Lorsqu'on demande à `Weka` de construire un classifieur, il produit en même temps un ensemble de mesures de l'erreur estimée. On se propose ici d'expliquer comment sont définies, calculées et utilisées ces différentes mesures, et dans quels cas on peut en préférer une à une autre.

# 1 Les valeurs calculées par `Weka`

La figure 1 représente les statistiques que `Weka` retourne par défaut lors de la construction d'un classifieur.

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         148               98.6667 %
Incorrectly Classified Instances         2                1.3333 %
Kappa statistic                          0.98
Mean absolute error                      0.0248
Root mean squared error                  0.0911
Relative absolute error                  5.5779 %
Root relative squared error             19.3291 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  1         0          1         1         1        Iris-setosa
  0.98      0.01       0.98      0.98      0.98     Iris-versicolor
  0.98      0.01       0.98      0.98      0.98     Iris-virginica

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 49  1 |  b = Iris-versicolor
  0  1 49 |  c = Iris-virginica
```

FIG. 1 – Statistiques de `Weka`

# 2 Les mesures générales

On reprend dans cette section les premières mesures données par `Weka`

## 2.1 Correctly Classified Instances

Le nombre d'exemples bien classés, en valeur absolue, puis en pourcentage du nombre total d'exemples.

## 2.2 Incorrectly Classified Instances

Sous le même format, le nombre d'exemples mal classés.

## 2.3 Kappa statistic

Le *coefficient Kappa* est censé mesuré le degré de concordance de deux ou de plusieurs *juges*. Dans `Weka`, on est toujours dans le cas de deux juges. On mesure la différence entre l'accord constaté entre les deux juges, et l'accord qui existerait si les juges classaient les exemples au hasard.

Dans `Weka`, le jugement, c'est la classe d'un exemple, et les deux juges sont le classifieur et la classe réelle de l'exemple.

L'accord/désaccord entre les deux juges se lit directement dans la matrice de confusion : c'est une mesure dont la valeur est d'autant plus grande que la matrice est diagonale.

Le *coefficient Kappa* se calcule de la façon suivante :

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

avec $P_o$ : La proportion de l'échantillon sur laquelle les deux juges sont d'accord (i.e. la diagonale principale de la matrice de confusion).

et

$$P_e = \frac{\sum_i p_i p_{.i}}{n^2}$$

où

– $p_i$ : somme des éléments de la ligne $i$
– $p_{.i}$ : somme des éléments de la colonne $i$
– $n$ : taille de l'échantillon

Sur l'exemple de la figure 1, dont la matrice de confusion était :

| | a | b | c | total |
|---|---|---|---|---|
| | 50 | 0 | 0 | 50 |
| | 0 | 49 | 1 | 50 |
| | 0 | 1 | 49 | 50 |
| total | 50 | 50 | 50 | |

On a :

$$P_0 = \frac{50 + 49 + 49}{150} = \frac{148}{150} = 0.98667$$

$$P_e = \frac{(50 \times 50) + (50 \times 50) + (50 \times 50)}{150 \times 150} = \frac{1}{3}$$

(cette valeur est constante pour tous les classifieurs sur cet ensemble d'exemples !)

et donc

$$\kappa = \frac{0.986667 - 0.333}{0.666} = 0.98$$

Le coefficient Kappa prend ses valeurs entre -1 et 1.

- Il est maximal quand les deux jugements sont les mêmes : tous les exemples sont sur la diagonale, et $P_0 = 1$
- Il vaut 0 lorsque les deux jugements sont indépendants ($P_0 = P_e$)
- Il vaut -1 lorsque les juges sont en total désaccord

Certains auteurs (Landis & Koch) ont proposé une échelle de degré d'accord selon la valeur du coefficient :

| Accord | Kappa |
|---|---|
| Excellent | >0.81 |
| Bon | 0.80-0.61 |
| Modéré | 0.4-0.41 |
| Médiocre | 0.4-0.21 |
| Mauvais | 0.20-0.0 |
| Très mauvais | <0 |

Quelques pages (en français) où je suis allé chercher ces informations : `http://kappa.chez.tiscali.fr/`

## 2.4 Mean absolute error

Erreur absolue en moyenne : pour chaque exemple, on calcule la différence entre la probabilité (calculée par le classifieur) pour un exemple d'appartenir à sa véritable classe, et sa probabilité initiale d'appartenir à la classe qui lui a été fixée dans l'ensemble d'exemples (en général, cette probabilité vaut 1). On divise ensuite la somme de ces erreurs par le nombre d'instances dans l'ensemble d'exemples.

Plus formellement :

- Soient $p_1, p_2, \cdots, p_n$ les probabilités calculées par le classifieur pour chaque exemple d'appartenir à sa *vrai classe*.
- Soient $a_1, a_2, \cdots, a_n$ les probabilités à priori pour chaque exemple d'appartenir à la classe qui leur a été fixée par définition (en général, les $a_i$ valent toujours 1, mais on peut imaginer qu'on soit un peu moins catégorique, et que la classe attribuée ne le soit qu'avec une certaine confiance).
- Alors on calcule :

$$\text{Mean Absolute Error} = \frac{|p_1 - a_1| + |p_2 - a_2| + \cdots + |p_n - a_n|}{n}$$

*A mon avis, il y a un bug dans* `Weka`*, et le calcul réellement effectué ne correspond pas à ce qui est écrit dans le livre.*

Dans le cas où le classifieur est un *prédicteur*, c'est-à-dire qu'il retourne une valeur réelle au lieu d'une classe discrète, c'est la différence entre la valeur

calculée et la valeur attendue qui sont utilisées pour $p_i$ et $a_i$ ; ça peut par exemple être le cas pour les réseaux de neurones.

### 2.5 Root mean-squared error

*Cette mesure d'erreur concerne principalement les* prédicteurs

Racine carrée de l'erreur quadratique moyenne : avec les mêmes notations que ci-dessus, elle correspond à :

$$\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}}$$

L'erreur quadratique avantage les solutions où il y a beaucoup de petits écarts, par rapport à celles qui sont exactes presque partout, mais qui font de grosses erreurs en un petit nombre de points.

Le fait de prendre la racine carrée permet de manipuler des quantités qui ont la même dimensions que les valeurs à prévoir.

### 2.6 Relative absolute error

*Cette mesure d'erreur concerne principalement les* prédicteurs

Erreur absolue relative : le nom paraît très mal choisi ...

On compare l'erreur absolue avec l'erreur absolue d'un prédicteur très simple, qui retournerait toujours la valeur moyenne des $a_i$, soit $\bar{a} = \frac{1}{n}\sum_i a_i$ :

$$\frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{|a_1 - \bar{a}| + \cdots |a_n - \bar{a}|}$$

### 2.7 Root relative squared error

*Cette mesure d'erreur concerne principalement les* prédicteurs

Racine carrée de l'erreur quadratique relative : rapport entre l'erreur quadratique et ce que serait l'erreur quadratique d'un prédicteur qui retournerait toujours la valeur moyenne des $a_i$ :

$$\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \cdots (a_n - \bar{a})^2}}$$

## 3 Les mesures d'exactitude par classe

Ces valeurs se trouvent dans la partie "Detailed Accuracy By Class". Pour chaque classe, `Weka` fournit cinq mesures.

Pour une classe donnée, un classifieur, et un exemple, quatre cas peuvent se présenter :

1. L'exemple est de cette classe, et le classifieur ne se trompe pas : c'est un *vrai positif*.

| | Classe prédite | |
|---|---|---|
| | prédite | non prédite |
| Exemples de cette classe | Vrais positifs | Faux négatifs |
| Exemple n'appartenant pas à cette classe | Faux positifs | Vrais négatifs |

FIG. 2 – Les quatre cas pour une classification binaire

2. L'exemple est de cette classe, mais le classifieur se trompe : c'est un *faux négatif*.
3. L'exemple n'est pas de cette classe, mais le classifieur la lui attribue quand même : c'est *faux positif*.
4. L'exemple n'est pas de cette classe, et le classifieur ne le range pas non plus dans cette classe : c'est un *vrai négatif*.

Le tableau de la figure 2 résume les différentes situations.

### 3.1 TP Rate

Rapport des vrais positifs. Il correspond à :

$$\frac{\text{Nombre de vrais positifs}}{\text{Nombre de vrais positifs} + \text{Nombre de faux négatifs}} = \frac{\text{Nombre de vrais positifs}}{\text{Nombre d'exemples de cette classe}}$$

C'est donc le rapport entre le nombre de bien classé et le nombre total d'éléments qui devraient être bien classés. Il se calcule en utilisant la première ligne du tableau de la figure 2.

### 3.2 FP Rate

Rapport des faux positifs. Il correspond, symétriquement à la définition précédente, à :

$$\frac{\text{Nombre de faux positifs}}{\text{Nombre de faux positifs} + \text{Nombre de vrais négatifs}} = \frac{\text{Nombre de faux positifs}}{\text{Nombre d'exemples n'étant pas de cette classe}}$$

Il se calcule en utilisant la deuxième ligne du tableau de la figure 2.

La donnée des taux `TP Rate` et `FP Rate` permet de reconstruire la matrice de confusion pour une classe donnée.

Symétriquement, la matrice de confusion permet de calculer `TP Rate` et `FP Rate`. Prenons l'exemple de la figure 1 :

- Les cinquante exemples de classe `Iris-setosa` sont bien classés (pas de faux négatifs) : donc `TP Rate`=1.
- 49 `iris-versicolor` sont bien classés, mais le dernier est mal classé : `TP Rate`=$\frac{49}{50} = 0.98$

- 2 exemples sont classés à tort parmi les `Iris-versicolor`, mais 98 exemples, qui ne sont pas des `Iris-versicolor`, n'ont pas été reconnus comme tels : `FP Rate`=$\frac{2}{98} = 0.0204$
- Un exemple est classé à tort parmi les `Iris-virginica`, pour 99 autres sur lesquels on ne s'est pas trompé : `FP Rate`= $\frac{1}{99} = 0.0101$
- 48 `iris-virginica` sont bien classés, sur un total de 50 : `TP Rate`= $\frac{48}{50} = 0.96$

Les notions de `Precision`, `Recall`, et `F-Measure` se rencontre plutôt dans le domaine de la classification de textes.

Les algorithmes de classification de textes peuvent être utilisés pour trouver tous les articles susceptibles d'intéresser leur utilisateur. Il est alors important de savoir :

- Si tous les articles intéressants ont été trouvés, si on n'en a pas oublié.
- Si tous les articles proposés à l'utilisateur sont pertinents. Combien lui en a-t-on proposés qui en fait ne l'intéresseront pas ?

La première notion est couverte par `Recall`, et correspond aussi exactement à `TP Rate`.

La seconde correspond à `Precision`.

On souhaite parfois obtenir une mesure globale regroupant ces deux valeurs : on définit alors la `F-Measure`

## 3.3 Precision

C'est le rapport entre le nombre de vrais positifs et la somme des vrais positifs et des faux positifs (la première colonne du tableau de la figure 2). Une valeur de 1 exprime le fait que tous les exemples classés positifs l'étaient vraiment.

## 3.4 Recall

Un `Recall` de 1 signifie que tous les exemples positifs ont été trouvés.

## 3.5 F-Measure

Cette quantité permet de regrouper en un seul nombre les performances du classifieur (pour une classe donnée) pour ce qui concerne le `Recall` et la `Precision` :

$$\text{F-Measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$