Comparing k-NN and Logistic Regression algorithms for Wine classification

Algorithms for classification of wine types and quality

Fábio Pinheiro & João Duro

Abstract

The idea of the project and this report is to see how different KNN and Logistic Regression are. Our main problem divides itself in two sub problems of supervised classification, with one being a binary supervised classification and the other a 7-class supervised classification, and we will see how both algorithms behave in each case.

"""""Com o almento do power computational começa a fazer sentido usar algoritons mais complaxesos para obetremos melhores resultados."""""

I. Introduction

The data we're given consists of a set of measures from 6000 different wines. The attributes of the wines are "fixedAcidity", "volatileAcidity"," citricAcid", "residualSugar", "chlorides", freeSulfurDioxide, "totalSulfurDioxide", "density", "pH", "sulphates", "alcohol", "quality" and "type". Our first goal is try and predict the quality of the wine, which is classified as "White" or "Red, and later, try and predict based on the same attributes the quality of the wine which is a integer variable ranging from 1 to 7, 1 being really good, and 7 being really poor. Based on the results and our understading of how the algoithms work, we will try and see how different it is to predict to a binary class or a 7-value class.
"""""""""""""""ACRESCENTAR MERDAS AQUI""""""""""""""""""

II. Methods

Normalize data

k-nearest neighbors algorithm (k-NN)

K-nearest neighbors algorithm also known as k-NN is a non-parametric method used for classification and regression. In both cases the k-NN algorithm starts from the principle that similar data are closed one to another.

The algorithm is just given a point 'Y' to find the 'k' points from training data where distance where is smaller that all the anthers. Then choose the class more frequent, in case of a tie, pick one at random among them.

The pseudocode for k-NN can be seen in Algorithm 1 [2]

Euclidean Distance

For n dimensions Euclidean Distance is give by following formula:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}$$

(Dimensions in here is the number of feature/parameters of the data)

Logistic Regression

"It is assumed that we have a series of N observed data points. Each data point i consists of a set of m explanatory variables $x_{1,i}x_{2,i}\ldots x_{m,i}$ (also called independent variables, predictor variables, input variables, features, or attributes), and an associated binary-

Algorithm 1 KNN

1: Define the distance measure or similarity between two objects
2: Find k;
3: Compute the distance between the new object and all objects in the training set: $d(xi, x0)i = 1, 2, ..., n$;
4: Sort the distances in increasing numerical order and pick the first k elements (neighbours), let be $Vk(x0) \subseteq D$, the set of those neighbours;
5: Save the classifications of all neighbours;
6: Assign new object to the class based on majority vote of its neighbours 2, i.e.

$$\mathcal{Y}_0 = \overset{arg\ max}{C_j} \sum_{(x_i, y_i) \subset V_(x_0)} I(y_i = C_j)$$

valued outcome variable Y_i (also known as a dependent variable, response variable, output variable, outcome variable or class variable), i.e. it can assume only the two possible values "failure" or 1 "success". The goal of logistic regression is to explain the relationship between the explanatory variables and the outcome, so that an outcome can be predicted for a new set of explanatory variables."

III. Experiments

data:

- Number of parameters: 11
- Number of samples in training set: 5000
- #Red: 906
- #White: 4094
- #Quality1: 5
- #Quality2: 161
- #Quality3: 854
- #Quality4: 2197
- #Quality5: 1604
- #Quality6: 159
- #Quality7: 20
- Number of samples in test set: 1000

Wine type:

To classify the type of wine first we used the k-nearest neighbors algorithm. To choose the K we separate the training dataset in two set (3750 points for training and points 1250 points for validation).

After to improve the results we try to use the some method but with the data normalize.

Wine quality

IV. Results

Table 1: *My current knowledge of tables*

Table type	Likely location
Coffee table	Living room
Dining table	Dining room
Bedside table	Bedroom

I. Confusion matrix for the wine quality

Table 2: *Using k-NN with k=1*

Quality	1	2	3	4	5	6	7
1	0	0	0	0	0	0	0
2	0	13	7	5	1	0	0
3	0	9	114	49	9	0	0
4	0	6	44	272	78	8	0
5	0	0	11	99	225	9	0
6	0	0	3	11	12	10	0
7	0	0	0	1	3	1	0

Table 3: *Using logistic regression*

Quality	1	2	3	4	5	6	7
1	a	b	c	d	f	g	i
2	a	b	c	d	f	g	i
3	a	b	c	d	f	g	i
4	a	b	c	d	f	g	i
5	a	b	c	d	f	g	i
6	a	b	c	d	f	g	i
7	a	b	c	d	f	g	i

V. Discussion

space and computers power espaço ocupado pelo knn e pelo logistic regresion

The values of quality are not independent with this a mean if a wine and classify was quality 'x' with is not the correct one, but is much more probably that the right quality is $x-1$ or $x+1$ than be $x-2$, $x+2$, $x-3$, $x+3$ ...

We could have used this to classify the quality of wine.

One of the motives for the result be soo low for the quality of one wine is a subjective thing, two wine enthusiasts do not necessary give the same qualification for the same wine. This way it would be interesting to compare the result our classifier with two independent groups of human wine enthusiasts for each wine. Since the quality is a subjective we would expect the error between our result and the human result should be close to the error between the two independent groups of humans.

k par and k odd

References

[1] Alpaydin, Ethem - Introduction to machine learning, MIT press, 2004.

[2] Concei cão, Amado - Lecture Slides: Statistical Methods in Data Mining (2014). Instituto Superior Técnico, Portugal.

[3] Ng, Andrew - Machine Learning. Multiclass Classifier: One-vs-All, `https://class.coursera.org/ml-005/lecture/38`.

[4] Howe, Bill - k Nearest Neighbors, `https://class.coursera.org/datasci-001/lecture/161`

Appendix A