Data Analytics Assignment (for HW7)

Predict the Ethnicity of Individuals from their Genes
==========

It is now possible to get the DNA sequence of an individual at a reasonable cost. An individual's genetic make-up determines a number
of charactersistics - eye colour, propensity for certain diseases, response to treatment and so on. In this problem, you are given a subset of genetic information for several individuals. For some of the individuals you are also told their ethinicity. Your task is to figure out the ethnicity of the other individuals.

The information provided is as follows:

1. For each individual the presence (1) or absence (0) of a genetic variation at a particular position on chromosome 6 is provided. In some cases, information for an individual at a particular position is not available and this represented as ? (missing).

2. Information is provided for approximately 204000 positions. These are your features.

3. The training set has data for 139 individuals along with their ethnicity.

4. The test (prediction) set has data for 11 individuals. You have to predict the ethnicity for these individuals and enter your answers via HW7.

Data Sets

The training set is available here: genestrain.tab.zip (6.2 Mb)

The test set is available here: genesblind.tab.zip (1.2 Mb)

File Format

(Note: Data sets are .tab files in the tab-separated format that can be read into Orange):

Both the training and test data files have a header line which is a tab-separated line of column/feature names: For example '6_10000005' indicates that the column describes the presence or absence of variations at position 10000005 on chromosome #6.

Entries in the second header line indicate the type of column (in this case all features are 'discrete').

Entries in the third header line indicate the nature of each column:
A ' ' for most columns that contain a feature, and 'class' for the first column as it contains the actual class labels (i.e., ethnicities of the individuals in each row).

These header lines are followed by lines containing feature values (0, 1, or ?) for each genetic feature of an individual.

In the training set file the first column, which denotes the class label, is a three-letter code with one of the following values:

o CEU is Northern and Western European
o GIH is Gujarati Indian from Houston
o JPT is Japanese in Tokyo
o ASW is Americans of African Ancestry
o YRI is Yoruba in Ibadan, Nigera

In the test file the ethnicity column also exists but is blank.

==========

For the purposes of your HW answer *alone*, each three letter code is to be marked with a NUMERIC VALUE as indicated in the table below:

o CEU is Northern and Western European - **0**
o GIH is Gujarati Indian from Houston - **1**
o JPT is Japanese in Tokyo - **2**
o ASW is Americans of African Ancestry - **3**
o YRI is Yoruba in Ibadan, Nigera - **4**

YOU MUST USE THE ABOVE NUMERIC VALUES TO ENCODE YOUR ANSWER. Note: This numeric value has no presence in the test or training data.

Task: For each of the individuals in the test file, predict their ethnicity as CEU, GIH, JPT, ASW or YRI and enter your answers in HW7 in exactly the order that the 11 individuals appear in the test file. So, for ***example***, if your prediction is **CEU, GIH, JPT, ASW, YRI CEU, GIH, JPT, ASW, YRI, CEU**, you should enter your answer as **0 1 2 3 4 0 1 2 3 4 0** (i.e. numbers separated by a space - no commas, tabs or anything else, just as space between single digit numbers).

Created Thu 2 May 2013 2:19 PM CEST (UTC +0200)