Comparing k-NN and Logistic Regression algorithms for wine classification

Algorithms for classification of wine types and quality

Fábio Pinheiro & João Duro

Abstract

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

I. Introduction

TODO Something in this document. This paragraph contains no information and its purposes is to provide an example on how to insert white spaces and lines breaks.
Something in this document. This paragraph contains no information and its purposes is to provide an example on how to insert white spaces and lines breaks.

When a line break is inserted, the text is not indented, there are a couple of extra commands do line breaks.
This paragraph provides no information whatsoever. We are exploring line breaks.
And combining two commands

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 a random one random among the most frequented.

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)

???????? Distance

TODO

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 $x_{1,i} x_{2,i}$ ... $x_{m,i}$ (also called independent variables, predictor variables, input variables, features, or attributes), and an associated binary-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

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

F

IV. Results

- Number of parameters: 11
- Number of samples in training set: 5000?
- Number of samples in test set: 1000
- #Red: 906
- #White: 4094
- #Q1: 5
- #Q2: 161
- #Q3: 854
- #Q4: 2197
- #Q5: 1604
- #Q6: 159
- #Q7: 20

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

Appendix A