Bases de Données II, Charleroi

Jef Wijsen

18 janvier 2013

Cahier ouvert. Durée : 1 heure

Nom et prénom
Année

La figure 2 montre une base de données XML pour stocker les podiums des courses cyclistes annuelles. Chaque coureur est identifié par un code unique (attribut `id`). Pour enregistrer les podiums, on utilise le *document order* d'XML ; par exemple, le gagneur du Tour des Flandres en 2011 était Nick Nuyens, avant Sylvain Chavanel (deuxième) et Fabian Cancellara (troisième).

Les noms des courses ne contiennent pas de blancs ou symboles spéciaux à l'intérieur.

La figure 1 montre le DTD.

```
<!-- This file is called courses.dtd -->
<!ELEMENT CYCLISME (COUREURS, COURSES)>
<!ELEMENT COUREURS (COUREUR*)>
<!ELEMENT COURSES (COURSE*)>
<!ELEMENT COURSE (PODIUM*)>
<!ELEMENT PODIUM (COUREUR*)>
<!ELEMENT COUREUR (#PCDATA)>
<!ATTLIST COUREUR id CDATA #REQUIRED>
<!ATTLIST COUREUR naissance CDATA #IMPLIED>
<!ATTLIST COUREUR nat CDATA #IMPLIED>
<!ATTLIST COURSE nom CDATA #REQUIRED>
<!ATTLIST PODIUM annee CDATA #REQUIRED>
```

Figure 1 – DTD.

```
<CYCLISME>
<COUREURS>
   <COUREUR id="tb" naissance="1980" nat="B">Tom Boonen</COUREUR>
   <COUREUR id="pg" naissance="1982" nat="B">Philippe Gilbert</COUREUR>
   <COUREUR id="nn" naissance="1980" nat="B">Nick Nuyens</COUREUR>
   <COUREUR id="sc" naissance="1979" nat="F">Sylvain Chavanel</COUREUR>
   <COUREUR id="fc" naissance="1981" nat="CH">Fabian Cancellara</COUREUR>
   <COUREUR id="fp" naissance="1981" nat="I">Filippo Pozzato</COUREUR>
   <COUREUR id="ab" naissance="1979" nat="I">Alessandro Ballan</COUREUR>
   <COUREUR id="jv" naissance="1981" nat="B">Johan Vansummeren</COUREUR>
   <COUREUR id="mt" naissance="1977" nat="NL">Maarten Tjallingii</COUREUR>
   <COUREUR id="st" naissance="1984" nat="F">Sebastien Turgot</COUREUR>
   <COUREUR id="vn" naissance="1984" nat="I">Vincenzo Nibali</COUREUR>
   <COUREUR id="sg" naissance="1980" nat="AUS">Simon Gerrans</COUREUR>
   <COUREUR id="ib" naissance="1977" nat="I">Ivan Basso</COUREUR>
</COUREURS>
<COURSES>
<COURSE nom="TourDesFlandres">
   <PODIUM annee="2011">
      <COUREUR id="nn"/><COUREUR id="sc"/><COUREUR id="fc"/>
   </PODIUM>
   <PODIUM annee="2012">
      <COUREUR id="tb"/><COUREUR id="fp"/><COUREUR id="ab"/>
   </PODIUM>
</COURSE>
<COURSE nom="ParisRoubaix">
   <PODIUM annee="2011">
      <COUREUR id="jv"/><COUREUR id="fc"/><COUREUR id="mt"/>
   </PODIUM>
   <PODIUM annee="2012">
      <COUREUR id="tb"/><COUREUR id="st"/><COUREUR id="ab"/>
   </PODIUM>
</COURSE>
<COURSE nom="MilanSanRemo">
   <PODIUM annee="2012">
      <COUREUR id="sg"/><COUREUR id="fc"/><COUREUR id="vn"/>
   </PODIUM>
</COURSE>
</COURSES>
</CYCLISME>
```

FIGURE 2 – Fichier XML avec des informations sur les podiums des courses cyclistes.

Question 1 Écrivez une expression XPath (aussi simple que possible) qui rend chaque nœud de type `texte` dont la valeur est le nom d'un coureur ayant terminé deuxième dans Paris-Roubaix. Pour le document de la figure 2, la réponse consiste en `Fabian Cancellara` et `Sebastien Turgot`.

.../3

Question 2 Écrivez une expression XPath (aussi simple que possible) qui rend chaque nœud de type `texte` dont la valeur est le nom d'un coureur ayant atteint un podium à la fois en 2011 et en 2012. Pour le document de la figure 2, le seul coureur dans la réponse est `Fabian Cancellara`.

.../3

Question 3 Écrivez une expression XPath (aussi simple que possible) qui rend chaque nœud de type `attribute` dont la valeur est une année où un Italien a atteint le podium de Paris-Roubaix. Pour le document de la figure 2, la seule réponse est `annee="2012"`.

.../3

Question 4 Écrivez une expression XPath (aussi simple que possible) qui rend chaque nœud de type `texte` dont la valeur est le nom d'un coureur italien ayant déjà atteint un podium. La liste doit être sans doublons. Pour le document de la figure 2, la réponse consiste en `Filippo Pozzato`, `Alessandro Ballan` et `Vincenzo Nibali`.

.../3

Question 5 Écrivez un programme XSLT qui génère un document XML affichant les podiums de façon conviviale et en anglais, comme suit. La position des blancs et retours à la ligne n'a pas d'importance. Le programme ne peut pas contenir des `xsl:for-each` or `xsl:if`.

```
<?xml version="1.0" ?>
<CYCLING>
  <TourDesFlandres>
     <PODIUM year="2011"><WINNER nat="B">Nick Nuyens</WINNER>
                         <SECOND nat="F">Sylvain Chavanel</SECOND>
                         <THIRD nat="CH">Fabian Cancellara</THIRD>
     </PODIUM>
     <PODIUM year="2012"><WINNER nat="B">Tom Boonen</WINNER>
                         <SECOND nat="I">Filippo Pozzato</SECOND>
                         <THIRD nat="I">Alessandro Ballan</THIRD>
     </PODIUM>
  </TourDesFlandres>
  <ParisRoubaix>
     <PODIUM year="2011"><WINNER nat="B">Johan Vansummeren</WINNER>
                         <SECOND nat="CH">Fabian Cancellara</SECOND>
                         <THIRD nat="NL">Maarten Tjallingii</THIRD>
     </PODIUM>
     <PODIUM year="2012"><WINNER nat="B">Tom Boonen</WINNER>
                         <SECOND nat="F">Sebastien Turgot</SECOND>
                         <THIRD nat="I">Alessandro Ballan</THIRD>
     </PODIUM>
  </ParisRoubaix>
  <MilanSanRemo>
     <PODIUM year="2012"><WINNER nat="AUS">Simon Gerrans</WINNER>
                         <SECOND nat="CH">Fabian Cancellara</SECOND>
                         <THIRD nat="I">Vincenzo Nibali</THIRD>
     </PODIUM>
  </MilanSanRemo>
</CYCLING>
```

. . . /10

Bases de Données II, Charleroi, 18 janvier 2013

Cahier fermé. Durée : 2 heures

Nom et prénom
Année

Situez chaque terme dans le cursus et expliquez de façon succincte mais précise.

Question 1 MDL principle.

. . ./4

Question 2 $F_{k-1} \times F_{k-1}$ method.

. . ./4

Question 3 Model underfitting.

. . ./4

Question 4 Confidence-based pruning.

. . ./4

TID	a_1	a_2	a_3	a_4	a_5	b_1	b_2	b_3	b_4	b_5	c_1	c_2	c_3	c_4	c_5
1	**1**	**1**	**1**	**1**	**1**	0	0	0	0	0	0	0	0	0	0
2	**1**	**1**	**1**	**1**	**1**	0	0	0	0	0	0	0	0	0	0
3	**1**	**1**	**1**	**1**	**1**	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	**1**	**1**	**1**	**1**	**1**	0	0	0	0	0
5	0	0	0	0	0	**1**	**1**	**1**	**1**	**1**	0	0	0	0	0
6	0	0	0	0	0	**1**	**1**	**1**	**1**	**1**	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	**1**	**1**	**1**	**1**	**1**
8	0	0	0	0	0	0	0	0	0	0	**1**	**1**	**1**	**1**	**1**
9	0	0	0	0	0	0	0	0	0	0	**1**	**1**	**1**	**1**	**1**
10	0	0	0	0	0	0	0	0	0	0	**1**	**1**	**1**	**1**	**1**

FIGURE 1 –

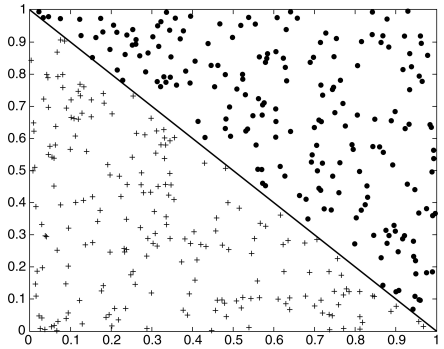


FIGURE 2 –

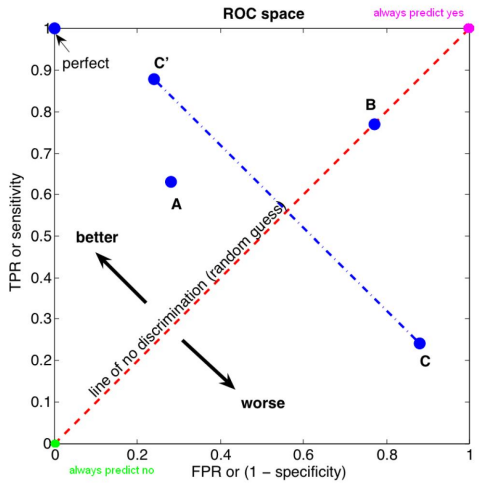
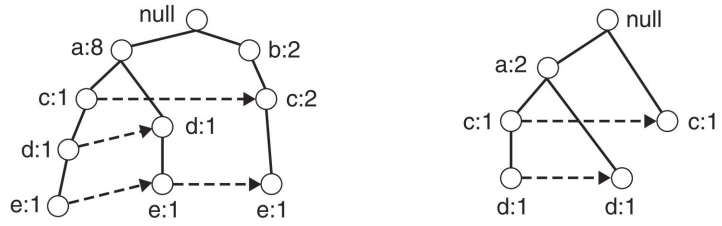


FIGURE 3 –



(a) Prefix paths ending in e (b) Conditional FP-tree for e

FIGURE 4 –

Question 5 Expliquez la figure 1 de façon détaillée. Dans quel contexte cette figure apparaît-elle ? Quel principe est illustré par cette figure ?

Question 6 Expliquez la figure 2 de façon détaillée. Dans quel contexte cette figure apparaît-elle ? Quel principe est illustré par cette figure ?

Question 7 Expliquez la figure 3 de façon détaillée. Dans quel contexte cette figure apparaît-elle ? Quel principe est illustré par cette figure ?

Question 8 Expliquez la figure 4 de façon détaillée. Dans quel contexte cette figure apparaît-elle ? Quel principe est illustré par cette figure ?

La question suivante est destinée exclusivement aux étudiants n'ayant pas presénté la partie XML de l'examen.

Question 9 Pour un problème de classification avec deux classes (yes et no), on veut comparer la performance des arbres de décision créés par deux algorithmes (C4.5 et ID3). Les matrices de confusion sur un ensemble de test sont montrées ci-après.

C4.5

		classe prédite	
		yes	no
classe observée	yes	18	2
	no	10	70

ID3

		classe prédite	
		yes	no
classe observée	yes	5	15
	no	5	75

1. En absence d'autres informations, expliquez une méthode pour tester si un des arbres est significativement meilleur que l'autre.
2. Supposons ensuite que l'on vous informe que le problème était de classer des courriels entrants en deux classes : si un courriel est classé "yes", il est considéré comme "spam" ; si un courriel est classé "no", il est considéré comme "non spam". Cette information supplémentaire peut-elle avoir un impact sur le choix du classificateur ? Détaillez votre réponse.

. . ./7

Réponse à la question 5.

. . ./7

Réponse à la question 6.

. . ./7

Réponse à la question 7.

. . ./7

Réponse à la question 8.

. . ./11

Réponse à la question 9 (pour les étudiants qui n'ont pas présenté la partie XML).