Validation “historique” de modèles, point de vue math-info

Peter, Vincent

STEEP, SET

Introduction

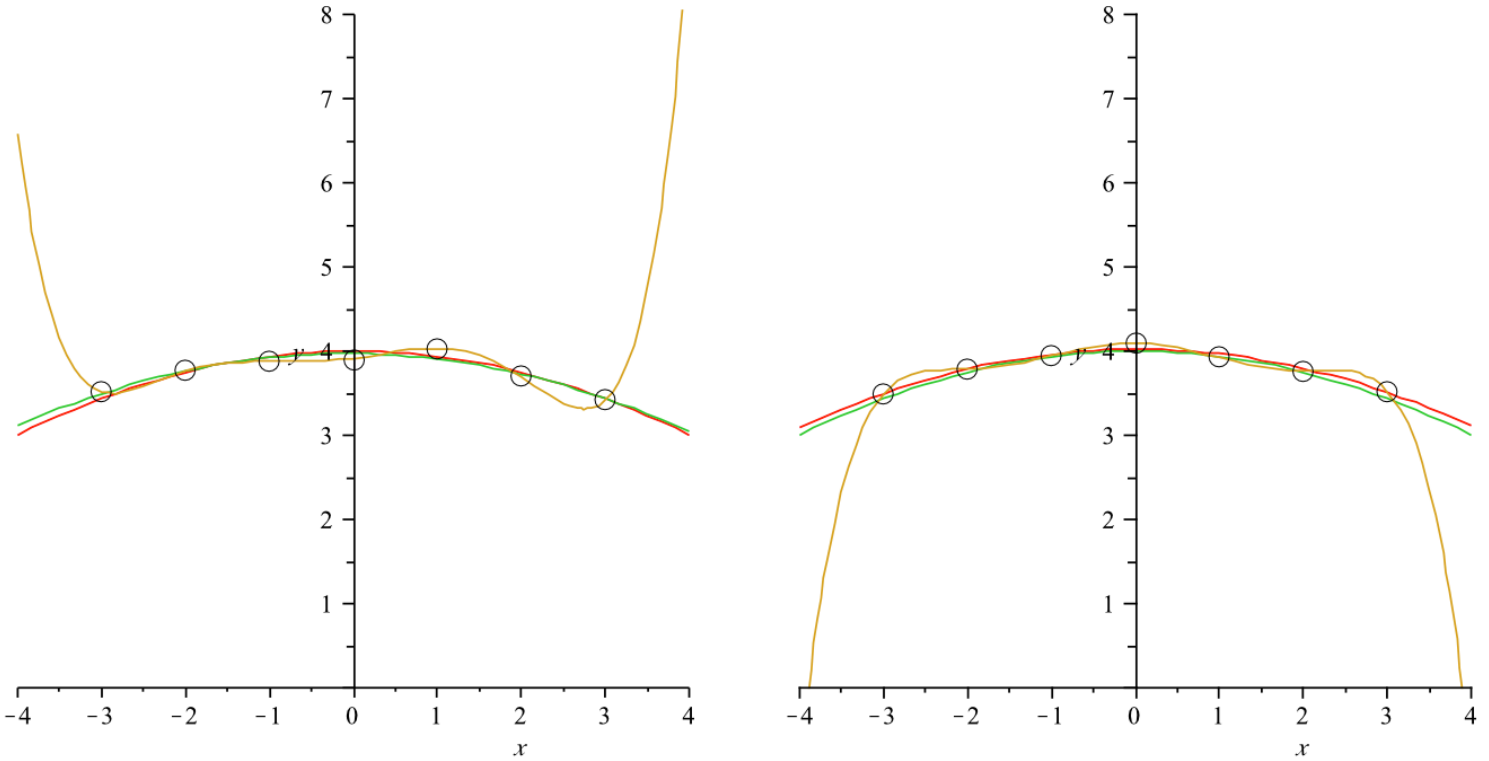
- Un modèle peut être utilisé de différentes manières – on peut poser différentes questions à un modèle
- Il est souhaitable de pouvoir valider si les réponses données par le modèle (modèle et ses paramètres) sont pertinentes
- Par validation "historique", nous entendons *grosso modo*, toute validation basée sur des observations
- $\longrightarrow$ confrontation de sorties du modèle avec ces observations

Introduction

- Une première validation s'effectue généralement lors du calage d'un modèle (e.g. en regardant la valeur de la fonction objectif).
- Cela n'est souvent pas suffisant, surtout si :
  - données de calage insuffisantes *vis-à-vis* de la complexité du modèle
  - données trop incertaines / bruitées
  - lors de l'utilisation du modèle, on "pose d'autres questions" que lors du calage.
    Par exemple, si on utilise un modèle pour de la prédiction alors que le calage s'effectue sur un seul instant de temps

Introduction

Illustration du problème classique de l'*over-fitting* / de la sur-paramétrisation : exemple de l'ajustement de polynomes de degrés différents



Introduction

- Validation “historique” :
  - Principe : séparer les données (observations) en deux ensembles, utiliser l’un pour le calage, l’autre pour la validation
  - E.g. séparation temporelle ($\rightarrow$ validation historique)
  - Mais pas que ça : toute autre partition des données, e.g. spatiale

Les spécificités des simulations à base d'agents

Les simulations :

- sont fréquemment multi-échelles,
- sont basés sur des entités autonomes et hétérogènes en interaction,
- l'environnement peut être fortement dynamique,
- ...

Quelques techniques

- Méthodes formelles (model checking/vérification) pour validation du modèle
- Animation/prototypage pour validation de comportement
- Décomposition du SMA (fonctionnelle, par but, comportementale, temporelle, ...) pour appliquer de manière efficace sur une partie du système
- Exploitation de la connaissance de l'expert (modèle, variables, process, ...)

Validation, point de vue stochastique

Clémentine PRIEUR

MOISE

Validation, point de vue stochastique

1. Validation, une approche bayésienne
2. Validation, une approche par analyse de sensibilité

Validation, une approche bayésienne

La validation peut passer par l'estimation de l'incertitude que l'on a sur la sortie du modèle.

- les modèles font des prédictions mais ... beaucoup d'incertitudes (et eventuellement d'alea) entrent dans la simulation
- il serait mieux d'avoir une idée de l'incertitude sur la sortie
- pourquoi pas en terme de distribution de probabilité sur la quantité d'intérêt ?

approche bayésienne pour la validation

Paradigme bayésien, généralités

principe

△ x^{obs} données observées;

△ modèle paramétrique $\mathcal{M}_0$: $X|\theta \sim L(x;\theta)$;

△ loi a priori $\theta \sim \pi(\theta)$.

Dans le paradigme bayésien, θ le paramètre inconnu n'est pas déterministe inconnu, mais il est est distribué selon une loi de probabilité.

loi a priori $\pi(\theta) \Rightarrow$ enrichie par l'information contenue dans x : $L(x;\theta)$

Paradigme bayésien, généralités

principe

△ x^{obs} données observées;

△ modèle paramétrique $\mathcal{M}_0$: $X|\theta \sim L(x;\theta)$;

△ loi a priori $\theta \sim \pi(\theta)$.

Dans le paradigme bayésien, θ le paramètre inconnu n'est pas déterministe inconnu, mais il est est distribué selon une loi de probabilité.

loi a priori $\pi(\theta) \Rightarrow$ enrichie par l'information contenue dans x : $L(x;\theta)$

Règle de Bayes

Paradigme bayésien, généralités

principe

△ x^{obs} données observées;

△ modèle paramétrique $\mathcal{M}_0$: $X|\theta \sim L(x;\theta)$;

△ loi a priori $\theta \sim \pi(\theta)$.

Dans le paradigme bayésien, θ le paramètre inconnu n'est pas déterministe inconnu, mais il est est distribué selon une loi de probabilité.

loi a priori $\pi(\theta) \Rightarrow$ enrichie par l'information contenue dans x : $L(x;\theta)$

Règle de Bayes

loi a posteriori : $\pi(\theta|x) \propto \pi(\theta)\, L(x;\theta)$.

Une application au modèle UrbanSim

Réf. : Assessing Uncertainty in Urban Simulations Using Bayesian Melding. H. Sevcikova, A. E. Raftery, P. A. Waddell.

Une application au modèle UrbanSim

Réf. : Assessing Uncertainty in Urban Simulations Using Bayesian Melding. H. Sevcikova, A. E. Raftery, P. A. Waddell.

Notations : $Y = f(\mathbf{X})$

- $\theta \subset \mathbf{X}$ ensemble des entrées incertaines,
- Φ sorties pour lesquelles on a de l'observation,
- $\Phi = M_\Phi(\theta)$,
- Ψ quantités qui nous intéressent $\Psi = M_\Psi(\theta, \Phi) = M_\Psi(\theta, M_\Phi(\theta))$,
- les observations qui donnent de l'info sur les sorties sont notées y.

$\rightarrow$ L'information disponible sur les entrées est représentée par un a priori $q(\theta)$.

$\rightarrow$ On spécifie la loi des observations y conditionnellement aux sorties Φ : $L(\Phi) = L(y; \Phi)$. On a alors :

$$[\Phi = M_\Phi(\theta)] \Rightarrow [L(\theta) = L(y; M_\Phi(\theta))] \ .$$

approche bayésienne pour la validation

Application du théorème de Bayes : loi a posteriori pour les entrées $\pi(\theta) \propto q(\theta)L(\theta)$.

Ψ s'exprime à partir des entrées $\Psi = M_\Psi(\theta, M_\Phi(\theta))$, du coup on obtient une loi a posteriori pour la quantité d'intérêt Ψ.

Application à UrbanSim

Données : Eugene-Springfield, Oregon.

Le modèle est lancé à partir de l'année 1980, année pour laquelle on dispose d'informations détaillées pour la ville.

But : prédire le nombre de ménages pour l'année 2000 dans chacune des $K = 295$ zones. On dispose des observations pour l'année 2000, mais on ne les garde que pour valider in fine.

On dispose des observations "nombre de ménages" dans chaque zone pour l'année 1994 $y = (y_1, \ldots, y_K)$.

Application à UrbanSim

Données : Eugene-Springfield, Oregon.

Le modèle est lancé à partir de l'année 1980, année pour laquelle on dispose d'informations détaillées pour la ville.

But : prédire le nombre de ménages pour l'année 2000 dans chacune des $K = 295$ zones. On dispose des observations pour l'année 2000, mais on ne les garde que pour valider in fine.

On dispose des observations "nombre de ménages" dans chaque zone pour l'année 1994 $y = (y_1, \ldots, y_K)$.

Résumé : "initialisation" : 1980, "présent" : 1994, "futur" : 2000.

Vraisemblance et loi a posteriori

$$L(y;\theta_i) = \Pi_{k=1}^{K} L(y_k;\theta_i).$$

Vraisemblance et loi a posteriori

$L(y;\theta_i) = \Pi_{k=1}^{K} L(y_k;\theta_i)$.

Sevcikova *et al.* proposent le modèle suivant :

$$\Phi_{i,j,k} = \mu_{i,k} + \delta_{i,j,k} \text{ avec } \delta_{i,j,k} \sim_{i.i.d.} \mathcal{N}(0, \sigma_\delta^2)$$

$$L(y_k;\theta = \theta_i) = \mu_{i,k} + a + \varepsilon_{i,k} \text{ avec } \varepsilon_{i,k} \sim_{i.i.d.} \mathcal{N}(0, \sigma_i^2).$$

$\phi_{i,j,k}$ (*resp.* $\psi_{i,j,k}$) désigne le nombre de ménages dans la zone k pour le $j^{\text{ème}}$ run et le $i^{\text{ème}}$ jeu de paramètres d'entrée pour l'année 1994 (*resp.* 2000).

Les différents paramètres sont estimés de façon empirique. O obtient alors :

$$L(y_k;\theta_i) \sim \mathcal{N}(\hat{a} + \hat{\mu}_{i,k}, \hat{\sigma}_i^2 + \frac{\hat{\sigma}_\delta^2}{J}).$$

Vraisemblance et loi a posteriori

Les paramètres ont été estimés à $t_1 = 1994$. La loi prédictive du nombre de ménages de la zone k au temps $t_2 = 2000$, Ψ_k, est donnée par :

$$L(\Psi_k) = \sum_{i=1}^{I} \omega_i \mathcal{N}(\hat{a} b_a + m_{i,k}, (\hat{\sigma}_i^2 + \frac{\hat{\sigma}_\delta^2}{J}) b_v), \quad k = 1, \ldots, K,$$

avec $m_{i,k} = \frac{1}{J} \sum_{j=1}^{J} \Psi_{i,j,k}$ et $\omega_i \propto L(y; \theta_i)$.

b_a et b_v sont des facteurs de propagation pour le biais et la variance sur la période $[t_1, t_2]$.

Vraisemblance et loi a posteriori

Les paramètres ont été estimés à $t_1 = 1994$. La loi prédictive du nombre de ménages de la zone k au temps $t_2 = 2000$, Ψ_k, est donnée par :

$$L(\Psi_k) = \sum_{i=1}^{I} \omega_i \mathcal{N}(\hat{a} b_a + m_{i,k}, (\hat{\sigma}_i^2 + \frac{\hat{\sigma}_\delta^2}{J}) b_v), \quad k = 1, \ldots, K,$$

avec $m_{i,k} = \frac{1}{J} \sum_{j=1}^{J} \Psi_{i,j,k}$ et $\omega_i \propto L(y; \theta_i)$.

b_a et b_v sont des facteurs de propagation pour le biais et la variance sur la période $[t_1, t_2]$.
Dans le papier, il y a aussi un paragraphe sur la loi des entrées a priori dans ce cas d'étude.

Base de la validation

Message important : il faut des observations, plus d'observations, pour faire de la validation (au moins 3 années : initialisation, présent, futur).

Validation, point de vue stochastique

1. Validation, une approche bayésienne
2. Validation, une approche par analyse de sensibilité

Validation par AS

Principe de la validation par analyse de sensibilité : proposer différents scénarios pour lesquels on attend une certaine réponse, voir si les paramètres influents sous tel ou tel scénario sont bien ceux attendus.

Validation par AS - un exemple de la litterature

T.W. Nicolai, L. Wang, K. Nagel, P. Waddell Coupling an urban simulation model with a travel model – A first sensitivity test. Working paper, TU Berlin. 2011

- Question posée : Comment l'usage du sol sur une parcelle évolue avec une modification de l'accessibilité de cette parcelle ?
- Modèle UrbanSim, aire métropolitaine de la région Puget Sound (Washington State)
- année de base : 2000 ; fin de la simulation : 2030.
- Analyse de sensibilité = vérifier que le modèle réagit de façon cohérente

Validation par AS - un exemple de la litterature

3 scénarios testés

- Scénario de base : le réseau de transport n'est pas modifié
- Scénario 1 : La connection par ferry est remplacée par un pont à grande capacité
- Scénario 2 : La connection par ferry est remplacée par un pont à faible capacité

Quantités d'intérêt :

- temps de trajet
- accroissement de la population habitant la parcelle
- prix du foncier sur la parcelle
- etc.

Les évolutions de ces quantités d'intérêt en fonction des différents scénarios sont tracées et analysées

Validation par AS - un exemple de la litterature

mais finalement, quel est vraiment le but de l'article ? ...

- Analyser les scénarios et faire de la prospective

  ou ...
- vérifier que le modèle réagit de façon cohérente (comme le suggère la présence du terme AS dans le titre) ?

Validation par AS

Si on veut vraiment vérifier que le modèle réagit de façon cohérente par analyse de sensibilité, il est nécessaire d'introduire des indicateurs numériques.

- définir les scénarios et les quantités d'intérêts.
- connaitre comment ces quantités d'intérêts devraient évoluer (tendances) : dires d'expert
- calculer les indicateurs de sensibilité (ou tendances) de manière numérique
- confronter les dires d'expert avec ces valeurs numériques

Base de la validation

Message important :

Validation historique : il faut des observations, plus d'observations, pour faire de la validation (au moins 3 années : initialisation, présent, futur).

Validation par AS : Ici, ce ne sont pas les données supplémentaires qui sont nécessaires, mais les connaissances des experts, pour définir les scénarios et prédire l'évolution des quantités d'intérêt.