidence Bound (UCB) :

- Idée : Choisit le point qui maximise la somme de la moyenne prédite et d'un terme d'incertitude.
- Rôle : Encourage à explorer les zones incertaines et à exploiter les zones prometteuses.

Probability of Improvement (PI) :

- Idée : Mesure la probabilité qu'une évaluation améliore le meilleur point connu.
- Rôle : Favorise les points où l'amélioration est probable, mais peut être trop prudent.

Thompson Sampling :

- Idée : Échantillonne la fonction à partir du modèle et choisit le point qui maximise l'échantillon.
- Rôle : Équilibre naturellement l'exploration et l'exploitation.

Avant de commencer la suite :

Analysons les données:

La table nous renseigne sur le rendement agricole en fonction de plusieurs paramètres.
Parmi ces paramètres, il y a 5 paramètres avec des paramètres quantitatives, et une seule qualitative qu'on va prendre soin de la transformer en entier.

On peut se demander: Faut-il standardier les données avant d'entrainer les 3 modèles (Bayésien, Random, et Grid)?

Non, car les arbres de décision et les forêts aléatoires sont insensibles à l'échelle des données. Ils effectuent des découpages successifs des données sur chaque caractéristique indépendamment de son échelle. Contrairement aux modèles basés sur des distances, ils ne sont pas affectés par des valeurs extrêmes ou des unités différentes.

Implémentation et applications

Question 4

Voici ci-dessous une exécution du modèle bayésien lors des premiers essais :

Listing 1: Exemple de code Python

iter	target	max_depth	min_sa...	n_esti...
1	-119.8	38.33	7.687	169.8
2	-131.6	17.7	1.117	183.3
3	-119.0	44.36	9.808	161.5
4	-120.3	5.941	4.543	77.43
5	-122.5	11.78	5.174	58.03
6	-131.9	19.0	1.229	117.8
7	-118.7	8.643	9.281	175.5
8	-120.2	6.683	7.517	69.61
9	-120.3	6.407	7.151	85.18
10	-124.9	31.07	3.021	184.5
11	-118.6	6.942	9.478	175.3
12	-118.6	10.5	10.0	161.5
13	-118.8	28.35	10.0	154.4
14	-132.2	50.0	1.0	145.3
15	-119.3	27.14	10.0	77.73
16	-119.1	46.21	10.0	70.7
17	-132.3	44.64	1.0	86.16
18	-119.0	35.36	10.0	60.15
19	-119.6	50.0	10.0	54.05
20	-132.8	38.26	1.0	37.51
21	-118.4	50.0	10.0	183.4
22	-131.3	50.0	1.0	200.0
23	-136.4	2.0	10.0	146.1
24	-118.8	21.95	9.594	165.6
25	-138.0	2.0	10.0	10.0
26	-123.2	50.0	10.0	10.0
27	-131.5	48.95	1.046	174.2
28	-133.7	26.55	1.702	67.76
29	-119.0	33.71	9.25	161.8
30	-119.2	19.71	10.0	86.63
31	-132.0	21.61	1.426	157.9
32	-120.7	44.75	6.171	61.64
33	-119.3	15.28	10.0	77.68
34	-119.1	4.144	2.534	167.2
35	-136.6	2.0	10.0	46.73
36	-118.8	32.48	10.0	143.5
37	-119.2	37.96	9.532	151.9
38	-136.4	2.403	8.857	100.7
39	-136.6	2.0	10.0	200.0

```
| 40        | -118.9    | 50.0      | 10.0      | 116.7     |
| 41        | -119.2    | 39.72     | 9.435     | 124.6     |
| 42        | -119.1    | 38.96     | 10.0      | 109.5     |
| 43        | -131.9    | 45.04     | 1.013     | 114.5     |
| 44        | -119.8    | 28.85     | 9.108     | 97.79     |
| 45        | -118.9    | 29.64     | 10.0      | 131.7     |
| 46        | -141.1    | 33.2      | 1.0       | 10.0      |
| 47        | -118.8    | 39.36     | 9.772     | 186.7     |
| 48        | -119.6    | 49.77     | 9.796     | 103.1     |
| 49        | -136.6    | 2.0       | 10.0      | 166.5     |
| 50        | -125.0    | 12.24     | 3.868     | 167.1     |
| 51        | -119.0    | 37.32     | 10.0      | 134.3     |
| 52        | -119.2    | 28.59     | 10.0      | 87.09     |
| 53        | -120.6    | 3.795     | 1.355     | 175.3     |
| 54        | -118.5    | 29.18     | 10.0      | 170.9     |
| 55        | -128.9    | 31.72     | 2.13      | 136.7     |
| 56        | -119.3    | 39.09     | 10.0      | 100.4     |
| 57        | -132.2    | 18.6      | 1.131     | 83.02     |
| 58        | -118.9    | 49.05     | 10.0      | 126.4     |
| 59        | -120.0    | 41.24     | 9.551     | 53.66     |
| 60        | -119.7    | 37.58     | 9.865     | 73.47     |
=============================================================
Meilleurs hyperparam tres : {'max_depth': 50.0, 'min_samples_leaf':
    ↪ 10.0, 'n_estimators': 183.3605857474121}
Meilleur score : -118.41309633086561
```

Nous avons comme paramètres pour le modèle Bayésien: à quelle itération on est, le score qu'on a calculé avec le MSE négatif 'neg_mean_squared_error' (somme des erreurs au carré), la profondeur maximale, le minimum d'échantillons de feuilles, et le nombre d'estimateurs.

Le but de notre modèle sera alors de maximiser le score (qui est negatif). Ainsi en repassant au négatif on aura le MSE le plus faible (le nombre d'erreurs au carré).

Le modèle va générer plusieurs couples d'hyperparamètres avec un score associé. Certains couples seront affichés en violet lorsque le score est plus haut que la normale. Le modèle decidera alors d'avoir d'avantages d'hyperparamètres simlaire à ceux qui ont donné de bon résultats durant cette itération.

Question 5

Comparons les résultats des trois méthodes d'optimisation des hyperparamètres (Bayesian Optimization, RandomizedSearchCV, et GridSearchCV), afin de trouver les hyperparametres avec le meilleur score et hyperparametres :

Listing 2: Utilisation de RandomizedSearch et GridSearch

```
Meilleurs hyperparametres (RandomizedSearchCV) : {'n_estimators': 164,
    ↪ 'min_samples_leaf': 8, 'max_depth': 4}
Meilleur score (RandomizedSearchCV) : -118.4201937817151

Meilleurs hyperparametres (GridSearchCV) : {'max_depth': 10, '
    ↪ min_samples_leaf': 10, 'n_estimators': 200}
Meilleur score (GridSearchCV) : -118.51634655826328
```

Listing 3: Temps d execution des 3 tests d hyperparametres:

```
Bayesian Optimization : 67.423 secondes
RandomizedSearchCV : 19.122 secondes
GridSearchCV : 45.349
```

Question 6

Voici les courbes de convergence pour l'ensemble des 3 modèoles utilisés :

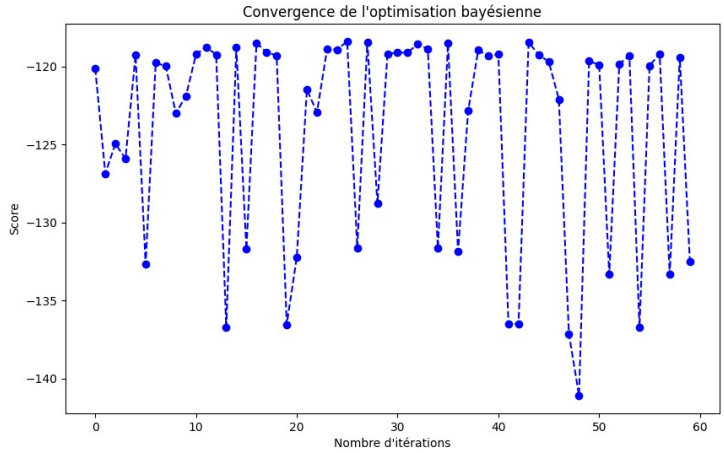


Figure 1: Graphique de convergence pour le modèle Bayésien avec peu d'entraînements

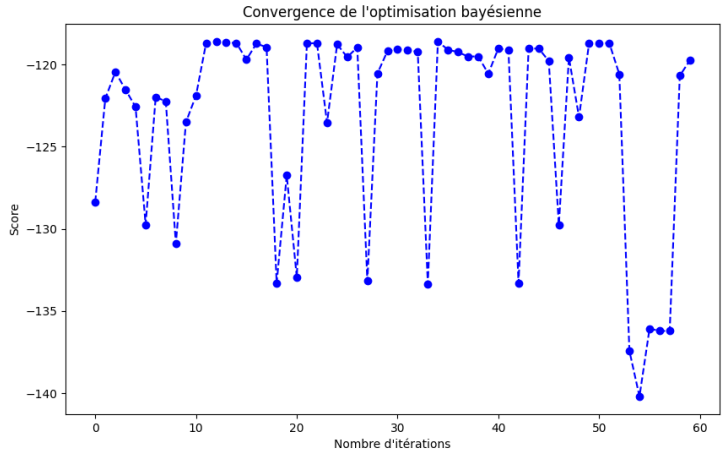


Figure 2: Graphique de convergence pour le modèle Bayésien avec beaucoup d'entraînements

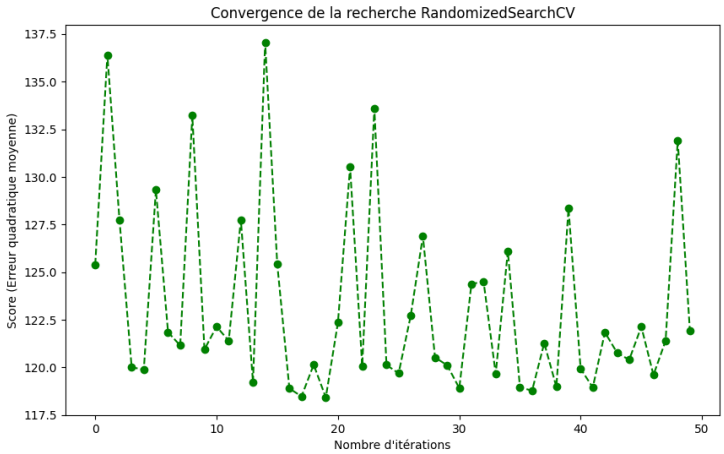


Figure 3: Graphique de convergence pour le modèle Random

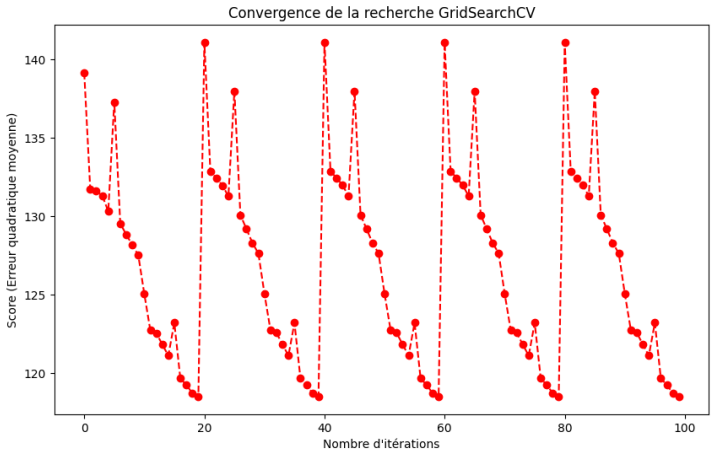


Figure 4: Graphique de convergence pour le modèle Grid

Analyse optimisation bayésienne

Pour la recherche bayésienne, on remarque une majorité de points proches du maximum (ou MSE minimum), mais également beaucoup de points qui témoignent de scores plus faibles.

Hypothèse : Le modèle va beaucoup chercher autour des bons paramètres pour améliorer peu à peu le score, mais essaiera de temps en temps des valeurs assez différentes.

Après entraînement, on remarque qu'il y a davantages de tests qui ont des bons scores et moins de tests ayant de mauvais scores.

On ne remarque pas de motifs qui se répètent. cela a l'air assez aléatoire.

Analyse optimisation random

Pour la recherche random, on remarque que les valeurs sont, comme le nom l'indique, aléatoires. On remarque d'avantages de valeurs intermédiaires. Mais une présence aléatoire de bon ou mauvais scores. L'optimisation est alors pas la meilleure solution mais est extrèmement rapide. Sur un coup de chance, on peut avoir de très bons résultats rapidement!

Analyse optimisation grid

Pour la recherche Grid, on reconnait un motif de répétition lorsqu'on teste l'ensemble des valeurs possible avec un certain pas. On remarque sur le graphique de convergence qu'il y a encore plus de valeurs intermédiaires. Mais dans le cas de recherche d'hyperparamètres optimiaux, ce sont des valeurs assez inutiles.

Question 7

Les avantages de l'optimisation Bayésienne sont alors :

- La pertinence : Les valeurs qu'on teste sont bien plus pertinentes. On teste directement les valeurs proches des anciennes bonnes valeurs, tout en testant de temps en temps des valeurs différentes.
- L'amélioration constante : Les performances ne peuvent quasiment que s'améliorer.

Cependant, on trouve également certains désavantages:

- Le temps d'exécution : On a un temps d'execution qui est significativement plus élevé, car on doit entrainer le modèle. C'est à dire qu'on a $nombred'entrainements \times tempsd'executiond'une fois$ secondes d'execution, ce qui revient à beaucoup de temps comparé aux autres modèles.
- Moins modulable : Si on veut par exemple chercher des hyperparamètres moins performants, ce sera impossible avec le modèle bayésien, mais possible avec les autres.

Fondements théoriques

Question 8

L'inférence bayésienne est une méthode pour mettre à jour nos croyances sur un modèle ou une hypothèse à la lumière de nouvelles données. Elle repose sur le théorème de Bayes, qui relie les probabilités conditionnelles.

Comment ça marche :

1. **Croyances initiales (prior)** : On commence avec une croyance initiale sur un paramètre ou une hypothèse, appelée distribution a priori.
2. **Nouvelles données** : On obtient de nouvelles données.
3. **Mise à jour des croyances (posterior)** : On utilise le théorème de Bayes pour mettre à jour nos croyances initiales en tenant compte des nouvelles données. La formule de Bayes est :

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

où $P(A|B)$ est la probabilité a posteriori, $P(A)$ est la probabilité a priori, $P(B|A)$ est la vraisemblance, et $P(B)$ est une constante de normalisation.

Question 9

Théorie des méthodes à noyau et leur lien avec les processus gaussiens

Les méthodes à noyau sont des techniques utilisées en apprentissage automatique pour transformer des données dans un espace de plus haute dimension, où elles peuvent être plus facilement séparées ou modélisées. **Lien avec les processus gaussiens :**

- **Noyau** : Dans un processus gaussien, le noyau (ou fonction de covariance) définit la similarité entre les points. Il permet de modéliser des relations complexes entre les données.
- **Modèle bayésien** : Les processus gaussiens sont des modèles bayésiens non paramétriques. Le noyau joue un rôle crucial en déterminant la structure de la fonction modélisée.

Pourquoi utiliser un noyau :

- **Flexibilité** : Les noyaux permettent de capturer des relations non linéaires entre les données.
- **Modélisation de l'incertitude** : Ils aident à modéliser l'incertitude dans les prédictions, ce qui est essentiel dans un cadre bayésien.

Question 10

Distribution a priori :

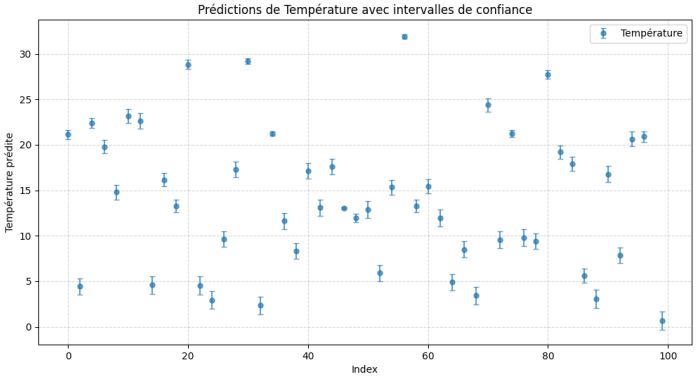
- **Définition** : C'est notre croyance initiale sur un paramètre avant d'observer les données. Elle reflète nos connaissances ou hypothèses antérieures.
- **Exemple** : Dans la prédiction de rendement agricole, la distribution a priori pourrait être basée sur des connaissances historiques ou des hypothèses sur les conditions climatiques.

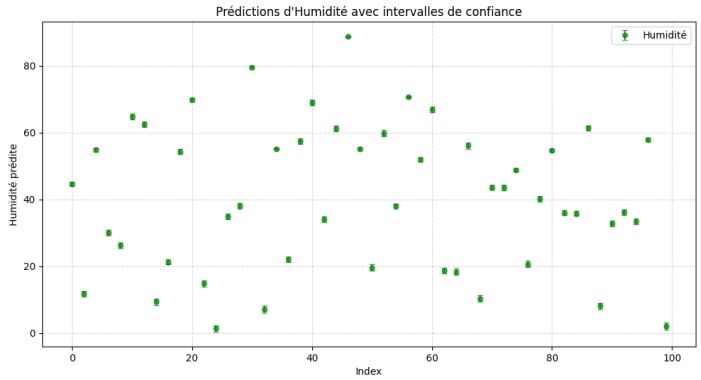
Distribution a posteriori :

- **Définition** : C'est notre croyance mise à jour sur un paramètre après avoir observé les données. Elle combine la distribution a priori avec les nouvelles données via le théorème de Bayes.
- **Exemple** : Après avoir collecté des données sur les conditions météorologiques actuelles et les pratiques agricoles, la distribution a posteriori reflétera une estimation plus précise du rendement agricole.

Implémentation et applications

Question 11





Question 12

La classification bayésienne à noyau (Gaussian Process Classifier) est un modèle probabiliste qui peut être utilisé pour classer des données dans des catégories comme le type de sol (argileux, sableux, limoneux). En la comparant à un **SVM classique** (Support Vector Machine), on peut observer que les deux modèles peuvent donner des résultats similaires pour certaines configurations de données, mais la classification bayésienne offre l'avantage supplémentaire de fournir une estimation de l'incertitude associée à chaque prédiction.

Question 13

L'incertitude dans les prédictions se manifeste par les intervalles de confiance plus larges, où le modèle est moins confiant. Les zones où l'incertitude est élevée sont souvent celles où les données sont peu denses ou là où il y a une grande variabilité. Par exemple, si le modèle est confronté à des points de données isolés ou peu représentatifs dans l'espace des caractéristiques, il aura plus de mal à faire des prédictions fiables et l'incertitude sera plus grande.

Question 14

Les noyaux sont utilisés pour transformer les données dans un espace de caractéristiques où elles sont plus facilement séparables ou ajustables. Voici la différence entre les noyaux que nous avons testés :

- **Noyau linéaire** : Utilisé lorsque la relation entre les variables d'entrée et la sortie est supposée linéaire.
- **Noyau RBF (Radial Basis Function)** : Ce noyau est plus flexible et adapté pour des relations non linéaires complexes.
- **Noyau polynomial** : Permet de capturer des relations non linéaires, mais peut être plus sensible au bruit.

L'impact de ces noyaux sur la précision du modèle dépend fortement des données. Le noyau RBF est souvent plus performant pour des données non linéaires.

Question 15

Le choix du noyau détermine comment les données sont projetées dans un espace de caractéristiques plus élevé, influençant ainsi la capacité du modèle à capturer les relations complexes. La **distribution a priori** joue également un rôle important : un a priori informatif peut guider le modèle, tandis qu'un a priori non informatif permet au modèle de s'adapter davantage aux données observées.

Listing 4: Resultats en probabilités (entre 0 et 1)

```
Precision de la classification bay sienne: 0.31
Precision du SVM: 0.34
Precision avec noyau Lin aire: 0.370
Precision avec noyau RBF: 0.310
Precision avec noyau Matern: 0.310
```