Milestone 3

Andres Calderon
acald013@ucr.edu

February 27, 2016

1 Introduction

This report describes the advances to accomplish the final project in the course. The main goal of the project is to perform a reliability analysis of a machine learning algorithm (kNN). This third report leads towards a more formal reasoning and understanding of the impact of error injection during the distance calculation of kNN. The main goal of the report is to discuss an initial formula that describe the behavior of the error rate, as well as its formulation and limitations. The analysis and reasoning will be supported by the previous implementation presented in previous milestones.

2 Formal reasoning.

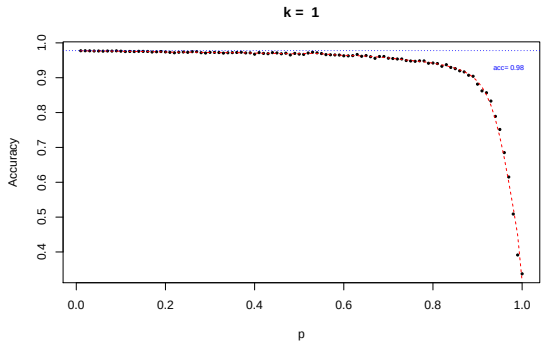
Table 1 summarizes some of the attributes of the studied datasets. *dr* refers to the default rate of the dataset and it is the ratio between the number of elements of the most frequent class and the size of the dataset. It is taken as a reference for classifiers because any resulting accuracy should be better than this, if not, it would be practically better to guess. N refers to the number of available points for training the classifier. C refers to the number of different classes in the dataset and D refers to the number of dimensions or attributes.

Figure 1a shows the difference in the accuracy results of kNN when the probability of error during distance computation increases. Figure 1b show the error rate of the same implementation by just subtracting the value of the original accuracy (without error injection) and the values obtained in figure 1a. We can see that the increasing error rate described by figure 1b can be modeled by an exponential function using the equation 1. α will determine the highest point of the function and β will describe the slope and shape of the curve.

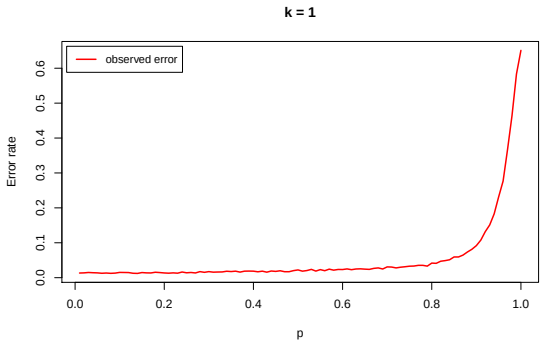
$$Err(p) = \alpha \times e^{\beta \times p} \tag{1}$$

	dr	**N**	**C**	**D**
Cancer	0.63	398	2	32
Iris	0.33	105	3	4
Seeds	0.33	147	3	7
Wine	0.40	125	3	13
Zoo	0.41	71	7	17

Table 1: Description of the datasets.



(a) Accuracies results with increasing error in distance computation.



(b) Error rate plot derived from figure 1a

Figure 1: Accuracy comparison and error rate for Iris dataset.

In our case, α represents the maximum error rate of the function and it relates to the default rate (dr) of the classifier. Even if we introduce a lot of errors, the chance to pick up a correct answer is not below to the default rate. So, we can model α as:

$$\alpha = (1 - dr) \tag{2}$$

As we already mention, β describes the shape of the curve and how fast the error rate increases together with the value of p. In the way that kNN works, the classification of each new point will depend on the number of available points (N) in the search space. Similarly, the probability to belong to a specific class (C) (and specifically to the correct one) depends on how many members of that class are close enough to the new point. For simplicity, we will consider that the number of available points N will distribute equally on the number of classes C, so the number of available points for each class will be $\frac{N}{C}$. Now, we will have to consider that this number is also affected by the default rate (dr). At the moment when $\frac{N}{C}$ would be below of dr, the classifier will start guessing. So, if many points from the correct class are affected for the error injection, the probabilities to misclassified the new point will increase. From this reasoning we can model β as:

$$\beta = \frac{N}{C} \times (1 - dr) \tag{3}$$

Finally, equation 4 illustrates an initial formula for error rate modeling after replacing equations 2 and 3 in equation 1.

$$Err(p) = (1 - dr) \times e^{\frac{N}{C} \times (1-dr) \times p} \tag{4}$$

3 Testing.

We extend table 1 by applying equations 2 and 3 to add columns α and β. Figures from 2 to 6 show the corresponding fitting model by applying equation 4 respectively. The figures compare the observed error rate for the studied datasets (continuous red line) and the plot of equation 4 using the corresponding parameters from table 1 (dashed blue line).

	dr	N	C	D	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$
Cancer	0.63	398	2	32	**0.37**	**74.20**
Iris	0.33	105	3	4	**0.67**	**23.35**
Seeds	0.33	147	3	7	**0.67**	**32.68**
Wine	0.40	125	3	13	**0.60**	**24.97**
Zoo	0.41	71	7	17	**0.59**	**6.00**

Table 2: α and β from values on table 1.

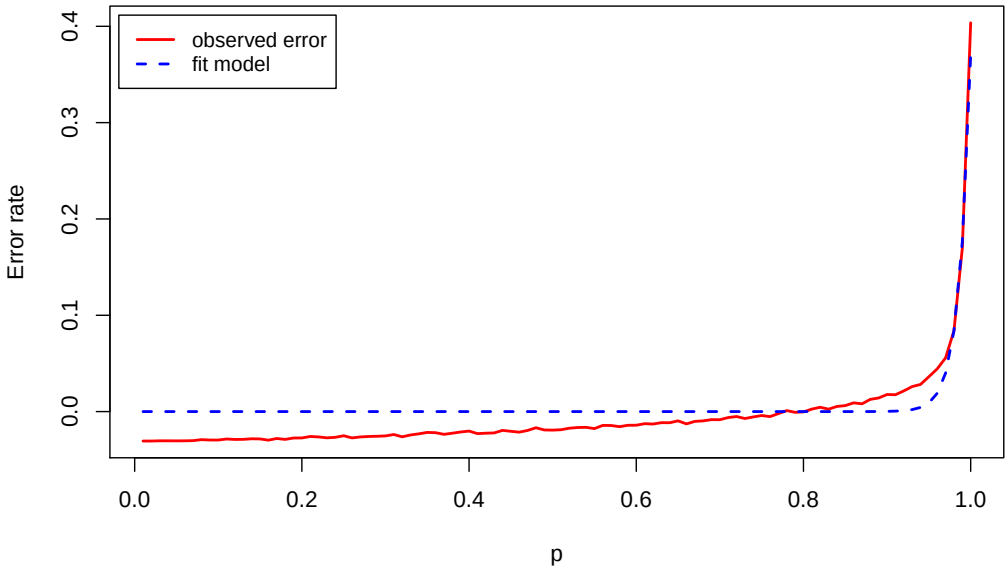


Figure 2: Observed error rate compared to the fitted model for Cancer.

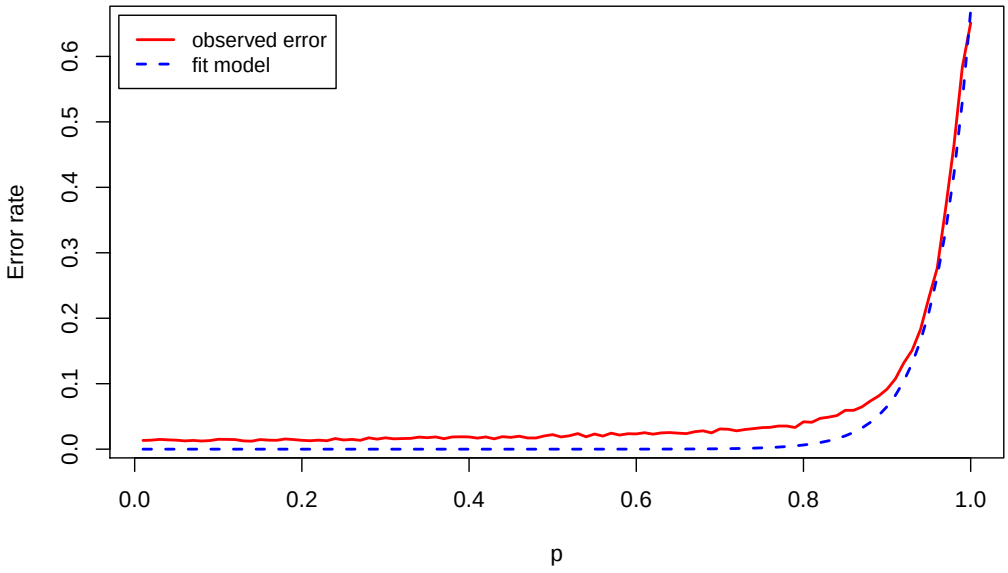


Figure 3: Observed error rate compared to the fitted model for Iris.

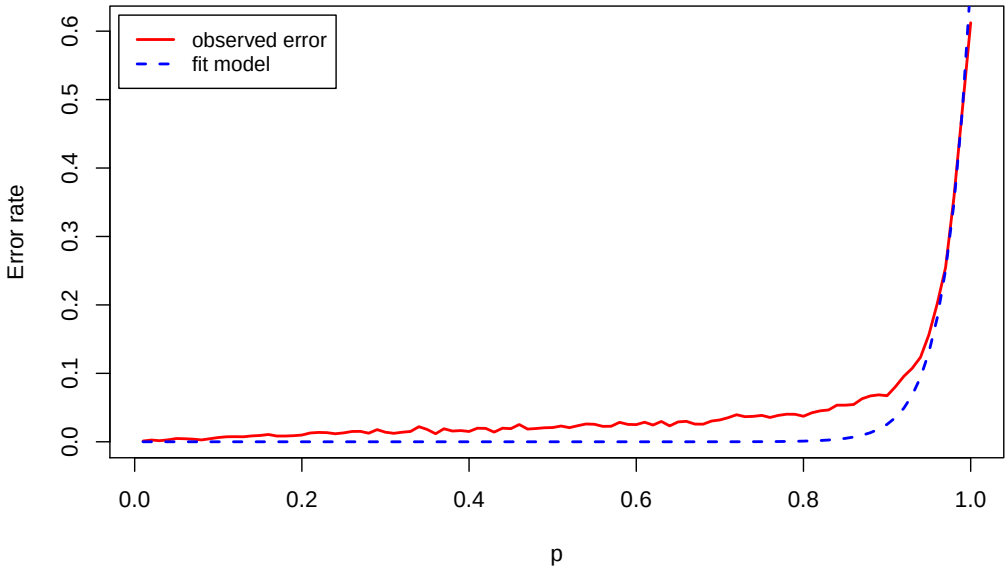


Figure 4: Observed error rate compared to the fitted model for Seeds.

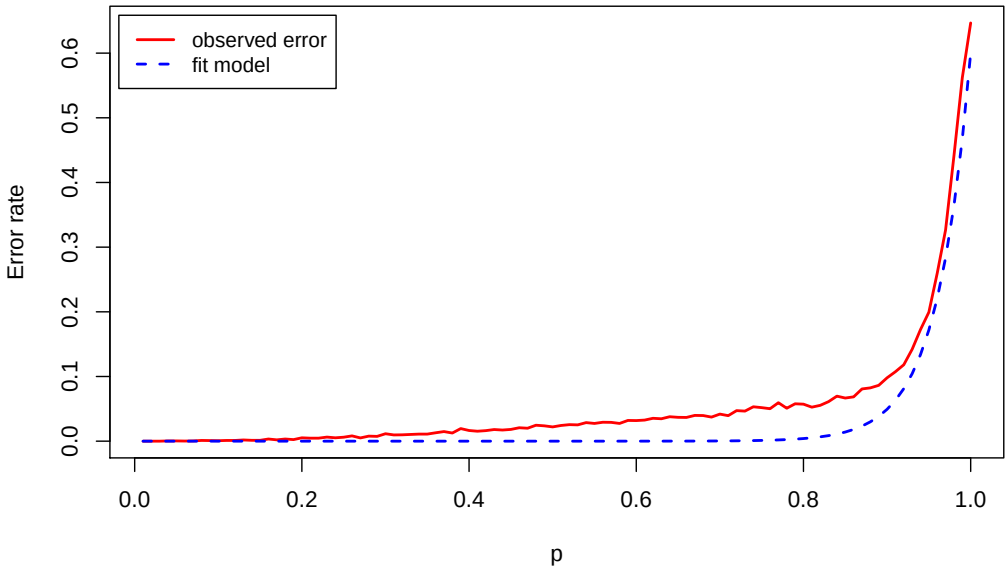


Figure 5: Observed error rate compared to the fitted model for Wine.

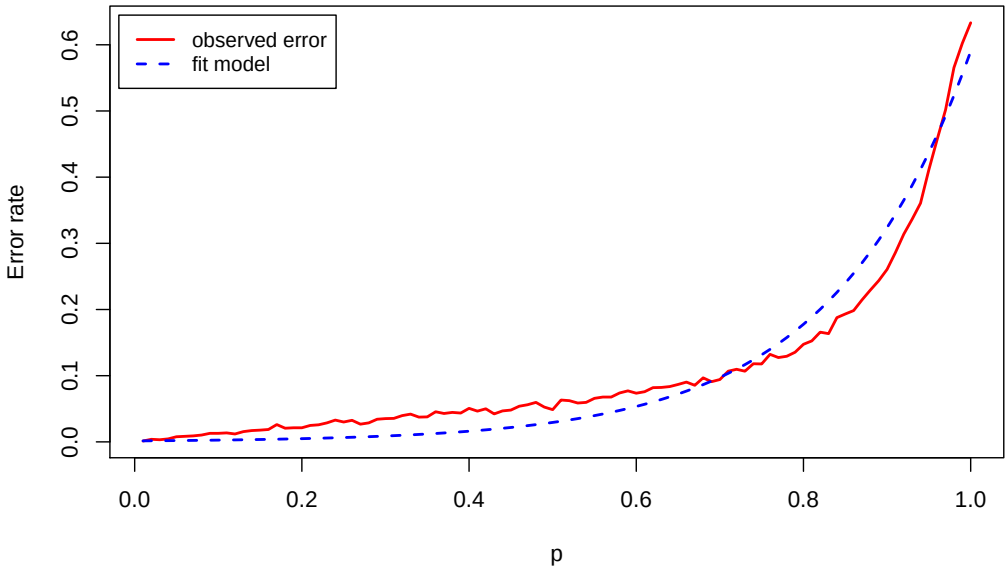


Figure 6: Observed error rate compared to the fitted model for Zoo.

4 Limitations

Analysis in section 2 just takes into account basic components and it is far for optimal. Among the additional considerations it would be important to include the impact of k into the equation 3. Previous analysis tend to indicate that higher values of k increase the impact of error rate modifying the slope of the curve. It seems reasonable to include k into equation 3 due to its role in the selection of the class for new points. If it is required more points to take a decision, the impact of losing potential valuable points by error injection is higher.

Another consideration is the probability distribution between different classes. In this analysis, we consider a uniform distribution of the available points among all the classes. However, in many cases, a particular class is more frequent than the others. So, the assumption of $\frac{N}{C}$ is a generalization which is prone of improvement.

Finally, the asymptotic nature of exponential functions makes difficult to model the initial states of the error rate. In general, the error starts to increase gradually at the beginning before to start growth exponentially. e^x equations run parallel to the x axis missing the initial slow increase of the rate.