Exam - December 15th 2016
(2 hours)

Data Mining

Documents are allowed. Take care of the presentation.

1 Principal component regression

The data considered here are related to $n = 3883$ movies from the database IMDb. For each movie, we know the mean rating given by the users and the following $p = 8$ quantitative variables:

1) `budget` : budget (US dollars),

2) `duration` : duration (minutes),

3) `facenumber_in_poster` : number of faces in poster,

4) `gross` : gross in US (US dollars),

5) `num_critic_for_reviews` : number of reviews outside IMDb,

6) `num_user_for_reviews` : number of reviews by IMDb users,

7) `num_voted_users` : number of votes,

8) `title_year` : year of release.

As examples, here are the 15 best-graded movies (note the good rankings obtained by *The Lord of the Rings*):

```
##  1 The Shawshank Redemption (Score : 9.3)
##  2 The Godfather (Score : 9.2)
##  3 The Dark Knight (Score : 9)
##  4 The Godfather: Part II (Score : 9)
##  5 The Lord of the Rings: The Return of the King (Score : 8.9)
##  6 Schindler's List (Score : 8.9)
##  7 Pulp Fiction (Score : 8.9)
##  8 The Good, the Bad and the Ugly (Score : 8.9)
##  9 Inception (Score : 8.8)
## 10 The Lord of the Rings: The Fellowship of the Ring (Score : 8.8)
## 11 Fight Club (Score : 8.8)
## 12 Forrest Gump (Score : 8.8)
## 13 Star Wars: Episode V - The Empire Strikes Back (Score : 8.8)
## 14 The Lord of the Rings: The Two Towers (Score : 8.7)
## 15 The Matrix (Score : 8.7)
```

The goal is to provide a regression model based on principal components to predict the rating of a movie from the observations of these p variables.

We start with a brief exploration of the data. Figures 1 and 2 represent respectively the distribution of centred data and the distribution of centred and reduced data.

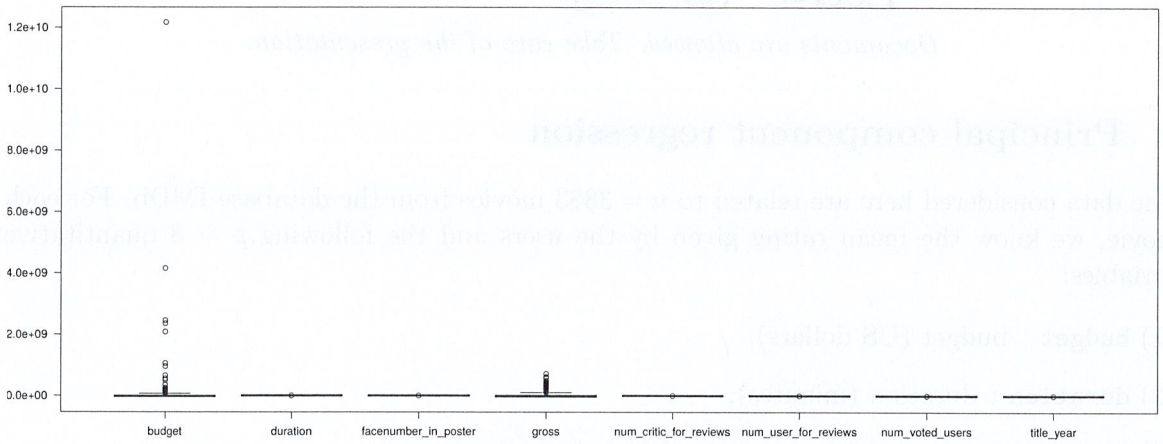


Figure 1: Box plots of centred data.

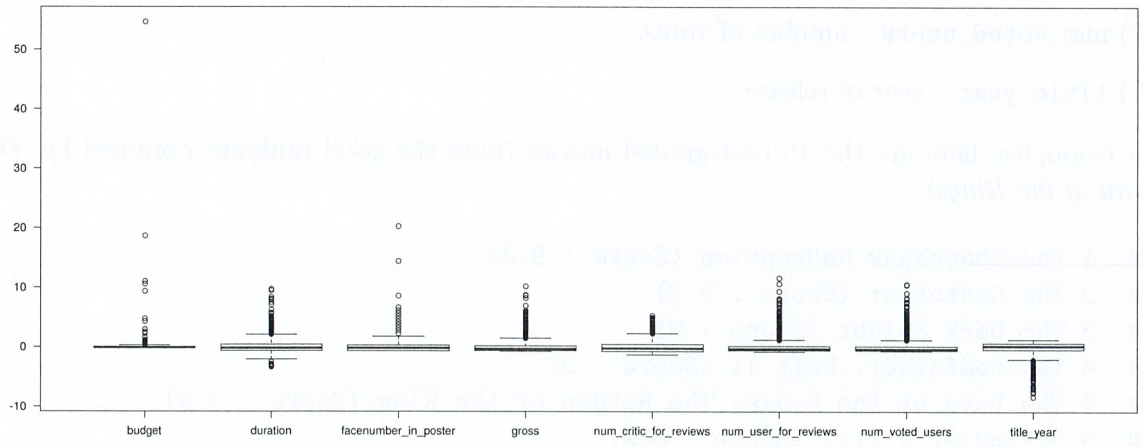


Figure 2: Box plots of centred and reduced data.

1.1 Does it seem reasonable to reduce the variances? What problem still remains after this normalization step?

Instead of initial data, we propose to handle the variables given through the function $t \mapsto \ln(1+t)$. Here are the variances of these transformed data:

```
##                budget                duration    facenumber_in_poster
##         2.3864221857              0.0335387936              0.4135809712
##                 gross num_critic_for_reviews    num_user_for_reviews
##         4.9515140481              0.7748816565              1.1716023304
##       num_voted_users                title_year
##         2.4059239275              0.0000251834
```

In the sequel, we consider the data set given by the centred and reduced versions of these transformed variables and we denote them by $x^1, \ldots, x^8$. Distributions of these data are represented in Figure 3.

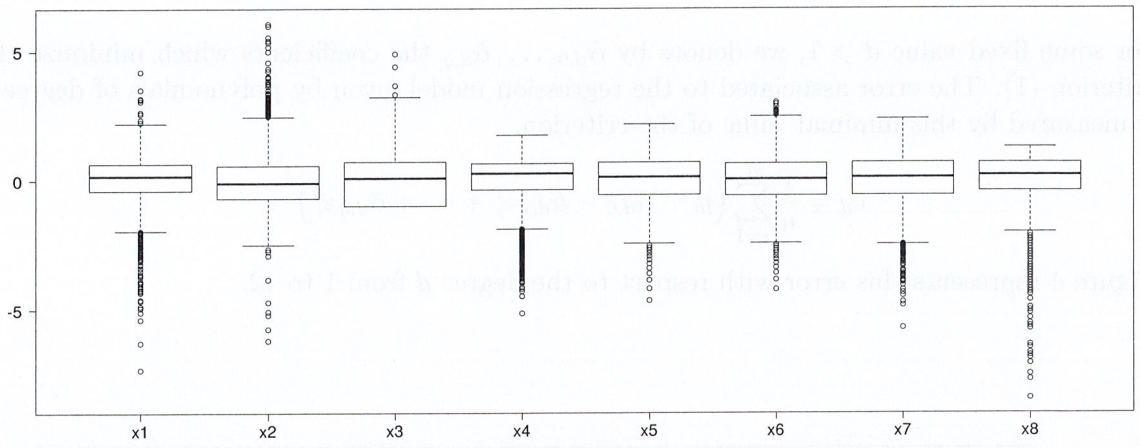


Figure 3: Box plots of centred and reduced transformed data.

1.2 Justify why we choose to reduce the variances of the transformed data. Compare the shapes of the distributions in Figures 2 and 3.

The proposed approach is based on linear regression. This means that if we denote by y the rating of a movie and by $z^1, \ldots, z^q$ the q predictor variables, we are looking for $q+1$ real numbers $\alpha_0, \ldots, \alpha_q$ to minimize the least squares criterion given by uniform weights,

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \alpha_0 - \alpha_1 z_i^1 + \cdots + \alpha_q z_i^q\right)^2 \tag{1}$$

where, for any $i \in \{1, \ldots, n\}$, y_i is the rating of the i-th movie and $z_i^1, \ldots, z_i^q$ are the observations for this movie. To get beyond a too simple model in the sequel, we consider polynomials of degree d according to the transformed variables. In other words, we handle the $q = pd$ real

variables given by

$$\begin{array}{llll} z^1 = x^1 & z^2 = (x^1)^2 & \dots & z^d = (x^1)^d \\ z^{d+1} = x^2 & z^{d+2} = (x^2)^2 & \dots & z^{2d} = (x^2)^d \\ \vdots & \vdots & & \vdots \\ z^{(p-1)d+1} = x^p & z^{(p-1)d+2} = (x^p)^2 & \dots & z^{pd} = (x^p)^d \end{array}$$

1.3 Explain the role of the coefficient α_0 in Model (1). What is the advantage in considering polynomials instead of a simple linear regression with respect to the initial variables?

For some fixed value $d \geqslant 1$, we denote by $\hat{\alpha}_{d,0}, \dots, \hat{\alpha}_{d,q}$ the coefficients which minimize the criterion (1). The error associated to the regression model given by polynomials of degree d is measured by this minimal value of the criterion,

$$e_d = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\alpha}_{d,0} - \hat{\alpha}_{d,1} z_i^1 + \dots + \hat{\alpha}_{d,q} z_i^q \right)^2 .$$

Figure 4 represents this error with respect to the degree d from 1 to 12.

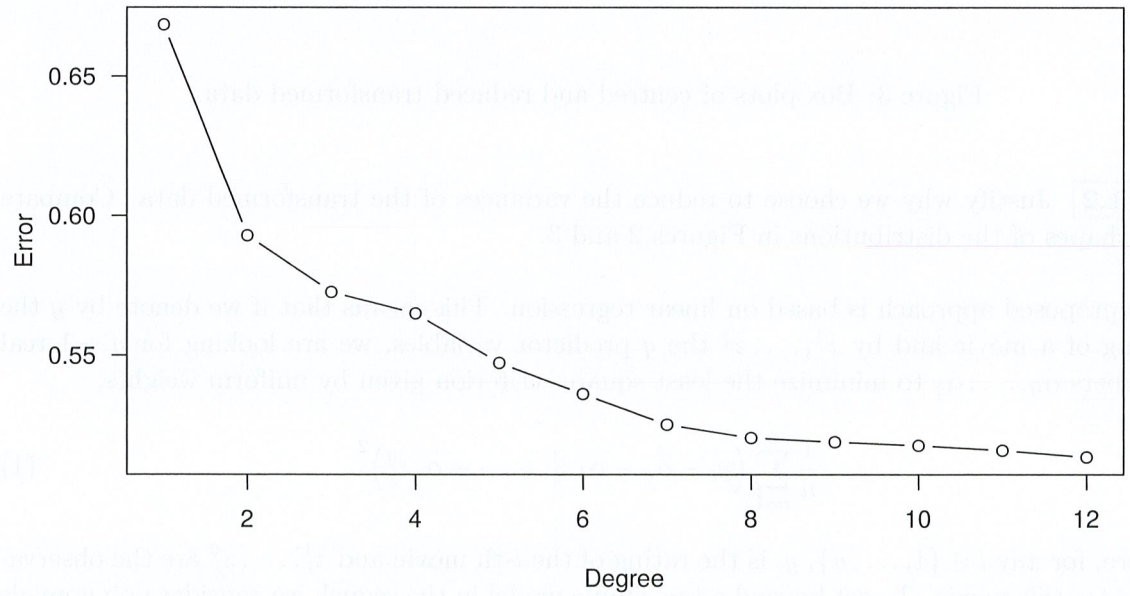


Figure 4: Error of regression models with respect to the degree of the polynomials.

1.4 How does behave the error e_d when d increases? Is it a wise choice to pick the value of d that makes this error minimal?

Instead of simply adjusting the degree d, we propose to compute the PCA for the covariance matrix associated to the $q = 96$ variables obtained by considering the polynomials of degree $d = 12$.

1.5 Defining appropriate matrices, explain how to compute the principal component matrix C for this PCA from the $q = 96$ variables $z^1, \ldots, z^{96}$.

Let $k \leqslant 96$, we can consider the linear regression of the rating according to the k first columns of matrix C. Figure 5 shows the error obtained with respect to k and the value previously achieved with $d = 12$ in Figure 4.

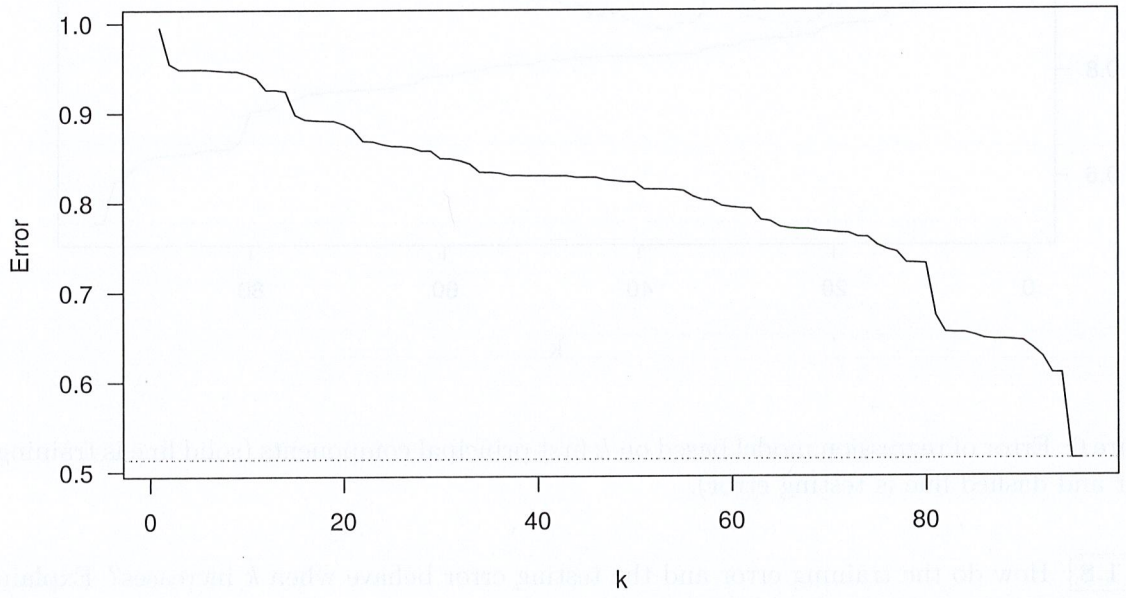


Figure 5: Error of regression model based on k first principal components (dotted line is the error for polynomials of degree 12).

1.6 How does the error behave when k increases? In particular, explain why this error is always larger than the one obtained with the regression model based on polynomials of degree 12 (dotted line in Figure 5).

1.7 Why are we only interested in the models given by the k first components and not by an arbitrary set of k principal components for our regression problem?

To pick a value of k, we use cross-validation. The data set is split into two parts, a training set to compute the regression model based on the k first principal components and a testing set to measure the error of the model. Results are given in Figure 6.

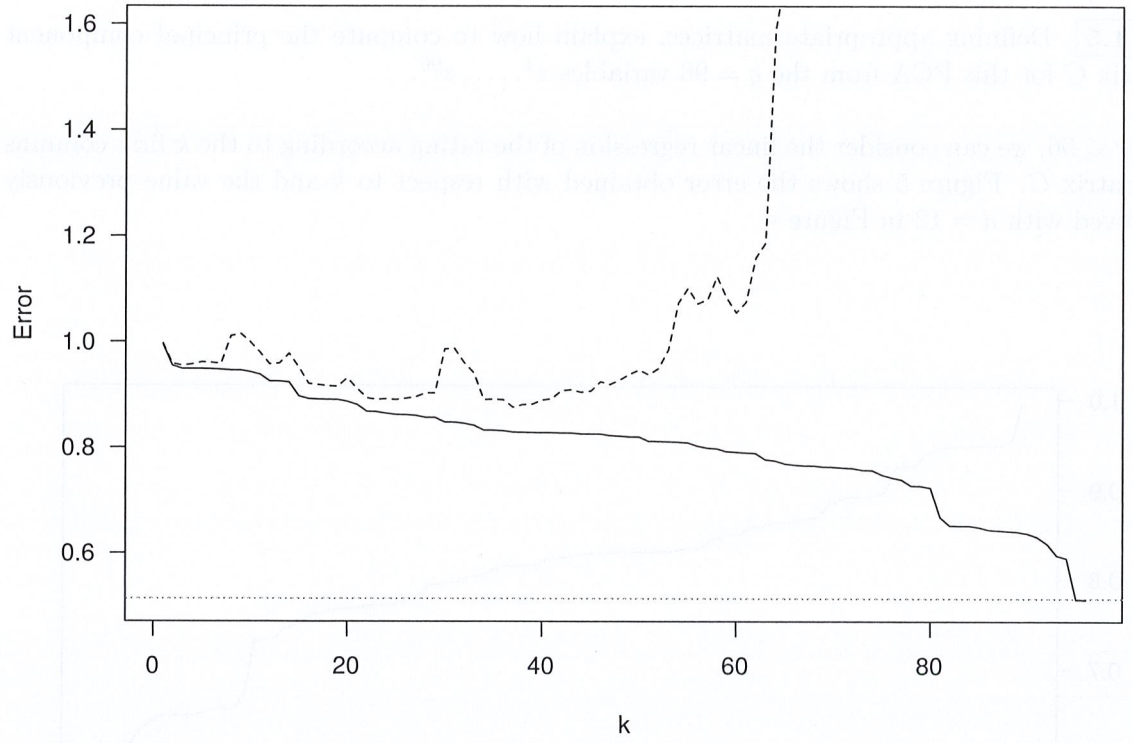


Figure 6: Error of regression model based on k first principal components (solid line is training error and dashed line is testing error).

1.8 How do the training error and the testing error behave when k increases? Explain the gap between these errors when k is large.

1.9 Roughly, what value is suggested by Figure 6 for k? Propose an automated method for such a choice.

To make the link with initial variables, we introduce the following quantities defined for any $i, k \in \{1, \ldots, q\}$,

$$r_i(k) = \sum_{j=1}^{k} \frac{\lambda_j^{3/2} \times |V_{ij}|}{\sqrt{\sigma^2(z^i)}}$$

where $\lambda_1, \ldots, \lambda_q$ are the eigenvalues from the PCA and the columns of V are the associated eigenvectors. From these values, we can also define, for any $\ell \in \{1, \ldots, p\}$,

$$s_\ell(k) = \sum_{i=12\times(\ell-1)+1}^{12\times\ell} r_i(k).$$

1.10 Noticing that $\lambda_j^{3/2} = \lambda_j \times \sqrt{\lambda_j}$, how to interpret the values $r_i(k)$? Deduce the meaning of $s_\ell(k)$.

Here are the obtained results for several values of k:

```
##                          k = 20 k = 30 k = 40 k = 50 k = 60 k = 70 k = 80
## budget                   0.0114 0.0143 0.0164 0.0182 0.0200 0.0218 0.0231
## duration                 0.0105 0.0132 0.0155 0.0178 0.0199 0.0216 0.0230
## facenumber_in_poster     0.0092 0.0123 0.0148 0.0173 0.0193 0.0210 0.0223
## gross                    0.0094 0.0125 0.0150 0.0174 0.0199 0.0218 0.0232
## num_critic_for_reviews   0.0109 0.0135 0.0158 0.0179 0.0197 0.0213 0.0227
## num_user_for_reviews     0.0099 0.0131 0.0155 0.0177 0.0198 0.0216 0.0230
## num_voted_users          0.0098 0.0126 0.0151 0.0176 0.0200 0.0220 0.0232
## title_year               0.0095 0.0126 0.0153 0.0175 0.0198 0.0217 0.0230
```

1.11 According to your answer at Question 9, what are the more important variables to predict the rating of a movie? What can we say about the number of faces in movie poster?

2 Your name says who you are

In this section, we consider the name list of the characters from J.R.R. Tolkien's works provided by *LOTR Project*. Each name is associated to a race and the database contains 783 characters distributed between 49 dwarves, 99 elves, 211 hobbits and 424 men. Here are few examples of names for these races:

`Dwarves` : Frór, Durin VI, Dís, Gróin, Thorin II, ...

`Elves` : Finwë, Galdor, Enel, Nerdanel the Wise, Egalmoth, ...

`Hobbits` : Gruffo Boffin, May Gamgee, Sméagol, Dodinas Brandybuck, ...

`Men` : Orchaldor, Ar-Pharazôn, Ingold, Tarcil, Lothíriel, ...

The goal is to predict the race of a character from his name. To this end, the letters of each name are converted to small letters, accented letters are replaced by their equivalents without accent and punctuation characters (space, hyphen, ...) are replaced by the capital letter 'X'. Thus, there are 27 allowed characters: small letters from 'a' to 'z' and capital letter 'X'. Here are some examples of this formatting:

Frór becomes `fror`,

Nerdanel the Wise becomes `nerdanelXtheXwise`,

Ar-Pharazôn becomes `arXpharazon`.

In a first time, we consider only the 27 variables given by the frequencies of each letter in the character's name. The example of *Frór* leads to the following vector:

```
##    a    b    c    d    e    f    g    h    i    j    k    l    m    n    o
## 0.00 0.00 0.00 0.00 0.00 0.25 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.25
##    p    q    r    s    t    u    v    w    x    y    z    X
## 0.00 0.00 0.50 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
```

2.1 Figure 7 represents the mean frequencies of each letter according to the race of the characters. Do these data seem relevant to set up a classification procedure? What are the races that could be hard to discriminate?

We consider an approach based on CART to deal with this classification problem. Let us first give the same weight to all the characters. The classification tree is given in Figure 9 and the misclassification rates according to the race are:

```
##    Dwarves      Elves    Hobbits        Men
## 1.00000000 1.00000000 0.09004739 0.04481132
```

2.2 Comment Figure 9 and these misclassification rates. Compute the global misclassification rate.

We now decide to assign weights to the characters as follows:

- $1/(4n_D)$ for the $n_D = 49$ dwarves,
- $1/(4n_E)$ for the $n_E = 99$ elves,
- $1/(4n_H)$ for the $n_H = 211$ hobbits,
- $1/(4n_M)$ for the $n_M = 424$ men.

2.3 Verify that these weights sum to 1. Explain why we consider these specific weights.

Figure 10 represents the classification tree obtained with these weights and the misclassification rates according to the race are:

```
##    Dwarves      Elves    Hobbits        Men
## 0.1428571 0.3939394 0.1042654 0.7500000
```

2.4 Comment Figure 10 and these misclassification rates. Compute the global misclassification rate and compare these results with what we obtained with uniform weights.

2.5 Apply the trees given by Figures 9 and 10 to your own name after having formatted it. Explain the steps.

We now consider the pairs of successive letters in the character's name. We have at our disposal the $27 \times 27 = 729$ variables associated to the frequencies of each pair which we can handle as a matrix. For example, name *Frór* leads to the pairs `fr`, `ro` and `or`:

$$\begin{array}{c} \\ a \\ \vdots \\ e \\ f \\ g \\ \vdots \\ n \\ o \\ p \\ q \\ r \\ s \\ \vdots \\ X \end{array} \begin{array}{c} \begin{array}{cccccccccc} a & \dots & n & o & p & q & r & s & \dots & X \end{array} \\ \begin{pmatrix} 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix} \end{array}$$

2.6 Figure 8 represents the mean frequencies of each pair of successive letters according to the race as intensity matrices (white for a null frequency and black for a frequency 1). Do these data bring more information than the previous ones? Does it allow us to think that the classification will be better?

As above, we use CART to perform the classification and here are the misclassification rates by race:

- with uniform weights:

```
##    Dwarves      Elves    Hobbits        Men
## 0.71428571 1.00000000 0.09478673 0.02358491
```

- with weights by race:

```
##   Dwarves     Elves   Hobbits       Men
## 0.2448980 0.1919192 0.1090047 0.5896226
```

2.7 Deduce the global mean errors in each case. Compare these results to what we have obtained with the frequencies of single letters.

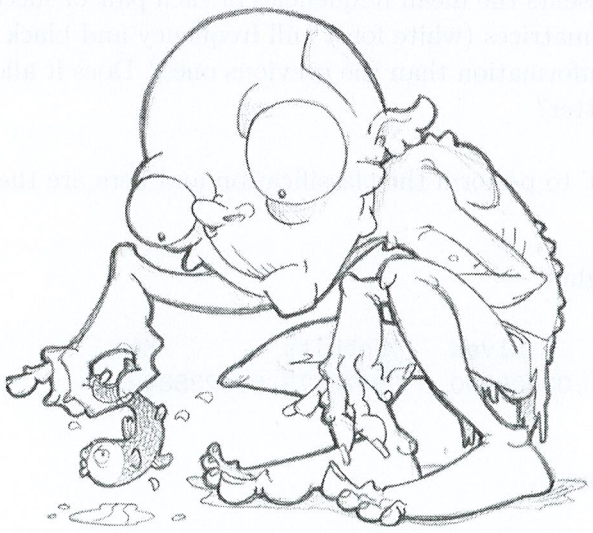
2.8 Apply the trees provided by Figures 11 and 12 to your own name and explain the steps.

2.9 What other method(s) would you propose to deal with this classification problem from these matrices of frequencies?

This procedure can be easily generalized to more than two consecutive letters even though there is no more simple visualisation for the data. For example, sequences of three successive letters in *Frór* are `fro` and `ror` with frequencies 1/2.

2.10 What are the difficulties with such an approach with more than 2 letters? Propose a way to bypass these difficulties.

2.11 What is the true name of *Gollum* and his race? Apply the trees from Figures 9, 10, 11 and 12 to verify the results. ☺



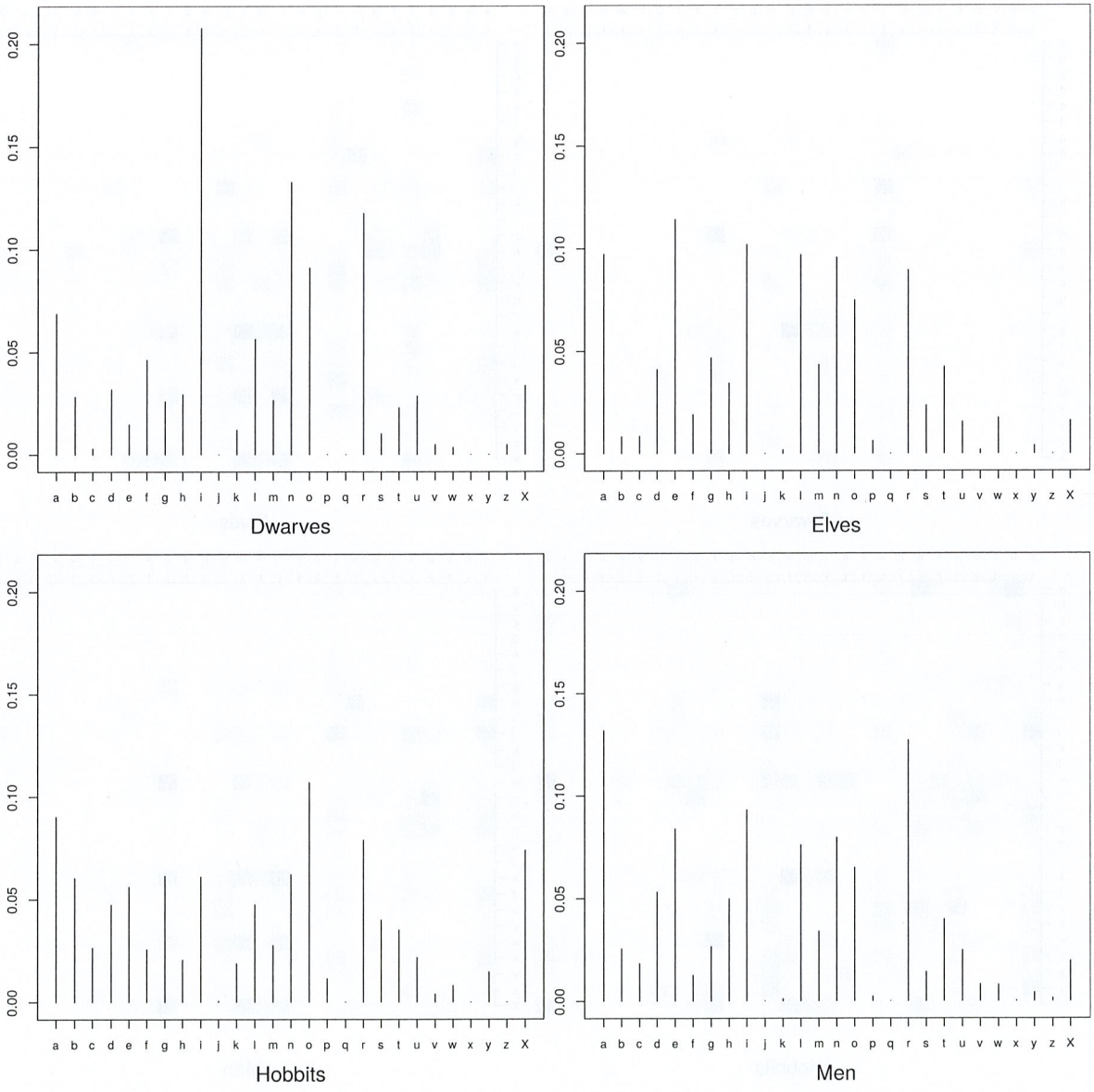


Figure 7: Mean frequencies of letters in character's name according to the race.

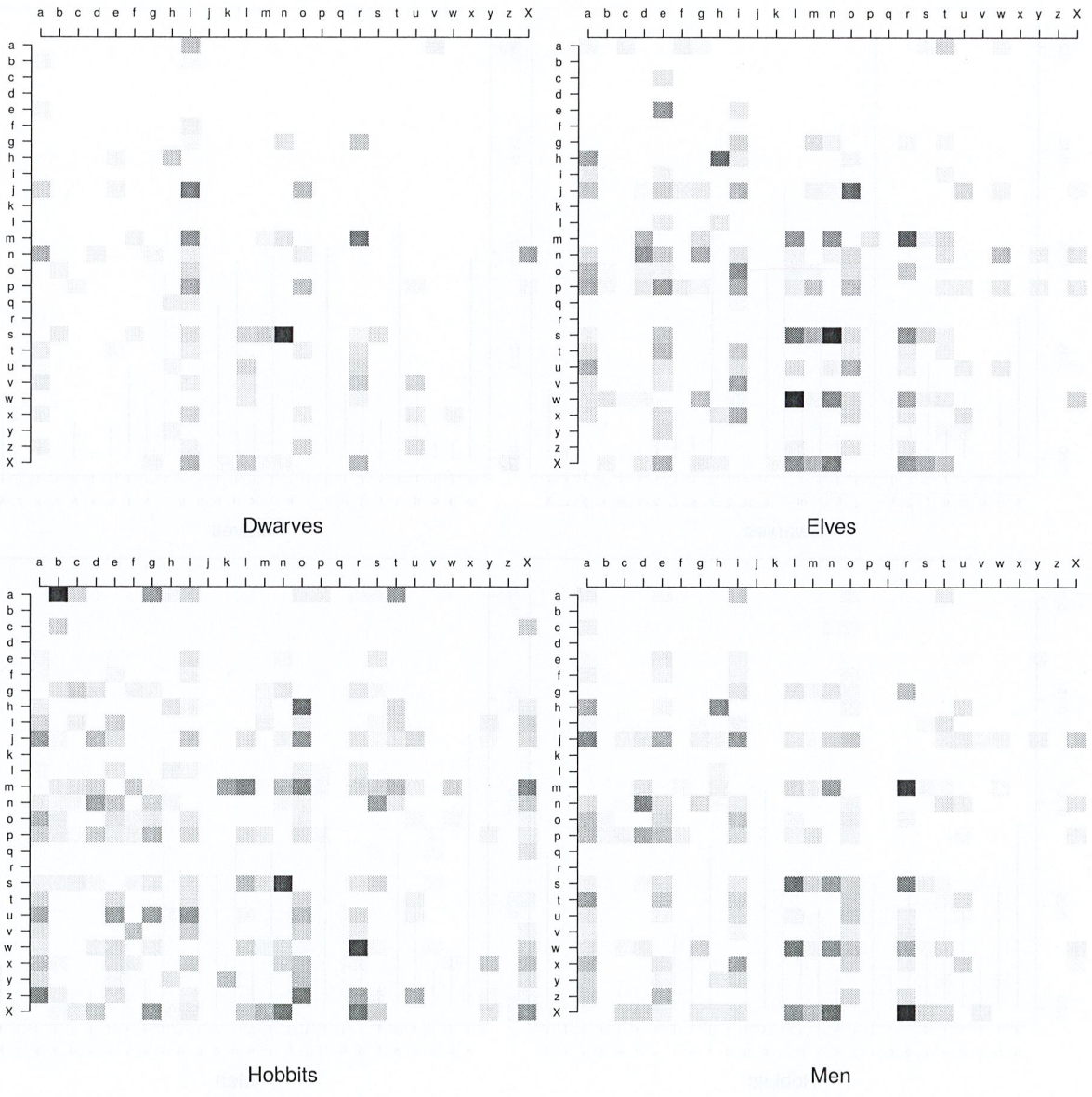


Figure 8: Mean frequencies from 0 (white) to 1 (black) of pairs of consecutive letters in character's name according to the race.

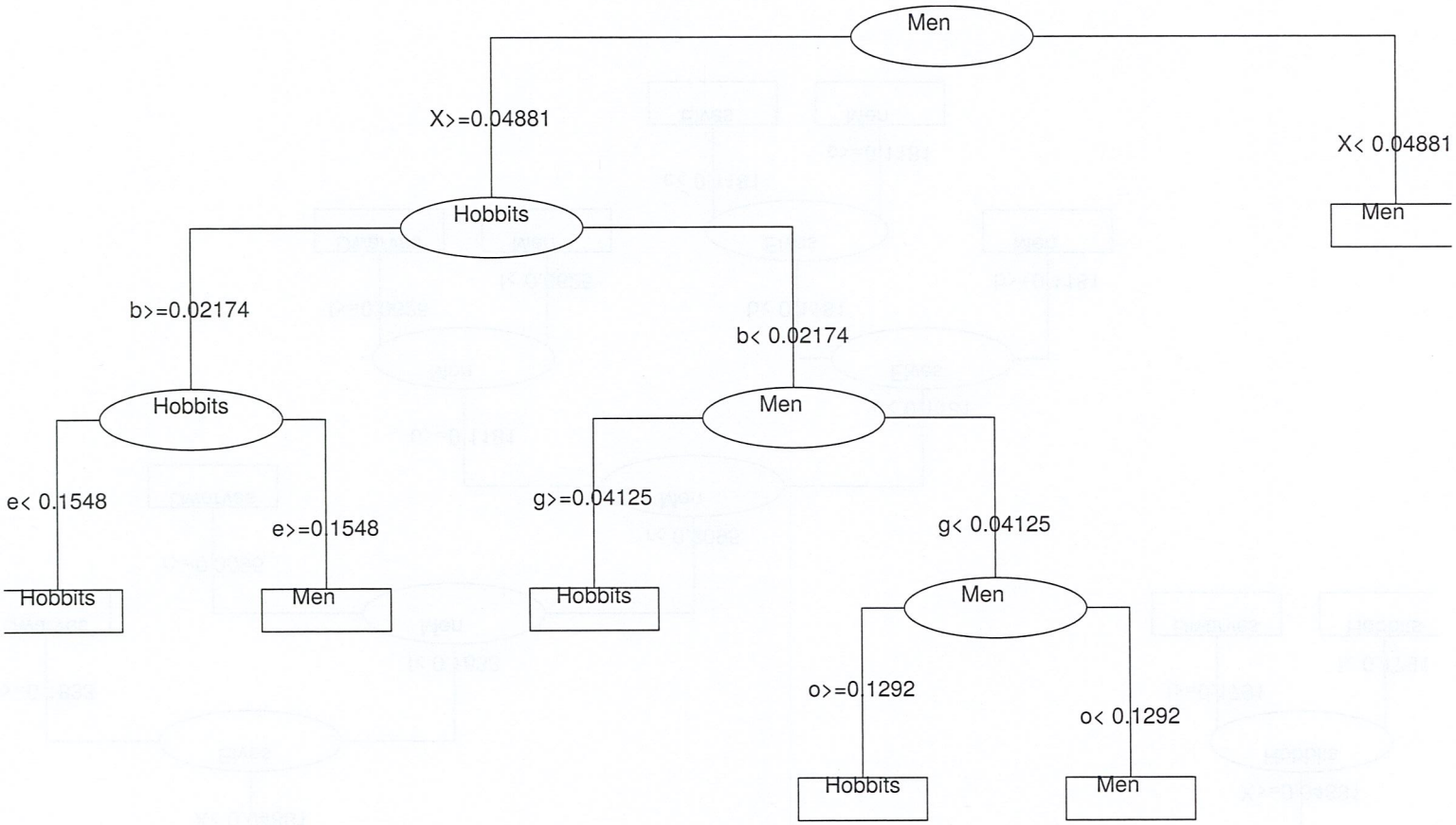


Figure 9: Classification tree based on single letter frequencies (uniform weights).

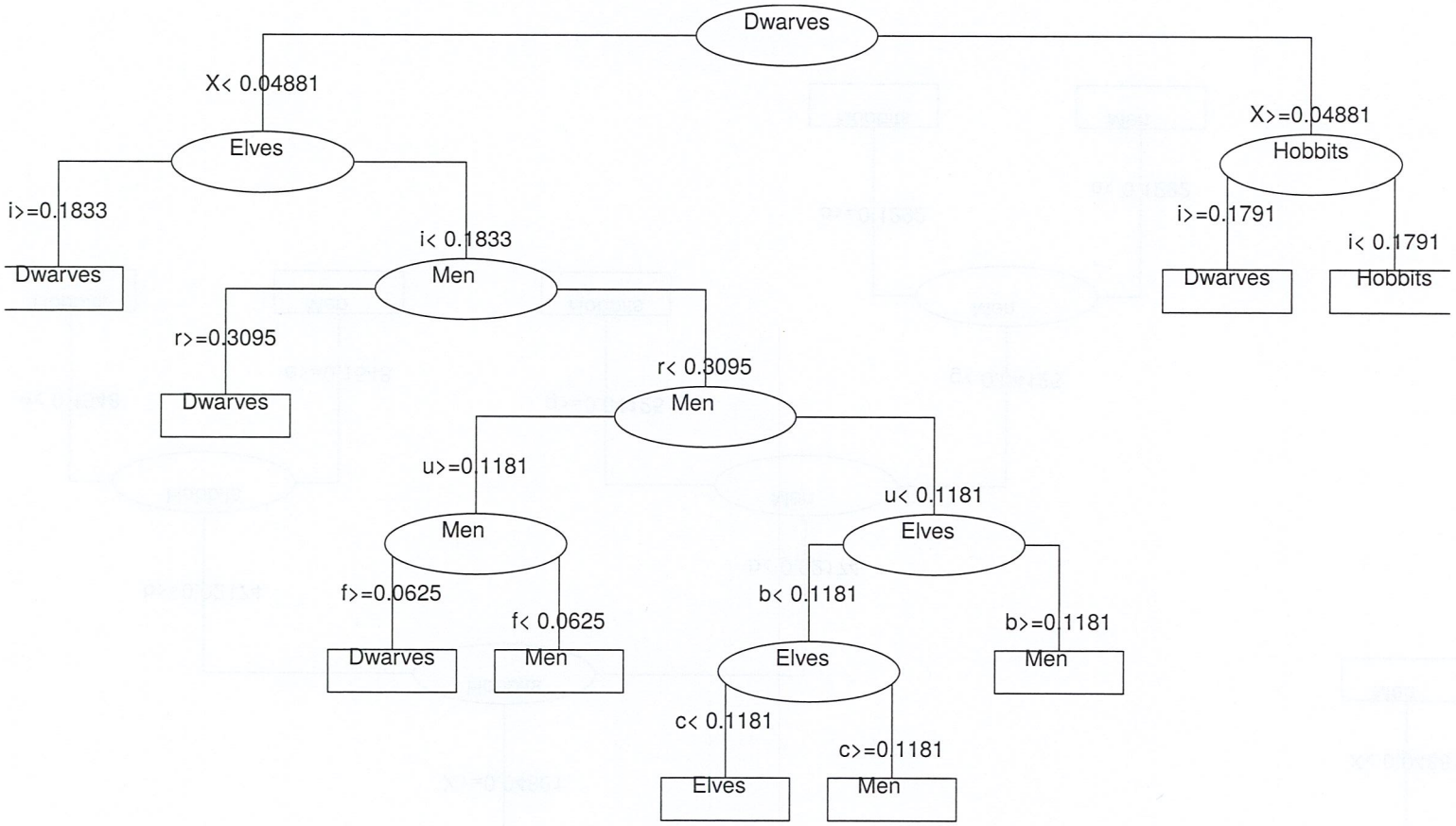


Figure 10: Classification tree based on single letter frequencies (weights by race).

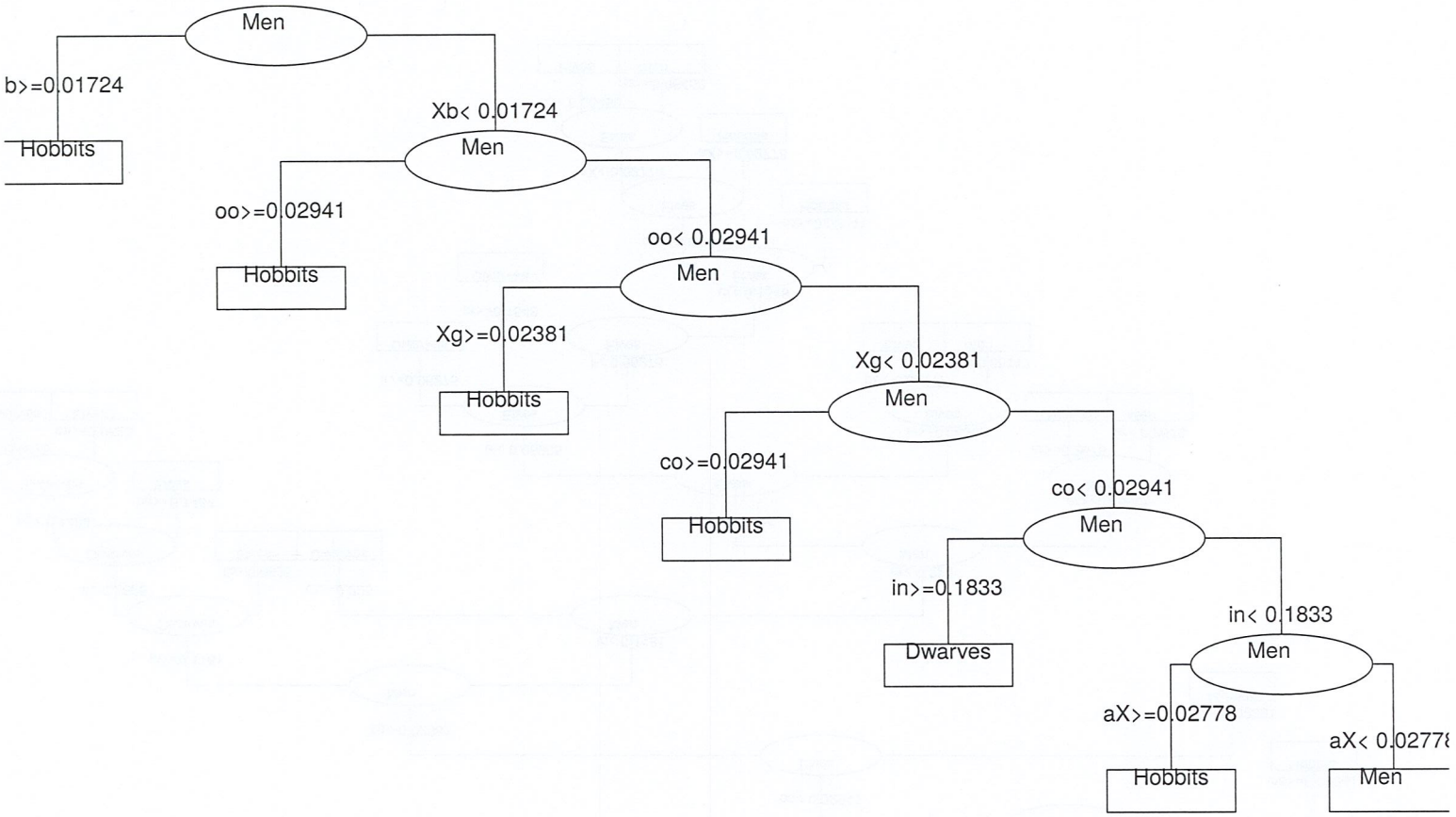


Figure 11: Classification tree based on frequencies of pairs of consecutive letters (uniform weights).

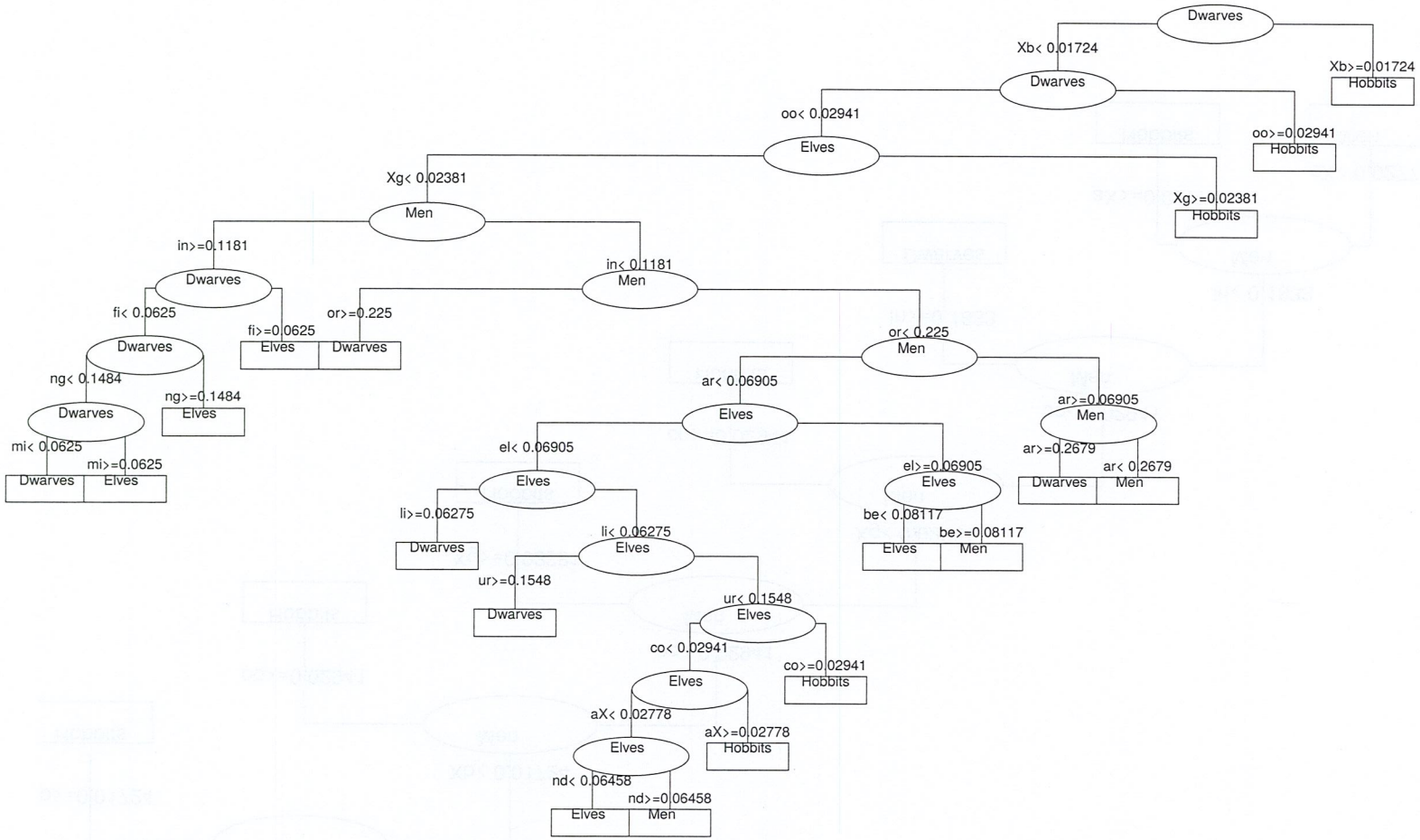


Figure 12: Classification tree based on frequencies of pairs of consecutive letters (weights by race).