Assignment 5: Logistic Regression

Learning Outcomes

- Prepare a dataset for logistic regression
- Preform a logistic regression

Motivation

Download this dataset from the UCI Machine Learning Repo. This dataset contains 5.7 million records. This is the same place where we found the Iris dataset.

In the German state of North Rhine-Westphalia, researchers were building a database of cancer patients at the Institute for Medical Biostatistics. Each patient had the following attributes.

1. Phonetic equality of first name and family name, equality of date of birth.
2. Phonetic equality of first name, equality of day of birth.
3. Phonetic equality of first name, equality of month of birth.
4. Phonetic equality of first name, equality of year of birth.
5. Equality of complete date of birth.
6. Phonetic equality of family name, equality of sex.

Here's the problem. Researchers were entering this data by hand and they were making mistakes along the way. After time passed, the researchers then decided to see if they could detect if duplicate people where in the system. They would take two people at random and record 11 attributes, such as did their last name match, did the postal codes match, etc. Finally, in the 12th attribute column, you'll see "TRUE" if the two people were a match and "FALSE" if they weren't.

1. id_1: Internal identifier of first record.
2. id_2: Internal identifier of second record.
3. cmp_fname_c1: agreement of first name, first component
4. cmp_fname_c2: agreement of first name, second component
5. cmp_lname_c1: agreement of family name, first component
6. cmp_lname_c2: agreement of family name, second component
7. cmp_sex: agreement sex
8. cmp_bd: agreement of date of birth, day component
9. cmp_bm: agreement of date of birth, month component
10. cmp_by: agreement of date of birth, year component
11. cmp_plz: agreement of postal code
12. is_match: matching status (TRUE for matches, FALSE for non-matches)

Your assignment.

Repeat these steps 3 times, each with a different RFormula:

1. Download and unzip the data. You'll notice that it's divided into 10 zip files.
2. Select one zip file as your training set and work through the steps in Chapter 26 to perform a logistic regression on the data to determine if the two people are a match or not.
3. Determine the accuracy of your training results and report it. Also report which zip container you used.
4. Using the same model that you trained, select a different zip file and run your tests again using the classifier that built previously.
5. Report the accuracy of your test results and the zip container that you used.

RFormulas

For each RFormula, you will be classifing the 12th column: is_match.

1. For the first RFormula, use classifier where you use everything except the target.
2. For the third RFormula, combine each part of the birthday into one part.
3. For the second RFormula, combine the two components of the German family name and combine the two components of the Germinal first name.

Turn it in

Prepare a document of each of your commands and turn it in.