Machine Learning: Homework 6

Laurent Hayez

November 13, 2016

Exercise 1. *Let's manually do a 5-fold cross-validation to get an error estimation of predicting the gender for the dataset below. You need multiple arff files. Let WEKA build the decision trees for each fold with the training data using the J48 algorithm, then compute the error of the test data. Now use WEKA for the whole 5-fold cross-validation and compare the results (accuracy and confusion matrix).*

Solution. The arff files were constructed as follows: the i-th test set consisted of the $2 \cdot (i-1)$-th and $(2 \cdot (i-1)+1)$-th samples and the i-th training test set consisted of all the samples except the ones selected in the i-th test set.

With this procedure, we created 10 arff files, 5 for training and 5 for testing. Every prediction was done using a decision tree constructed with the J48 algorithm.

1. The results of the first cross validation can be found in the file `results1.txt`. The confusion matrix is
$$C_1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}.$$
So the error is $err_1 = \frac{1}{2}$.
2. The results of the second cross validation can be found in the file `results2.txt`. The confusion matrix is
$$C_2 = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}.$$
So the error is $err_2 = \frac{1}{2}$.
3. The results of the third cross validation can be found in the file `results3.txt`.

The confusion matrix is

$$C_3 = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}.$$

So the error is $err_3 = \frac{1}{2}$.

4. The results of the fourth cross validation can be found in the file `results4.txt`. The confusion matrix is

$$C_4 = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}.$$

So the error is $err_4 = \frac{1}{2}$.

5. The results of the fifth cross validation can be found in the file `results5.txt`. The confusion matrix is

$$C_5 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

So the error is $err_5 = 0$.

The resulting confusion matrix is

$$C = C_1 + C_2 + C_3 + C_4 + C_5 = \begin{pmatrix} 4 & 2 \\ 2 & 2 \end{pmatrix}$$

and the resulting error is

$$\widehat{err} = \frac{1}{k}\sum_{i=1}^{k} err_i = \frac{1}{5}\left(4 \cdot \frac{1}{2}\right) = \frac{2}{5}.$$

We can compare this result with WEKA's auto 5-fold cross validation (which can be found in `results_whole_set.txt`). With WEKA's auto 5-fold cross validation, we obtain the following confusion matrix

$$\begin{pmatrix} 4 & 2 \\ 2 & 2 \end{pmatrix}$$

which means that 6 instances were correctly classified and 4 were not correctly classified. Also note that this matrix is equal to C. From this we obtain that the accuracy is $\frac{4}{6+4} = \frac{2}{5}$ which is the same that we found. □

Exercise 2. *A dataset about 76 booklovers shows some information (gender, age, number of books, likes Dan Brown, ..., bought your bestseller). You want to send publicity to other customers who might be interested in your book. Based on the known readers* *you build decision trees using the k-fold cross validation. Which of these trees do you use to determine which person to send advertisement to?*

What's the size of the first, sixth, seventh, and tenth test set if you are using 10-fold cross-validation?

Solution. Assuming that the new book is similar to the bestseller (in the sense that it is the same genre, about the same length and so on), I would make predictions based on if people liked the bestseller. I would use the trees which minimize the error over the training set and the test set. Using these trees would mean that the attribute selected based on the training set are good, and if they also work well for the test set, then we are sure that by using these attributes, we will have a good chance to reach people who will like the book.

For the test set sizes, the 6 first test sets will be of size 8 and the 4 remaining test sets will be of size 7. Indeed we have 78 observations and $8 \cdot 6 + 7 \cdot 4 = 48 + 28 = 76$. □