

BROSSARD Florian
Master 1 ILC
Fouille de données

Compte rendu Titanic

Table des matières

I)	Préparation des données	3
1.	Les données originales	3
2.	Rendre les données lisible par WEKA.....	3
3.	Nettoyer les données	4
II)	Prédictions.....	5
1.	Choix du classifieur	5
2.	Test du perceptron multicouche	5
3.	Test de J48.....	5
4.	Le modèle donné par J48	6
III)	Envois des prédictions sur Kaggle	6

I) Préparation des données

1. Les données originales

Pour ce TP, il nous a été recommandé d'utiliser le logiciel WEKA afin de pouvoir faire des prédictions sur les données fournis par le site Kaggle.

Les données d'origines n'étant pas directement compatibles il a fallu les nettoyer.

Ces fichiers sont au format CSV, en utilisant le séparateur ';' et en mettant les chaîne des caractères entre '"'. Chacun de ces fichiers contiennent 12 colonnes représentant les attributs de chaque passager :

- PassengerId : l'identifiant du passager
- Survived : indique si le passage à survécu (1) ou non (0)
- Pclass : la classe du passage (1ère, 2ème et 3^{ème} classe, respectivement 1,2 et 3)
- Name : Le nom du passage, une chaîne de caractère entre '"'
- Sex : le sexe du passage (male ou female)
- Age : l'âge du passager
- SibSp : nombre de (demi-)frères, (demi-)sœur, épouse/mari
- Parch : nombre de père, mère, (beau) fils, (belle) fille
- Ticket : numéro du ticket
- Fare : tarif du ticket
- Cabin : cabine du passager
- Embarked : pont d'embarcation

2. Rendre les données lisible par WEKA

La première étape consiste à rendre le fichier CSV lisible par WEKA, on peut voir directement d'où vient le problème en ouvrant le fichier avec un éditeur de texte :

La colonne d'attribut 'Name' contient des chaînes de caractère comme celle-ci : "John, Mr. Smith".

Or WEKA fait le lien entre le nombre de champ sur la première ligne (les noms des attributs) et le nombre de champ sur les autres lignes (les instances). Ici, WEKA comprend qu'il y a pour chaque instance un champ supplémentaire du fait du formatage de ce champ, il faut donc ou le modifier, ou le supprimer.

Étant donné que le nom d'une personne à peu de chance d'influer sur le fait qu'il survive ou pas, j'ai décidé de supprimer ce champ.

Le fichier étant maintenant visible, toute les manipulations peuvent se faire avec WEKA.

3. Nettoyer les données

La seconde étape consiste à supprimer les champs non-pertinent. Parmi ceux-ci j'ai jugé que ces champs n'étaient pas nécessaires durant le traitement des données :

- PassengerId : cet attribut ne sera qu'utile que lors de l'envoi de résultats au site Kaggle et peu se retrouver très facilement avec un logiciel comme Excel car il n'y a qu'une incrémentation à partir de l'Id du premier passager.
- Fare : quasiment tous les tarifs ont une valeur différente, il est donc peu probable que cet attribut soit utile
- Ticket : tous les tickets ont une valeur différente en revanche on pourrait se demander si le numéro de ticket peut éventuellement donner des informations sur l'emplacement de la cabine du passager, cependant rien ne garantis que le dit passager soit effectivement dans sa cabine lors de l'incident.
- Cabin : cet attribut n'est qu'assez peu renseigné (204 sur 891), on ne peut donc pas l'utiliser
- Embarked : Il est peu probable que le pont d'embarcation (sauf malédiction avérée) ai une quelconque incidence sur les chances de survie d'un individu.

Ces 5 attributs étant éliminés, il ne nous reste plus que les attributs Survived, Pclass, Sex, Age , SibSp et Parch.

Sur ces attributs on peut remarquer que Survived ne vaut que 0 ou 1, Pclass 1,2 ou 3 et Sex male ou female. On peut donc utiliser un filtre pour les convertir en tant qu'attributs nominal en ayant au préalable enregistré nos données dans un fichier au format ARFF, utilisé par WEKA.

De même que nos données d'entrainements, il faut préparer les données de test. Le fichier de tests subis donc le même traitement (suppression des colonnes Name, PassengerId, Fare, Ticket, Cabin et Embarked et conversion des colonnes Pclass et Sex en nominal) mais avec l'ajout de la colonnes Survived, sans mettre de valeur. WEKA se chargera de remplir la colonne Survived du jeu de test, que nous enverrons ensuite sur Kaggle.

Les données sont maintenant préparé nous pouvons donc commencer à faire des prédictions.

II) Prédiction

1. Choix du classifieur

Suite aux cours et à différents tests nous avons déjà connaissance de différents algorithmes et méthode permettant de faire de prédiction. De plus nos prédictions porteront sur l'attribut Survived, attribut nominal.

Deux choix s'offrent directement à nous :

- L'algorithme J48
- Le perceptron multicouche

2. Test du perceptron multicouche

Comme vu en cours le perceptron multicouche utilise un réseau de neurones. Nous allons utiliser la configuration de base proposée par WEKA et ne modifier que le nombre de neurones :

- 3 neurones (par défaut) : 82.15% de réussite
- 4 neurones (nombre de colonnes de données) : 82.04%
- 10 neurones (plus de neurones => plus efficace ?) : 81.48%

On peut constater que le meilleur taux de réussite s'obtient avec 3 neurones.

Comme l'on pouvait s'y attendre, plus de neurones n'implique pas forcément une meilleure précision mais en revanche, cela fait croître grandement le temps d'entraînement du perceptron : 15s avec 10 neurones contre 0.8s avec 3 neurones pour construire le modèle dans WEKA.

3. Test de J48

Cet algorithme nous a été présenté lors d'un précédent TP. Il se base sur le fonctionnement des arbres de décision.

Son efficacité sans modifier de paramètre atteint 82.04% de réussite, soit le même taux de réussite qu'avec un perceptron multicouche de 4 neurones. En revanche la construction du modèle est bien plus rapide.
















Du fait de notre meilleure compréhension du résultat fournis par J48, et de la faible différence de réussite entre J48 et le perceptron multicouche de 3 neurones, nous avons décidé d'utiliser l'algorithme J48.

4. Le modèle donné par J48

Voici le résultat donné par l'algorithme :

```
Sex = male
| Age <= 13
| | SibSp <= 2
| | | Parch <= 0: 0 (11.47/2.14)
| | | Parch > 0: 1 (19.25/0.08)
| | SibSp > 2: 0 (16.41/1.0)
| Age > 13: 0 (529.87/86.69)
Sex = female
| SibSp <= 2
| | Parch <= 1: 1 (247.0/50.0)
| | Parch > 1
| | | Parch <= 3: 1 (39.0/10.0)
| | | Parch > 3: 0 (7.0/1.0)
| SibSp > 2
| | Age <= 16: 0 (14.0/1.0)
| | Age > 16: 1 (7.0/2.0)
```

III) Envois des prédictions sur Kaggle

Overview	Data	Kernels	Discussion	Leaderboard	Rules	Team	My Submissions	Submit Predictions
3137	▲1583	Akshaykumar Salunke		0.78947	5	4h		
3138	new	Gordon Zheng		0.78947	12	6h		
3139	new	Petr Sykora		0.78947	16	5h		
3140	new	Eric Hamers		0.78947	11	4m		
3141	new	Florian BROSSARD		0.78947	1	now		
Your Best Entry ↑							Your submission scored 0.78947, which is not an improvement of your best score. Keep trying!	
3142	▼399	LWHuang		0.78468	7	2mo		
3143	▼399	Jason Mak		0.78468	10	2mo		
3144	▼399	Julia Telia		0.78468	1	2mo		
3145	▼399	wklamad		0.78468	7	2mo		
3146	▼399	dvaliu		0.78468	9	2mo		
3147	▼399	cherrycht		0.78468	5	2mo		
3148	▼399	MirageChung		0.78468	1	2mo		
3149	▼399	xiaobai		0.78468	17	1mo		
3150	▼399	TOM LAU		0.78468	21	1mo		
3151	▼399	Alice Tong		0.78468	10	2mo		