ASSESSMENT: BREAST CANCER TYPE CLASSIFICATION

Iris Manola
manola.irida@gmail.com
18 January 2019

CONTENTS

This presentation includes:

1. An introduction to the subject and the tools used.
2. Having a look at the data
3. Find the most influential features
4. Select the set of Machine Learning (ML) models
5. Apply and test the accuracy of those models
6. Decide on the most appropriate one

Detailed line-by-line commenting on the reasoning of each step of this presentation can be found in the Jupyter notebook file.

1. SUBJECT AND TOOLS

- Data from 157 patients diagnosed with three different types of breast cancer, including their DNA measurements.
- We are interested in the type of breast cancer, and whether it can be predicted from other variables in the dataset.

- The language used in this assessment is Python 3.5.2 on Jupyter notebook.

2. THE DATASET

- Loading the 'BreastCancerAll.reduced.using.csf.missing.arff' data in a pandas dataframe for a quick look of the first 5 rows and 5 columns.

	302_chrom1_reg44816452-46047207_probnorm	1195_chrom2_reg15573647-20346474_probgain	1441_chrom2_reg126478953-126745976_probloss	1471_chrom2_reg131945577-132042781_probgain	1673_chrom3_reg12735694-14873985_probloss	1966_chrom3_r1286647
0	0.979	0.005	0.009	0.006	0.008	
1	0.979	0.006	1.000	0.000	0.009	
2	0.978	0.006	0.007	0.007	0.007	
3	0.000	0.651	0.069	0.001	1.000	
4	0.978	0.009	0.017	0.003	1.000	

5 rows × 60 columns

- There are 59 DNA features and 1 cancer feature of 157 sample observations. This is in total 9420 features in the dataset.
- The cancer feature is found under the variable name 'class' and includes 3 different cancer types with names:
  1. 'HER2+'
  2. 'HR+'
  3. 'TN'

2. THE DATASET

Are the data clean?

- A df.isnull().sum() counts that there are two features that have one missing value each.
- Those two observations should be omitted from our analysis to avoid errors.
- The cancer labels are categorical data. They need to be transformed in integers with a label encoder so that pythons models can understand them. Now they are called 0, 1 and 2 instead of HER2+, HR+ and TN.
- Now the dataset is reduced to 155 observations and 9300 features and is ready to be processed.
- Additionally, the frequency distribution table below is showing that the 3 cancer types are of almost equal chances of occurrence in the 155 samples:

Cancer Type	Frequency of occurrence
HER2+	49
HR+	53
TN	53

3. THE MOST INFLUENTIAL FEATURES

- The most influential/informative features among the 59 DNA measurements are those that can explain most of the variance in the data.
- The SelectKBest command of scikit-learn library can easily give the list of the top 5 most informative features and their F-scores (the ration between the explained and unexplained variance):

5 Most important predictors	
Feature names	**F-Scores**
9594_chrom17_reg34940330-35067992_probnorm	44.06747
9602_chrom17_reg35286565-35336158_probnorm	41.505915
9598_chrom17_reg35076296-35282086_probnorm	35.178244
9725_chrom17_reg41458670-41494331_probloss	25.97394
8917_chrom16_reg68838368-68874167_probloss	17.628219

4. Select the set of models

- We should now select a multiclass supervised classification ML models (algorithms) that would allow us to make trustworthy predictions on the cancer class if such new DNA samples are given.

- Since the dataset consists of only 9300 features we will not be constrained by memory issues and therefore we can decide freely between fast and light ML techniques or heavier and more memory consuming.

- Since we start with the already reduced dataset we do not need to apply any feature reduction technique. If we had used the 'BreastCancerAll.missing.arff' dataset then such as action would have been required because then the number of the different DNA features would have been much larger than the total number of sample observations.

- The multiclass supervised classification ML techniques that can fitting in our case and will be tested here are the following:
    1. K Nearest Neighbour
    2. Naive Bayes
    3. Decision Tree
    4. Random Forest

5. Apply and test the models

- The data will be divided in the X independent variables (the 59 DNA categories) and the Y dependent variable (the 3 cancer classes), which is the one that we would like to be able to predict given the set of X variables.

- The data will be separated in two, random, non-overlapping sets, which are the training set and the test set.

- The training set will define with the help of the ML algorithms the relations between the data and learn from them. The test set will perform an unbiased evaluation of the fit of each model on the training data. In this case we divide the data in a training set that contains the 80% of the data and the test set that contains the rest of 20%. A sensitivity test (repeating the process for slightly different test sets of 20, 22 and 25% showed that the outcome does not change considerably in the current analysis).

- Because the DNA features are of different magnitudes, in order to be compared they should be transformed to the same level of magnitude. Here we used the StandardScaler module for this.

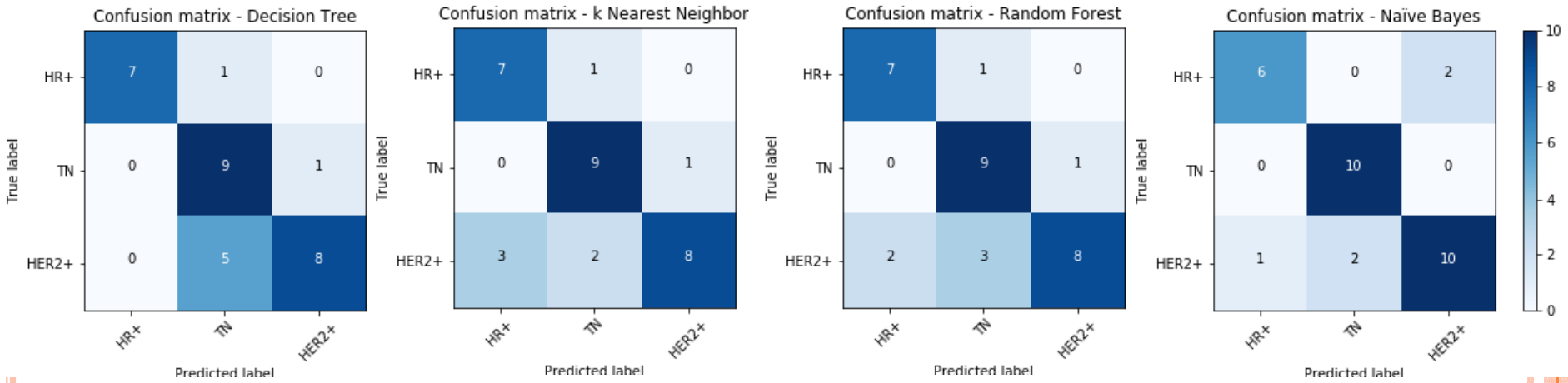
5. Apply and test the models

There are many tests to assess the accuracy of a ML model. A highly explanatory one is the **Confusion Matrix** algorithm that predicts how many times each model makes a right or a wrong prediction by giving details for the precision and the recall of the model as:

Precision = True Positive / (True Positive + False Positive)

Recall = True Positive / (True Positive + False Negative)

Let's plot the confusion matrix for each of the four models:



So, for example, if we make 31 predictions using the Decision Tree model we will have 7 times correct prediction for a HR+ cancer, 1 time it will be predicted as HR+ but it will truly be TN and 0 times it will be predicted as HR+ but it will truly be a HER2+. Respectively the TN will be predicted as HR2+ zero times, 9 times will be predicted correctly and 1 time will be predicted as TN but it will be a HER2+.

(keep in mind that with a different random selection of the train and test data these numbers will change slightly, but they give a good estimation)

5. Apply and test the models

- In the table below a more general and easier way to read a confusion matrix is shown were the outcome is normalized for simplicity.
- The F1-Score a is the harmonic average of precision and recall, so it also takes into account the false positives and false negatives but can give the overall estimation of the precision of each tested model.

Classification report k Neighbors			
type	precision	recall	f1-score
HER2+	0.7	0.88	0.78
HR+	0.75	0.9	0.82
TN	0.89	0.62	0.73
weighted avg	0.8	0.77	0.77

Classification report Decision Tree			
type	precision	recall	f1-score
HER2+	1	0.88	0.93
HR+	0.6	0.9	0.72
TN	0.89	0.62	0.73
weighted avg	0.82	0.77	0.77

Classification report Naive Bayes			
type	precision	recall	f1-score
HER2+	0.86	0.75	0.8
HR+	0.83	1	0.91
TN	0.83	0.77	0.8
weighted avg	0.84	0.84	**0.84**

Classification report Random Forest			
type	precision	recall	f1-score
HER2+	0.78	0.88	0.82
HR+	0.69	0.9	0.78
TN	0.89	0.62	0.73
weighted avg	0.8	0.77	0.77

6. THE MOST PRECISE MODEL

According to the F1-Score model that performs the best is the **Naïve Bayes** model with precision of 84% ! This is in general a pretty high percentage of accuracy and therefore the Naïve Bayes is assumed a quite trustworthy and well fitted model to predict the type of Breast Cancer if the set of 59 DNA features is given.

P.S. We should keep in mind that the Naïve Bayes model assumes that the predictors are independent of each other given the class. It is important that this assumption is verified, which is randomly assumed as true for the current analysis.

Thank you for your time!

I look forward to your evaluation.