Modélisation et Régularisation avec MaxEnt

Table des matières

1 Introduction

Ce document décrit la modélisation probabiliste de la présence d'une espèce dans un paysage d'intérêt à l'aide de MaxEnt, ainsi que la régularisation et l'interprétation des résultats sous forme de probabilités logistiques.

2 Modélisation

2.1 Définitions

- $y = 1$: présence de l'espèce.
- $y = 0$: absence de l'espèce.
- $\mathbf{z}$: vecteur des covariables (prédicteurs).
- L : paysage d'intérêt.
- $f_1(\mathbf{z})$: densité de probabilité des covariables dans la zone où l'espèce est présente.
- $f_0(\mathbf{z})$: densité de probabilité des covariables dans la zone où l'espèce est absente.
- $f(\mathbf{z})$: densité de probabilité des covariables dans L.

2.2 Objectif

L'objectif est d'estimer la probabilité conditionnelle $\mathbb{P}(y = 1 \mid \mathbf{z})$ sous des *conditions environnementales* :

$$\mathbb{P}(y = 1 \mid \mathbf{z}) = \frac{f_1(\mathbf{z}) \cdot \mathbb{P}(y = 1)}{f(\mathbf{z})},$$

avec :

$$f(\mathbf{z}) = f_1(\mathbf{z}) \cdot \mathbb{P}(y = 1) + f_0(\mathbf{z}) \cdot \mathbb{P}(y = 0).$$

2.3 Remarques

- On peut estimer $f_1(\mathbf{z})$ à partir des données de présence (on calcule la densité dans la zone où l'espèce est présente).
- $f(\mathbf{z})$ est également accessible via un échantillonnage simple.
- Cependant, $\mathbb{P}(y = 1)$ reste inidentifiable avec les seules données de présence.
- D'autres subtilités techniques, telles que la probabilité de détection (*erreurs techniques*), peuvent intervenir. Ces erreurs s'annulent si on utilise des données de présence-absence.

3 MaxEnt (Maximum Entropy)

3.1 Définition du Rapport

MaxEnt cherche à estimer le rapport :

$$\frac{f_1(\mathbf{z})}{f(\mathbf{z})},$$

appelé *suitability* ou *Relative suitability*, qui est la sortie du modèle.

3.2 Hypothèses

Dans le cadre des données de présence seulement :

— Il est impossible de connaître la prévalence de l'espèce (proportion des sites occupés par l'espèce dans le paysage total).

— Pour contourner cette limitation, MaxEnt fixe cette probabilité de présence à 50%. Ce choix arbitraire permet de passer de la *suitability* à une probabilité.

3.3 Score de MaxEnt

On travaille avec le score suivant :

$$g(\mathbf{z}) = \log\left(\frac{f_1(\mathbf{z})}{f(\mathbf{z})}\right).$$

4 Description du Modèle

4.1 Modélisation

— $f(\mathbf{z})$ est obtenue via échantillonnage.

— $f_1(\mathbf{z})$ est estimée à partir des données de présence.

— Sans données de présence, l'espèce est supposée uniformément distribuée sur L, définissant ainsi un modèle nul.

On modélise :

$$\begin{cases} f_1(\mathbf{z}) = f(\mathbf{z})e^{\eta(\mathbf{z})}, \\ \eta(\mathbf{z}) = \alpha + \beta \cdot h(\mathbf{z}), \end{cases}$$

avec :

— α : constante telle que $\int f_1(\mathbf{z}) = 1$,

— β : poids des covariables,

— $h(\mathbf{z})$: vecteur des covariables transformées (*features*).

4.2 Minimisation

La minimisation de la distance entre $f_1(\mathbf{z})$ et $f(\mathbf{z})$ conduit à :

$$\frac{f_1(\mathbf{z})}{f(\mathbf{z})} = e^{\eta(\mathbf{z})}.$$

5 Régularisation de MaxEnt

5.1 Définition

On définit le terme de régularisation :

$$\lambda_j = \lambda \sqrt{\frac{s^2[h_j]}{m}},$$

où :
- $s^2[h_j]$: variance de h_j,
- m : nombre de sites de présence,
- λ : terme de régularisation.

5.2 Optimisation

On maximise :

$$\max_{\alpha,\beta} \left\{ \frac{1}{m} \sum_{i=1}^{m} \ln(f(\mathbf{z}_i)) + \eta(\mathbf{z}_i) - \sum_{j=1}^{n} \lambda_j \beta_j \right\},$$

sous la contrainte :

$$\int_L f(\mathbf{z}) e^{\eta(\mathbf{z})} d\mathbf{z} = 1.$$

6 Logistic Output

Une fois que α et β ($\eta(\mathbf{z})$) sont déterminés, on calcule :

$$\mathbb{P}(y = 1 \mid \mathbf{z}) = \frac{\tau\, e^{\eta(\mathbf{z})-r}}{1 - \tau + \tau\, e^{\eta(\mathbf{z})-r}},$$

avec :
- r : entropie relative (*Kullback-Leibler divergence*) entre $f_1(\mathbf{z})$ et $f(\mathbf{z})$,
- τ : probabilité de présence sous des conditions typiques, par défaut 0.5 (ajustable avec des connaissances supplémentaires sur l'espèce).