Coursework Assignement 1:

Data Mining And Machine Learning (F21DL) - Lecturer : David Corne

Guillaume Lemaitre

1 Introduction

This report presents a work dealing with various type of data. We study the link between normalization and accuracy of classification using on the one hand *max-min normalization* ans *z-normalization* and on the other hand accuracy of one nearest neighbour algorithm.

To perform different processing and algorithm, I used *C++* programming language. This choice can seem not judicious because on the one hand I had to rewrite every scripts which were given. On the other hand, *C++* is not as flexible as *awk*. Indeed, we need strictly the same organisation inside dataset (two spaces instead of one space can cause a lot of troubles). But I used *C++* for another reasons. I wanted remind the syntax and this exercice was a good exercice for that. Then, it occurred to me that applications created are faster than application implemented in *awk*.

2 General datasets

We work with different dataset and we want to know the impact of the normalization on the accuracy of one nearest neighbour algorithm. We present results of accuracy of one nearest neighbour on non normalized and normalized datasets in table 1.

We can note that the accuracy is better when datasets is normalized. Normalization allows to compare like with like. Indeed, one nearest neighbour algorithm compared the difference between two values in the same field. If one nearest neighbour is apply on dataset non normalized, we compare two values without know the meaning of these values. With normalization, you transform these values to give a meaning knowing the distribution of the field. *min-max normalization* is used to map values of each field between 0 and 1. *z-normalization* don't map values between 0 and 1. It transforms on a value representing the distance (*standard deviation*) of this value compared at the average of all values in the field (*mean*).

Looking more precisely different results, we can see that the difference of accuracy for communities and crimes dataset is closed. In fact, the non normalized dataset is a *min-max normalization* while the normalized dataset is a *z-normalization*. So we can conclude that for this dataset, the *z-normalisation* allow to have a best accuracy.

Regarding authors dataset, the non normalized dataset is a *z-normalization* while the normalized dataset is a *min-max normalization*. Unlike communities and crimes dataset, authors dataset present a best accuracy with *min-max normalization*.

For the others datasets, non normalized datasets are "real" non normalized. Hence, these results confirm that normalization on large dataset is useful.

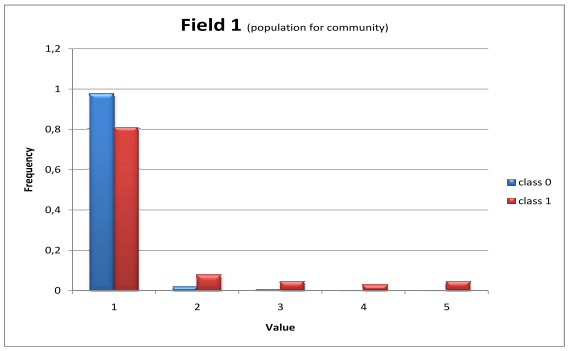
3 Reduced datasets

In this part, we study the specificities of datasets reduced at the first five fields. In first, we study distribution of each datasets and then we peform datasets with two fields.

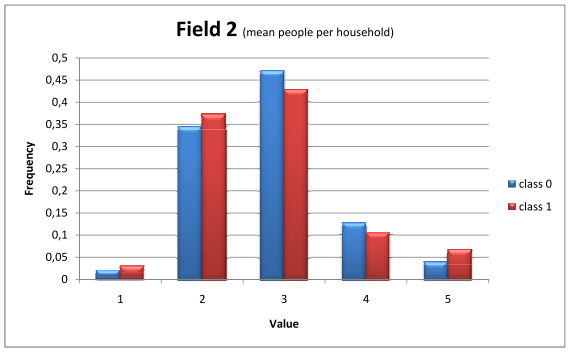
Type of dataset / Name dataset	Accuracy with non normalized dataset	Accuracy with normalized dataset
Communities and crimes	84.65%	84.85%
Pima Indians Diabetes	67.97%	70.58%
Yeast	69.72%	70.33%
Authors	92.31%	98.46%

Table 1: Results of accuracy of nearest neighbour algorithm

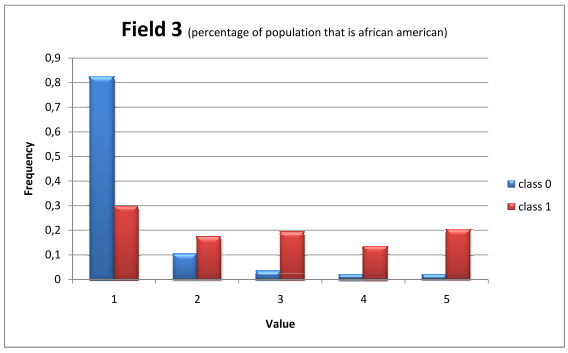
3.1 Histograms of communities and crimes dataset



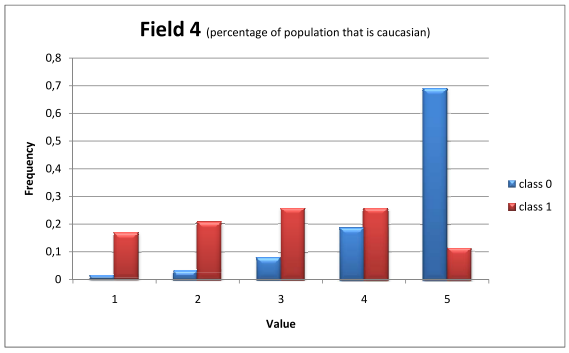
(a) Distribution histogram of population for community



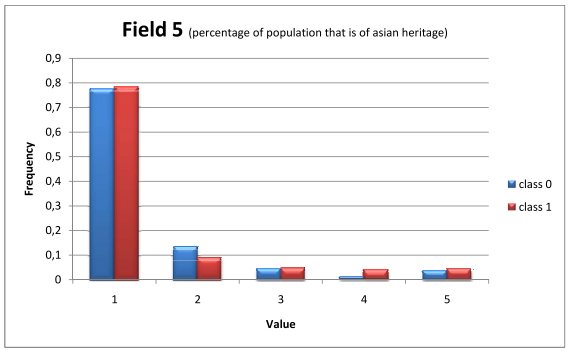
(b) Distribution histogram of mean people per household



(c) Distribution histogram of percentage of population that is african american



(d) Distribution histogram of percentage of population that is caucasian

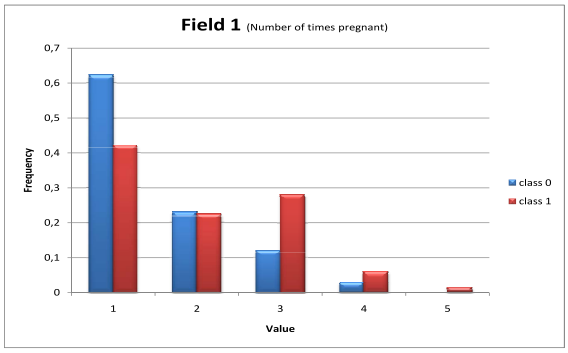


(e) Distribution histogram of percentage of population that is of asian heritage

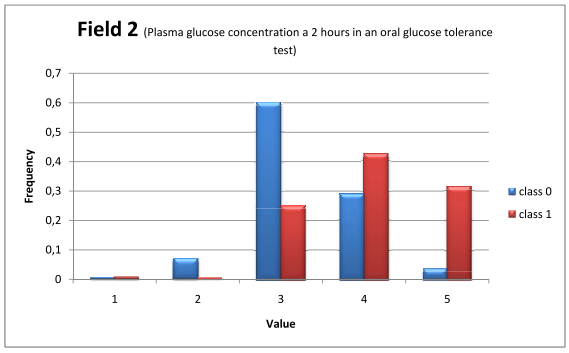
Figure 1: Distribution histograms of five first fields of communities dataset

We can see that distributions of class 0 and class 1 are very different for the field 3 1(c) and 4 1(d). Regarding the other fields, distributions are most homogenous and the difference between the both class is not flagrant.

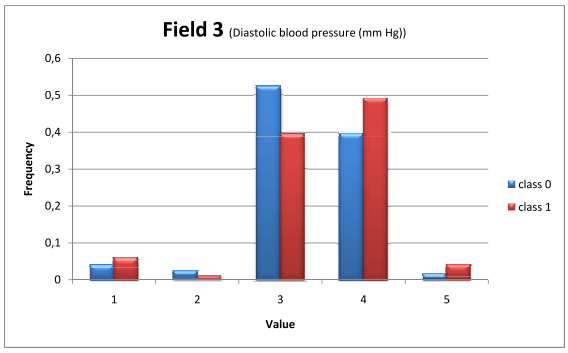
3.2 Histograms of pima indians diabetes dataset



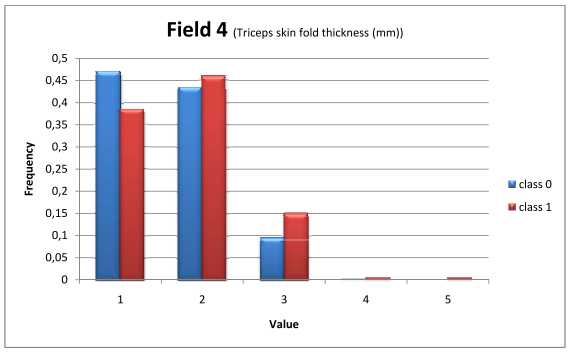
(a) Distribution histogram of number of times pregnant



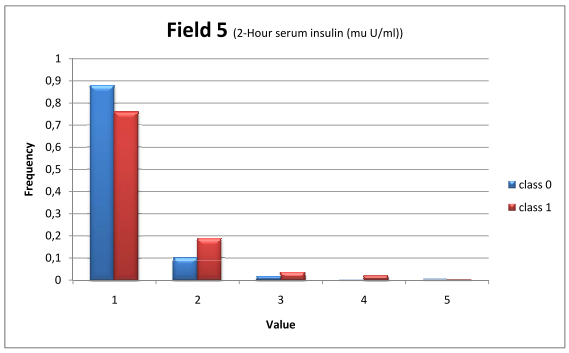
(b) Distribution histogram of plasma glucose concentration a 2 hours in an oral glucose tolerance test



(c) Distribution histogram of diastolic blood pressure (mm Hg)



(d) Distribution histogram of triceps skin fold thickness (mm)

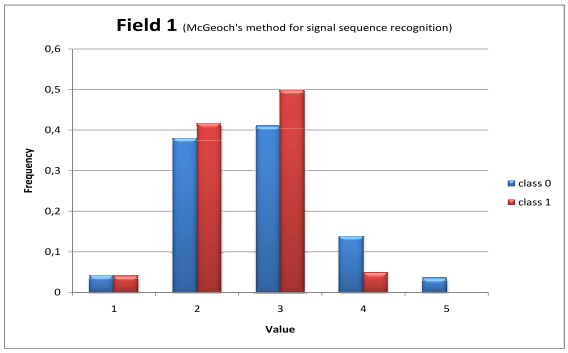


(e) Distribution histogram of 2-hour serum insulin (mu U/ml)

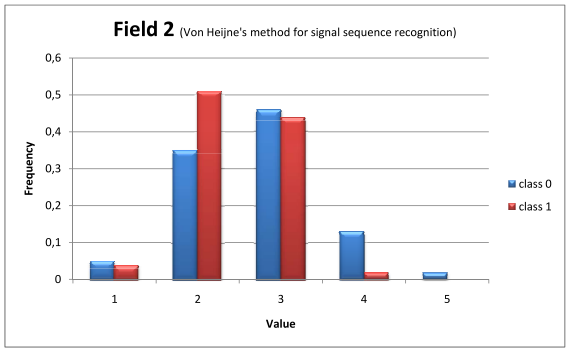
Figure 2: Distribution histograms of five first fields of pima indians diabetes dataset

We can see that distributions of class 0 and class 1 are very different for the field 2 2(b). Regarding the other fields, distributions are most homogenous and the difference between the both class is not flagrant.

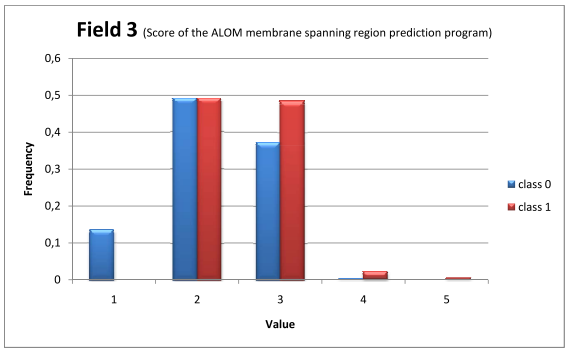
3.3 Histograms of yeast dataset



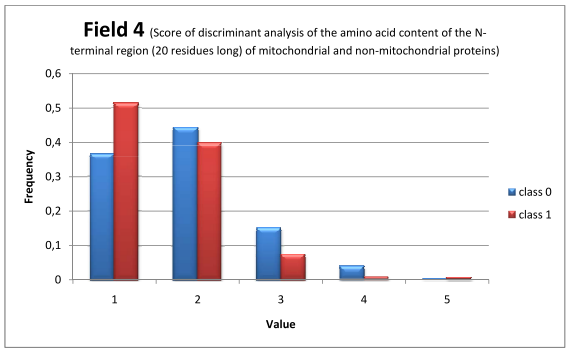
(a) Distribution histogram of McGeoch's method for signal sequence recognition



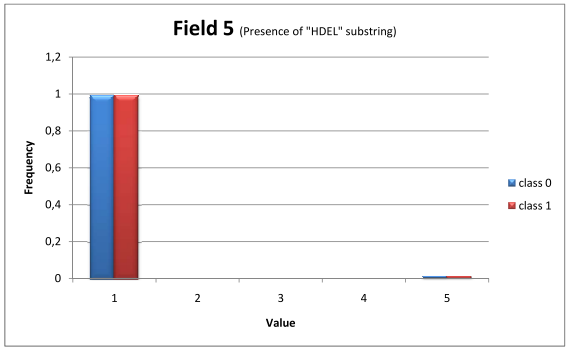
(b) Distribution histogram of Von Heijne's method for signal sequence recognition



(c) Distribution histogram of Score of the ALOM membrane spanning region prediction program



(d) Distribution histogram of Score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins

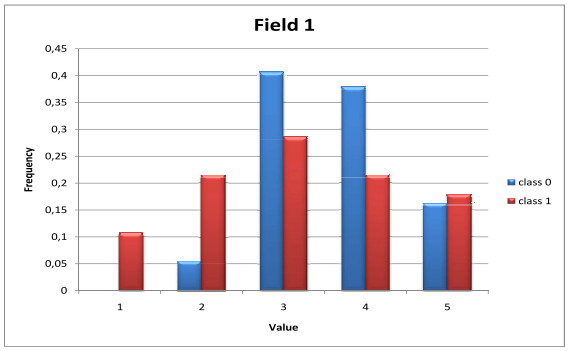


(e) Distribution histogram of Presence of "HDEL" substring

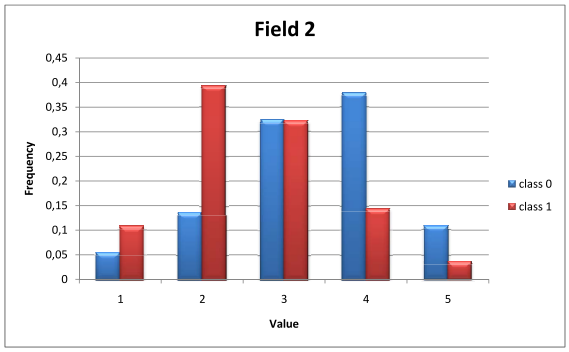
Figure 3: Distribution histograms of five first fields of yeast dataset

We can see that distributions of class 0 and class 1 are not very different for each fields. However, fields 2 3(b) and 3 3(c) are the fields presenting the most differences.

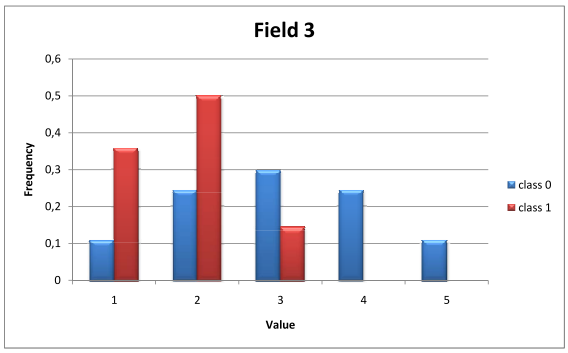
3.4 Histograms of authors dataset



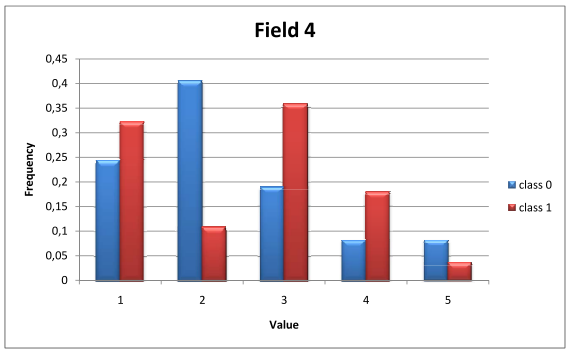
(a) Distribution histogram of the first field



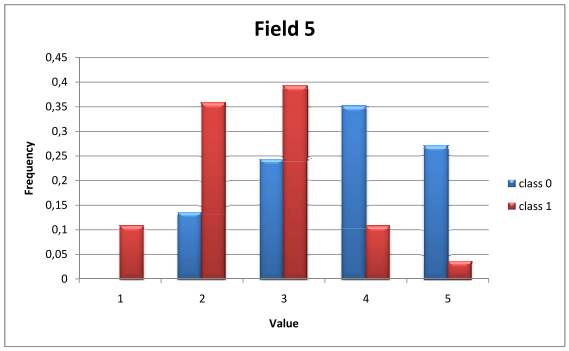
(b) Distribution histogram of the second field



(c) Distribution histogram of the third field



(d) Distribution histogram of the fourth field



(e) Distribution histogram of the fifth field

Figure 4: Distribution histograms of five first fields of authors dataset on general datasets

We can see that every distribution of class 0 and class 1 are very heterogenous. It indicates that all fields are a good indicator for the classification. If we delete fields, we risk to lose accuracy.

3.5 Production of reduced dataset

In this part, we will explain how we produce reduced datasets with distribution histograms previously perform. The datasets used is composed of six fields with the last field corresponding to the class field. We must produce a datasets of three fields with the last field corresponding to the class field. We must find a method to select two most important fields.

Each histogram of the distribution of a field is divided in 5 bins. Hence each bin represents $\frac{1}{5^{th}}$ of the range of the field. We assume that one field is more important if the sum of the difference of each distribution on each class is especially important.

Euclidean distance allows to perform this parameter. We can formalize as follow:

$$df = \sum_{i=1}^{5} (p_{c0}(i) - p_{c1}(i))^2 \quad (1)$$

where $p_{c0}(i)$ is the distribution of the class 0 in the bin i and $p_{c1}(i)$ is the distribution of the class 1 in the bin i.

Hence for each dataset, we calculate the Euclidean distance for each field. We keep the two fields with the value of Euclidean distance the most important. We present the results of the computation and the fields selected in the table 2.

4 Accuracy and reduction

Results of accuracy of nearest neighbour are presented in the following table 3. We can note that the accuracy of reduced database is worse than non reduced database. Indeed, we remove fields which had weight in classification with nearest neighbour algorithm. Hence, it is preferable to have large quantity of relevant fields. To clear dataset, we should remove all fields which are not revelant.

Unlike results for general communities and crimes dataset, we can see that *min-max normalization* is more efficient than *z-normalization*. However, the difference of the accuracy between the two types of datasets is very closed. This remark show that both fields are an identic distribution with *mean* and *standard deviation* closed.

Unlike results for general authors dataset, we can see that *z-normalization* is more efficient than *min-max normalization*.

For the others datasets, the results are very closed. Hence, normalization on this both datasets do not have influence.

Number field / Name dataset	Field 1	Field 2	Field 3	Field 4	Field 5	Number of fields selected
Communities and crimes	0,188	0,063	0,595	0,651	0,054	Field 3 - Field 4
Pima Indians Diabetes	0,259	0,472	0,164	0,105	0,150	Field 1 - Field 2
Yeast	0,136	0,197	0,176	0,175	0,002	Field 2 - Field 3
Authors	0,280	0,360	0,472	0,367	0,445	Field 3 - Field 5

Table 2: Results of the Euclidean distance computation for each field and fields selected for reduced datasets

Type of dataset / Name dataset	Accuracy with non normalized dataset	Accuracy with normalized dataset
Communities and crimes	77.73%	77.58%
Pima Indians Diabetes	64.84%	64.32%
Yeast	63.59%	63.79%
Authors	73.85%	67.69%

Table 3: Results of accuracy of nearest neighbour algorithm on reduced datasets