Prédiction de scores de golf
-
Algorithme SAS

Sylvain Demars

Club de Québec
des utilisateurs SAS

29 octobre 2015

Qu'est-ce qu'Actulab ?

Une solution pour connecter les milieux pratiques et académiques,

en organisant des évènements d'**innovation ouverte**.

Qu'est-ce qu'Actulab ?

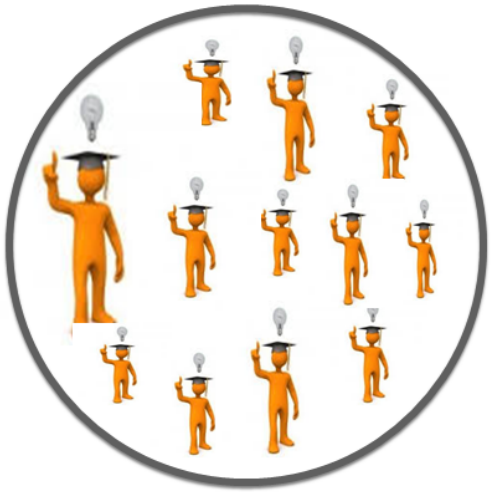
Une solution pour connecter les milieux pratiques et académiques,

en organisant des évènements d'**innovation ouverte**.

Problèmes d'industrie

Actulab

Plusieurs solutions créatives



Milieu universitaire

Pouvez-vous prédire mon pointage au golf ?

A tout moment durant son parcours, un joueur souhaite prédire son score final à partir des points qu'il a réalisé jusqu'à maintenant :

	③	④	⑤	⑥	⑦
joueur 1	4	3	4	7	4

	⑯	⑰	⑱	①	②	③	④	⑤
joueur 2	6	3	3	4	4	5	3	4

Le **score final** est la somme des 18 scores bruts obtenus durant le parcours.

Jeu d'apprentissage

Le **jeu d'apprentissage** contenait 2000 joueurs ayant réalisé le même parcours et dont tous les scores étaient connus :

	①	②	③	...	...	⑰	⑱	Score final
joueur 1	4	5	3	...	...	6	4	75
joueur 2	7	5	6	...	...	6	7	90
joueur 3	3	5	3	...	...	4	4	71
⋮				⋮				⋮
⋮				⋮				⋮
joueur 2000	5	3	4	...	...	5	4	84

Jeu de validation

Le **jeu de validation** contenait 600 joueurs ayant réalisé le même parcours que les 2000 précédents mais dont seule une partie des scores étaient connus :

	①	②	③	...	...	⑰	⑱	Score final
joueur 2001	3	3	7	...	...	-	-	?
joueur 2002	-	-	6	...	...	-	-	?
joueur 2003	-	-	-	...	...	6	3	?
⋮				⋮				⋮
⋮				⋮				⋮
joueur 2600	6	4	5	...	...	-	-	?

L'objectif est d'apporter une **prédiction aux scores finaux** pour ces 600 nouveaux joueurs.

Principe du plus proche voisinage

Pour un nouvel individu fixé, on se restreint aux trous complétés, que l'on compare aux 2000 joueurs de la base d'apprentissage.

	③	④	⑤	⑥	⑦	Score final
joueur 2001	4	3	4	7	4	?

	③	④	⑤	⑥	⑦	Score final
joueur 1	4	5	3	4	5	75
joueur 2	3	5	3	4	3	71
joueur 3	4	3	4	6	4	90
⋮			⋮			⋮
joueur 2000	5	3	4	4	6	84

Son score prédit sera basé sur le score des joueurs de la base qui lui ressemblent le plus !

Principe du plus proche voisinage

Pour un nouvel individu fixé, on se restreint aux trous complétés, que l'on compare aux 2000 joueurs de la base d'apprentissage.

	③	④	⑤	⑥	⑦	Score final
joueur 2001	4	3	4	7	4	?

	③	④	⑤	⑥	⑦	Score final
joueur 1	4	5	3	4	5	75
joueur 2	3	5	3	4	3	71
joueur 3	4	3	4	6	4	90
⋮			⋮			⋮
joueur 2000	5	3	4	4	6	84

Son score prédit sera basé sur le score des joueurs de la base qui lui ressemblent le plus !

Eléments du modèle

Les choix suivants ont été déterminés à l'aide d'un processus de validation croisée.

- Distance de Manhattan.

	③	④	⑤	⑥	⑦	distance	Score final
joueur 2001	4	3	4	7	4	-	?
joueur 1	4	5	3	4	5	7	75
joueur 2	3	5	3	4	3	8	71
joueur 3	4	3	4	6	4	1	90
joueur 4	3	3	4	7	4	1	88

- Moyenne de tous les joueurs dont la distance est minimale.

Eléments du modèle

Les choix suivants ont été déterminés à l'aide d'un processus de validation croisée.

- Distance de Manhattan.

	③	④	⑤	⑥	⑦	distance	Score final
joueur 2001	4	3	4	7	4	-	?
joueur 1	4	5	3	4	5	7	75
joueur 2	3	5	3	4	3	8	71
joueur 3	4	3	4	6	4	1	90
joueur 4	3	3	4	7	4	1	88

- Moyenne de tous les joueurs dont la distance est minimale.

Calcul de distance

1. Réunir le nouveau joueur avec la base d'apprentissage.

```
data compar;
   set joueur training;
run;
```

Calcul de distance

1. Réunir le nouveau joueur avec la base d'apprentissage.

```
data compar;
   set joueur training;
run;
```

2. Calcul de la distance.

```
proc distance data=compar method=CITYBLOCK
              out=dist(keep=dist1);
    var interval(H1--H18);
run;
```

Calcul de distance

1. Réunir le nouveau joueur avec la base d'apprentissage.

```
data compar;
   set joueur training;
run;
```

2. Calcul de la distance.

```
proc distance data=compar method=CITYBLOCK
              out=dist(keep=dist1);
    var interval(H1--H18);
run;
```

```
data compar;
    merge compar dist;
    if ID="1B" then delete;
run;
```

Dist1
0
59
49
53
54
57
53
48
51
62

Ajustement de la distance

	...	②	③	④	⑤	⑥	⑦	⑧	...	distance
joueur 2001	...	0	4	3	4	7	4	0	...	
joueur 1	...	6	4	5	3	4	5	4	...	59

Ajustement de la distance

	...	②	③	④	⑤	⑥	⑦	⑧	...	distance
joueur 2001	...	0	4	3	4	7	4	0	...	
joueur 1	...	6	4	5	3	4	5	4	...	59

- Récupération des n° des trous non joués.

```
proc transpose data=joueur out=joueur_t; run;
proc sql noprint;
    select _NAME_ into :list_nonjoue separated by ','
           from joueur_t(where=(COL1 = 0));
run;
```

Ajustement de la distance

	...	②	③	④	⑤	⑥	⑦	⑧	...	distance
joueur 2001	...	0	4	3	4	7	4	0	...	
joueur 1	...	6	4	5	3	4	5	4	...	59

- Récupération des n° des trous non joués.

```
proc transpose data=joueur out=joueur_t; run;
proc sql noprint;
    select _NAME_ into :list_nonjoue separated by ','
           from joueur_t(where=(COL1 = 0));
run;
```

- On ajuste la distance précédente.

```
data compar (keep=id somme_nonjoue dist_joue);
    set compar;
    SOMME_NONJOUE = sum(&list_nonjoue.)
    DIST_JOUE = DIST1-SOMME_NONJOUE;
run;
```

Le voisinage

3. Identification des joueurs les plus proches.

```
proc sql noprint;
   select min(DIST_JOUE) into :mini
   from compar;
run;
```

4. Moyenne de ces joueurs **sur les trous non joués**.

```
proc means data=compar(where=(DIST_JOUE=&mini.)) mean;
    var SOMME_NONJOUE;
    output out=resultat mean=PRED_NONJOUE;
run;
```

5. Prédiction finale = score actuel
   + prédiction sur les trous non joués.

Conclusion

- Plus le joueur a de données manquantes, plus la prédiction tend vers la moyenne.
- La **base d'apprentissage** impacte très fortement la prédiction !

Conclusion

- Plus le joueur a de données manquantes, plus la prédiction tend vers la moyenne.
- La **base d'apprentissage** impacte très fortement la prédiction !

Conclusion

- Plus le joueur a de données manquantes, plus la prédiction tend vers la moyenne.
- La **base d'apprentissage** impacte très fortement la prédiction !

Merci pour votre attention.