ENSAI École nationale de la statistique et de l'analyse de l'information

TP 1 Apprentissage supervise 2A

Septembre 2019

Exercice 1 : règle et risque de Bayes en discrimination binaire

Voici trois problèmes de discrimination binaire :

Cas 1 :

On a simulé un échantillon de taille n=1000 selon le modèle :

Pour i=1 à 1000, $X_i \sim \mathcal{N}(0,1)$ $U_i \sim \mathcal{U}[0,1]$ et $Y_i = \begin{cases} \mathbb{1}_{U_i \leq 0.1}, & si \quad X_i \leq 0 \\ \mathbb{1}_{U_i > 0.2}, & si \quad X_i > 0 \end{cases}$

Cas 2 :

On a simulé un échantillon de taille n=1000 selon le modèle :

Pour i=1 à 1000, $X_i \sim \mathcal{N}(0,1)$ $U_i \sim \mathcal{U}[0,1]$ et $Y_i = \begin{cases} \mathbb{1}_{U_i \leq 0.2}, & si \quad X_i \leq 0 \\ \mathbb{1}_{U_i > 0.4}, & si \quad X_i > 0 \end{cases}$

Cas 3 :

On travaille sur des données réelles issues d'une enquête, à partir d'un échantillon tiré au hasard de n=1000 consommateurs de café. La variable à expliquer Y est qualitative binaire et prend les modalités "sucré" et "non sucré". La variable explicative à notre disposition est X qui représente le sexe. On dispose de la table avec en lignes les 1000 consommateurs et en colonnes les 2 variables X et Y. Des résultats de statistique bivariée nous donnent :

- Parmi les femmes, on a 20% qui prennent du sucre dans leur café Y="sucré".
- Parmi les hommes, on a 10% qui prennent du sucre dans leur café Y="sucré".

1. Donner dans chacun des trois cas, si c'est possible, la règle de Bayes et le risque de Bayes.
2. Est-il possible de donner un indicateur de la complexité de ces problèmes et ainsi de les ordonner en fonction de leur complexité ?
3. Simuler les données du cas 2 et créer en plus un échantillon test de taille 200 (avec le même modèle).
4. Quelle est la règle de décision associée à l'algorithme des k plus proches voisins ? Mettre en oeuvre les k plus proches voisins. Justifier le choix des paramètres et commenter les résultats en validation et sur l'échantillon test.
5. Quelle est la règle de décision associée au bayésien naïf ? Mettre en oeuvre le Bayésien naïf. Justifier le choix des paramètres et commenter les résultats en validation et sur l'échantillon test.

Exercice 2 : discrimination à plusieurs classes

On utilisera les K plus proches voisins et Bayésien naif pour discriminer les 4 types de véhicules (variable à expliquer à 4 modalités) à l'aide de différentes variables décrivant la silhouette d'un véhicule.

1. Charger la table Vehicle après avoir chargé le package mlbench. Lire la description de la table. Combien de prédicteurs potentiels dans cette table ?
2. Quelle fonction de coût utiliser ?
3. Quelle est la règle de décision associée à l'algorithme des k plus proches voisins ? Mettre en oeuvre les k plus proches voisins. Justifier le choix des paramètres et commenter les résultats.
4. Quelle est la règle de décision associée au bayésien naïf ? Mettre en oeuvre le Bayésien naïf. Justifier le choix des paramètres et commenter les résultats.

Exercice 3 : Problème de régression

" **Context :** Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019.

Content : This data file includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions.

Acknowledgements This public dataset is part of Airbnb, and the original source can be found on this website."

source : kaggle et Airbnb.

— id : listing ID
— name : name of the listing
— host_id : host ID
— host_name : name of the host
— neighbourhood_group : location
— neighbourhood : area
— latitude : latitude coordinates
— longitude : longitude coordinates
— room_type listing : space type
— price : price in dollars
— minimum_nights : amount of nights minimum
— number_of_reviews : number of reviews
— last_review : latest review
— reviews_per_month : number of reviews per month
— calculated_host_listings_count : amount of listing per host
— availability_365 : number of days when listing is available for booking

La variable à expliquer est ***price***.

1. Quelles variables explicatives proposez-vous d'utiliser si vous êtes un particulier et que vous souhaitez prédire à quel prix vous pourriez proposer votre logement en Airbnb en vous positionnant "au prix du marché".
2. Quelle fonction de coût utiliser ?
3. Quelle est la règle de décision associée à l'algorithme des k plus proches voisins ? Mettre en oeuvre les k plus proches voisins. Justifier le choix des paramètres et commenter les résultats.
4. Quelle est la règle de décision associée au bayésien naïf ? Mettre en oeuvre le Bayésien naïf. Justifier le choix des paramètres et commenter les résultats.