Coursework Assignement 2 : Correlation, Selection Features and Confusion Matrix

Guillaume Lemaitre - ID student : 09295005
Heriot-Watt University, Universitat de Gerona, Universite de Bourgogne
g.lemaitre58@gmail.com

I. Processing

We used *C++* to perform all process below.

A. Pre-treatment

The first task was to prepare the dataset. We computed the dataset like in the first assessment. However, we had to add one modification. The values in class field should be values between 0 and 9 instead of 0 and 1. That is why, we create a script to map values. The table I represents the mapping used.

B. Randomisation

The second task was to randomise the instances of the dataset. In this part, we created a vector containing the number of instances of the dataset. For instance, a dataset containing five instances, the vector which we want perform will be $v = [0, 1, 2, 3, 4]$. Then, we used *STL* library and more precisely *random_shuffle* function to randomise the values contained in the vector. For instance, a solution of randomisation using this function will be $v_{rand} = [3, 0, 2, 5, 1]$. Then, we had to permute each instance of the dataset following the order of the randomised vector.

C. Correlation

1) Theory: The third task was to create a function allowing to compute the correlation between two fields. To implement this function, we used the following formula:

$$r_{xy} = \frac{1}{N} \sum_{i=0}^{N-1} \frac{(x_s(i) - \overline{x})(y_s(i) - \overline{y})}{s_x s_y} \tag{1}$$

where N is the number of instances,x_s and y_s are respectively one sample of the field x and y, $\overline{x}$ and $\overline{y}$ are respectively the mean of the field x and y and s_x and s_y are respectively the standard deviation of the field x and y. The mean of a field can be formalised:

$$\overline{x} = \frac{1}{N} \sum_{i=0}^{N-1} x_s(i) \tag{2}$$

where N is the number of instances and x_s one sample of the field x. The standard deviation of a field can be formalised:

$$s_x = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} (x_s(i) - \overline{x})} \tag{3}$$

where N is the number of instances, x_s one sample of the field x and $\overline{x}$ is the mean of the field x.

2) Correlation between each non-class field and class field: To perform the correlation between each non-class field and class field, we applied the equation (1). We introduce now the fact that we took the absolute value of the correlation to sort fields.

3) Reduction of dataset using correlation between each non-class field: To perform this reduction, we computed the correlation between each non-class field using equation (1). We removed one of both fields where the value of the correlation is maximum. We still carried this manipulation to have only five non-class fields.

II. Reduction of dataset using correlation between non-class fields and class field

A. Top 5 non-class fields

To perform this ranking, we computed the correlation and sort fields with the absolute value of the correlation of each field. For the top 5, we selected only five fields with the highest values of absolute value of the correlation. We show the results in the table IV. We created a reduced dataset with only top 5 fields and the class field to perform the Naive Bayes classifier. The accuracy of the classifier is 42.11%. The confusion matrix is presented in the table II.

Interval	$0-0.1$	$0.1-0.2$	$0.2-0.3$	$0.3-0.4$	$0.4-0.5$	$0.5-0.6$	$0.6-0.7$	$0.7-0.8$	$0.8-0.9$	$0.9-1$
Mapping	0	1	2	3	4	5	6	7	8	9

Table I
TABLE REPRESENTING THE MAPPING REALISED TO TRANSFORM THE CLASS FIELD

Class	1	2	3	4	5	6	7	8	9	10
1	108	22	7	0	0	0	1	0	0	0
2	26	25	17	4	0	7	3	0	0	0
3	5	24	21	5	2	7	3	0	0	3
4	2	4	14	1	1	6	1	0	1	6
5	0	3	4	1	1	6	2	1	0	2
6	0	1	2	3	0	3	4	1	0	0
7	1	0	0	0	0	3	1	0	1	4
8	0	0	1	0	0	3	2	0	0	1
9	0	0	0	0	0	1	2	1	0	3
10	0	1	2	0	0	1	2	0	1	8

Table II
CONFUSION MATRIX OF THE TOP 5 FIELDS

Class	1	2	3	4	5	6	7	8	9	10
1	108	22	7	0	0	0	1	0	0	0
2	22	25	18	5	3	7	1	0	1	0
3	4	23	22	6	2	6	5	0	0	2
4	1	3	13	3	2	6	2	0	2	4
5	0	1	5	0	4	3	6	0	0	1
6	0	0	3	3	2	4	1	1	0	0
7	0	1	0	1	0	3	2	0	0	3
8	0	0	0	0	2	2	3	0	0	0
9	0	0	0	0	1	0	3	1	0	2
10	0	0	1	1	0	3	0	0	2	8

Table III
CONFUSION MATRIX OF THE TOP 10 FIELDS

B. Top 10 non-class fields

To perform this ranking, we computed the correlation and sort fields with the absolute value of the correlation of each field. For the top 10, we selected only ten fields with the highest values of absolute value of the correlation. We show the results in the table IV. We created a reduced dataset with only top 10 fields and the class field to perform the Naive Bayes classifier. The accuracy of the classifier is 44.11%. The confusion matrix is presented in the table III.

C. Top 20 non-class fields

To perform this ranking, we computed the correlation and sort fields with the absolute value of the correlation of each field. For the top 20, we selected only twenty fields with the highest values of absolute value of the correlation. We show the results in the table IV. We created a reduced dataset with only top 10 fields and the class field to perform the Naive Bayes classifier. The accuracy of the classifier is 43.86%. The confusion matrix is presented in the table V.

Class	1	2	3	4	5	6	7	8	9	10
1	109	16	11	2	0	0	0	0	0	0
2	21	26	14	7	3	5	5	0	0	1
3	1	22	21	7	5	6	5	0	1	2
4	2	1	11	3	5	6	5	0	1	2
5	0	1	4	2	3	5	4	0	0	1
6	0	0	3	3	1	2	4	1	0	0
7	0	1	0	0	0	2	4	0	0	3
8	0	0	0	0	0	3	2	2	0	0
9	0	0	0	0	2	1	2	0	1	1
10	0	0	0	2	1	1	0	1	6	4

Table V
CONFUSION MATRIX OF THE TOP 20 FIELDS

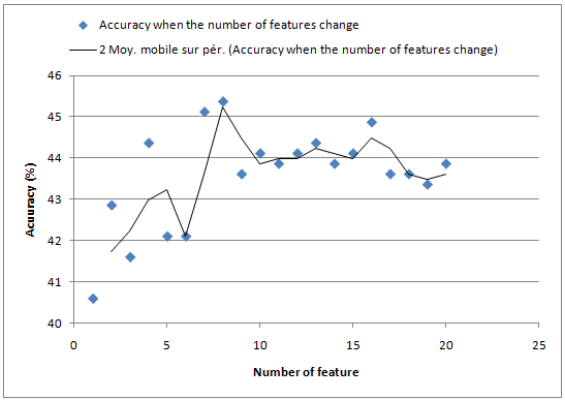

Figure 1. Accuracy when the numbers of features change

D. Discussion

In this part, we will discuss about the accuracy results of the Naive Bayes classifier. We summarize the results on the table VI. This results are presented on the figure 1. The accuracy should increase until a maximum for top 7 fields and decrease and then be monotonous. Basically, we can explain two phenomena. First, the accuracy increases because each additional new field bring more details regarding classification. These fields are a large value of correlation. Then, the accuracy

Ranking field	1^{st}	2^{sd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}
N field	50	44	43	4	45	46	3	18	16	40
Correlation	0.734	−0.730	−0.698	−0.682	−0.659	−0.653	0.628	0.569	−0.567	0.547
Absolute value	0.734	0.730	0.698	0.682	0.659	0.653	0.628	0.569	0.567	0.547
Ranking field	11^{th}	12^{th}	13^{th}	14^{th}	15^{th}	16^{th}	17^{th}	18^{th}	19^{th}	20^{th}
N field	41	28	67	38	32	77	73	30	49	69
Correlation	0.544	0.517	−5162	0.5159	0.499	0.486	0.479	0.478	0.4662	0.4655
Absolute value	0.544	0.517	5162	0.5159	0.499	0.486	0.479	0.478	0.4662	0.4655

Table IV
TABLE REPRESENTING THE TOP 20 NON-CLASS FIELD

Selection top fields	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6	Top 7	Top 8	Top 9	Top 10
Accuracy results	40.60%	42.86%	41.60%	44.36%	42.11%	42.11%	45.11%	45.36%	43.61%	44.11%
Selection top fields	Top 11	Top 12	Top 13	Top 14	Top 15	Top 16	Top 17	Top 18	Top 19	Top 20
Accuracy results	43.86%	44.11%	44.36%	43.86%	44.11%	44.86%	43.61%	43.61%	43.36%	43.86%

Table VI
SUMMUARIZED OF ACCURACY RESULTS OF THE NAIVE BAYES CLASSIFIER

decreases because, each additional new field do not bring more information. The value of correlation is weak. Moreover, we add noise.

III. REDUCTION OF TOP 10 FIELDS USING CORRELATION BETWEEN EACH FIELD

In this part, we reduced the top 10 fields dataset using correlation between each non-class field. We computed the correlation between each non-class field, and we removed one field of the most correlated pair of non-class fields. We repeated this manipulation until have only 5 fields. The next section presents which fields are selected and the results regarding accuracy of classification with the Naive Bayes classifier.

A. Results

After the manipulation, we have only five non-class fields and one class field. We present on the table VII, the number of remaining fields. We also remind the correlation of each remaining field with the class field. We trained the Naive Bayes classifier. The accuracy of the classifier is 46.67%. The confusion matrix is presented in the table VIII.

B. Discussion

We show that with this manipulation, we have a better result than every top fields tested before. Basically, with this manipulation we removed non-class fields which are very similar. We keep only fields which have the best correlation with the class field and all non-class fields are the most different as possible.

Class	1	2	3	4	5	6	7	8	9	10
1	112	16	8	1	0	1	0	0	0	0
2	29	24	11	5	4	6	2	0	1	0
3	9	19	25	5	4	4	2	0	0	2
4	2	5	8	6	4	3	1	0	3	4
5	0	1	6	2	6	2	3	0	0	0
6	0	4	2	1	4	3	0	0	0	0
7	0	1	0	2	2	3	0	0	0	2
8	0	0	0	2	2	2	0	1	0	0
9	0	0	0	0	2	0	1	1	0	3
10	0	0	2	2	0	2	0	0	0	9

Table VIII
CONFUSION MATRIX OF THE REDUCED DATASET

IV. CALCULATING CORRELATION VALUES FOR CATEGORICAL DATA

A. Type of variables

We can define three different types of variable.

a) Quantitative variables: This type of variables could be called numeric variables. It can be represented with number like $1, 0.2, -0.3, -4$.

b) Ordinal variables: This type of variables is representative of ranking. It can be represented like $1_{st}, 2_{sd}, 3_{rd}, 4_{th}$.

c) Nominal variables: This type of variables is represented as a name like *dead, live*.

B. Calculation of correlation coefficient

In this section, we will present different method to compute the correlation depending of the type of

N field	4	46	3	18	40
Correlation	-0.682	-0.653	0.628	0.569	0.547
Absolute value	0.682	0.653	0.628	0.569	0.547

Table VII
TABLE REPRESENTING THE FIELDS OF THE REDUCED DATASET

Variable X \ Variable Y	Quantitative	Ordinal	Nominal
Quantitative	Pearson r_{xy}	Biserial r_b	Point Biserial r_{pb}
Ordinal	Biserial r_b	Spearman ρ	Rank Bisereal r_{rb}
Nominal	Point Biserial r_{pb}	Rank Bisereal r_{rb}	Pearson's contingency coefficient C

Table IX
METHODS WITH DIFFERENT TYPE OF VARIABLES

variables. The table IX shows the methods used with different types of variables.

1) Spearman's rank correlation coefficient: To compute the correlation between two ordinal data, we use Spearman's rank correlation coefficient. The formula allowing the computation of this coefficient is:

$$\rho = 1 - \frac{6 \sum_{i=0}^{n} d_i^2}{n(n^2-1)} \tag{4}$$

where n is the number of elements in field of a set and d_i is the difference between two ranks in field X and Y as $d_i = X_i - Y_i$. The meaning of the result is the same than r_{xy}. The result is included between -1 and 1.

2) Point Biserial r_{pb}: Point Biserial r_{pb} is used to compute correlation between dichotomous nominal variable and quantitative variable. The formula allowing the computation of this coefficient is:

$$r_{pb} = \frac{(M_1 - M_0)}{\sigma_n} \sqrt{\frac{n_0 n_1}{n^2}} \tag{5}$$

where σ_n is the standard deviation (eq. 3) of the numeric field, M_1 is positive member, M_0 is negative member, n_0 is the number of negative sample, n_1 is the number of positive sample and n is the total number of samples. The meaning of the result is the same than r_{xy}. The result is included between -1 and 1.

3) Biserial r_b: Biserial r_b is used to compute correlation between dichotomous ordinal variable and quantitative variable. The variable are the same that for Point Biserial r_{pb}. The formula allowing the computation of this coefficient is:

$$r_b = \frac{(M_1 - M_0)}{\sigma_n} \frac{n_0 n_1}{n^2 u} \tag{6}$$

where u is the ordinate of the normal distribution with zero mean and unit variance at the point which divides the distribution into proportions $\frac{n_0}{n}$ and $\frac{n_1}{n}$. The meaning of the result is the same than r_{xy}. The result is included between -1 and 1.

4) Rank-Biserial r_{rb}: Rank-Biserial r_{rb} is used to compute correlation between dichotomous nominal variable and ordinal variable. The variable are the same that for Point Biserial r_{pb}. The formula allowing the computation of this coefficient is:

$$r_{rb} = 2\frac{(M_1 - M_0)}{n} \tag{7}$$

where M_1 is positive member, M_0 is negative member and n is the total number of samples. The meaning of the result is the same than r_{xy}. The result is included between -1 and 1.

5) Pearson's contingency coefficient C: Pearson's contingency coefficient C is used to compute correlation between two nominal variables. The formula allowing the computation of this coefficient is:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} \tag{8}$$

where n is the total number of samples. χ^2 is defined as:

$$\chi^2 = \sum_{i=0}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{9}$$

where O_i is an observed frequency, E_i is an expected frequency and n is the number of possible outcomes of each event. The result is included between 0 and 1 or the value 1 corresponds to a perfect correlation.