Évaluation, Comparaison et Sélection de Variables Prédictives

Cette notice concerne la question de l'évaluation, la comparaison, l'ordonnancement et la sélection des variables prédictives lors d'une application de classification supervisée.

1. Évaluations Basées sur l'Entropie

Le calcul d'*Entropie* sur un ensemble de données fournit une valeur numérique indiquant la quantité d'information disponible dans les données afin de prédire la classe. Plus la valeur obtenue est élevée, plus l'apprentissage d'un modèle de prédiction sera difficile.

Les mesures *Information Gain*, *Gain Ratio*, *Symmetrical Uncertainty* et *Relief* calculent pour chaque variable prédictive une valeur indiquant à quel point cette variable sera utile pour prédire la classe, c-à-d sa contribution à réduire l'*Entropie*. Plus la valeur obtenue est élevée, plus la variable sera utile.

2. Évaluations des Liens de Corrélation et Covariance

Les mesures de *corrélation* entre deux variables numériques continues donnent une valeur réelle décrivant la force du lien de dépendance entre les valeurs des variables. Une valeur positive indique une dépendance positive, une valeur de 0.0 indique l'indépendance et une valeur négative indique une dépendance négative. Plus la valeur (positive ou négative) obtenue est élevée, plus la dépendance entre variables est forte.

La mesure de *covariance* entre deux variables numériques continues génère une valeur exprimant à quel point les valeurs des variables évoluent de manière identique, c-à-d si une augmentation (resp. diminution) des valeurs de l'une implique une augmentation (resp. diminution) des valeurs de l'autre. Plus la valeur obtenue est élevée, plus ce lien linéaire est fort. Une valeur négative indique un comportement inverse (l'augmentation de l'une implique la diminution de l'autre).

La force des liens de corrélation (plus général) ou covariance (lien linéaire) entre deux variables peut indiquer la présence d'une « redondance » parmi les variables prédictives. Par exemple, la présence d'une variable dont les valeurs sont calculées directement à partir des valeurs de l'autre variable. Cela équivaut à avoir l'information correspondante représentée deux fois dans les données, sur des échelles de valeurs ou avec des représentations différentes.

Les mesures de *corrélation* entre deux variables discrètes (catégorielles, ordinales, binaires) fournissent principalement une valeur de probabilité d'indépendance entre les valeurs des variables, appelée *p-value*. Une valeur inférieure à 0.05 (seuil de 5% usuel) suppose une dépendance entre les variables, et une valeur supérieure suppose une indépendance. Plus la valeur obtenue est faible, plus la probabilité d'une dépendance entre les variables est importante. Les mesures les plus communes sont les tests du χ^2, de *Fisher* et de *Yule*.

3. Librairies et Fonctions du Logiciel R

Les méthodes classiques d'évaluation des variables prédictives selon leur utilité pour la prédiction des classes consistent à réaliser :

- Le calcul de mesures d'évaluation basées sur l'*Entropie* pour chaque variable prédictive.
- Le calcul de tests de corrélation, et éventuellement covariance, entre variables.
- Le filtrage des variables prédictives inutiles ou le filtrage des variables prédictives les moins utiles en présence de très nombreuses variables (cf. problème de *Curse of dimensionality*).

Plusieurs librairies R fournissent différentes fonctions pour ces opérations :

- Librairie **FSelectorRcpp** (https://www.rdocumentation.org/packages/FSelectorRcpp/) :
  - Fonction `information_gain()` de calculs des mesures d'évaluation de l'utilité des variables prédictives par *Information Gain*, *Gain Ratio* et *Symmetrical Uncertainty* basées sur l'*Entropie*.
  - Fonction `relief()` de calcul de la mesure d'évaluation *Relief* d'utilité des variables prédictives basée sur l'*Entropie* et la répartition des classes des exemples par voisinage dans l'espace des données.
  - Fonction `cut_attrs()` de filtrage des variables prédictives les moins utiles en fonction des résultats d'une mesure d'évaluation d'utilité ci-dessus.
- Librairie **stats** (https://www.rdocumentation.org/packages/stats/) de RBase :

- Fonction `cor()` de calcul des mesures de *corrélation* de *Pearson* (par défaut), *Kendall* ou *Spearman* entre deux variables numériques continues.
- Fonction `cov()` de calcul de la mesure de *covariance* entre deux variables numériques continues.

• Librairie **questionr** (https://www.rdocumentation.org/packages/questionr/) :
- Fonction `rprop()` de représentation sous forme de pourcentages des valeurs d'une matrice de contingence entre deux variables discrètes.
- Fonction `chisq.residual()` de calcul des valeurs du χ^2 pour l'évaluation de la corrélation entre les valeurs de deux variables discrètes, à partir de la matrice de contingence en pourcentages.
- Fonction `fisher.test()` de calcul des valeurs d'évaluation par *Fisher's Test* de la corrélation entre les valeurs de deux variables discrètes, à partir des vecteurs des variables.

• Librairie **graphics** (https://www.rdocumentation.org/packages/graphics/) de RBase :
- Fonction `mosaicplot()` d'affichage graphique des résultats de l'évaluation par *Fisher's Test* de la corrélation entre les valeurs de deux variables discrètes.

• Librairie **psych** (https://www.rdocumentation.org/packages/psych/) :
- Fonction `YuleCor()` de calcul de mesure de *corrélation* entre deux variables binaires uniquement.