Exercice - Session 3

Une start-up technologique Rennaise travaillant pour l'industrie agro-alimentaire et cosmétique développe un outil d'imagerie pour détecter de manière quasi-instantanée des bactéries dans des échantillons de produits.

Sur chaque image, des amas de pixels lumineux apparaissent (que l'on appelle des événements), certains étant identifiés par des experts comme des bactéries (les positifs), d'autres pas (les négatifs). L'expert s'appuie sur des règles de décision qui lui sont propres et qui intègrent la forme de l'amas de pixels lumineux, l'homogénéité des pixels, etc.

Cette start-up souhaite mettre au point un outil d'aide à la décision reproduisant le plus fidèlement possible et de manière automatique l'évaluation de l'expert. Pour cela, elle a constitué un tableau de données rassemblant des caractéristiques géométriques d'événements (diamètre, élongation, luminosité moyenne, médiane, écart-type, etc.) et le classement par l'expert de cet événement en bactérie ou événement indésirable. L'expert a travaillé sur deux images.

Des données d'apprentissage sont disponibles dans le fichier `bacttrain.txt` et des données test dans le fichier `bacttest.txt`.

La problématique est donc la suivante : comment prédire le statut positif ou négatif d'un événement à partir des variables extraites de l'analyse d'image ?

1. Dans cette problématique, quelle est la variable réponse ? Quelles sont les variables explicatives ? Vous donnerez la nature, quantitative ou catégorielle, de ces variables.

2. Proposez un modèle permettant de répondre à la problématique.

3. Donnez une estimation des probabilités d'être un événement positif pour chaque événement contenu dans les données test.

4. Quelle est la proportion de bons classements pour la règle de Bayes appliquée aux probabilités estimées à la question 3 ?

5. L'utilisation d'une méthode d'estimation pénalisée du modèle de régression de la question 2 conduit-elle à une amélioration de la performance de prédiction de la règle de Bayes ?

6. Donnez une sélection de variables que vous recommanderiez pour prédire le statut d'un événement.

7. Cette sélection est-elle sensible à l'échantillonnage (au choix des données d'apprentissage) ?