NOM BOUKHELOUA

Prénom Rémi

Promo 2020

Date 10.01.19

20150257: BOUKHELOUA Rémi
M1:
ST2ML1-DE (10/01/2019)
Amphi orange

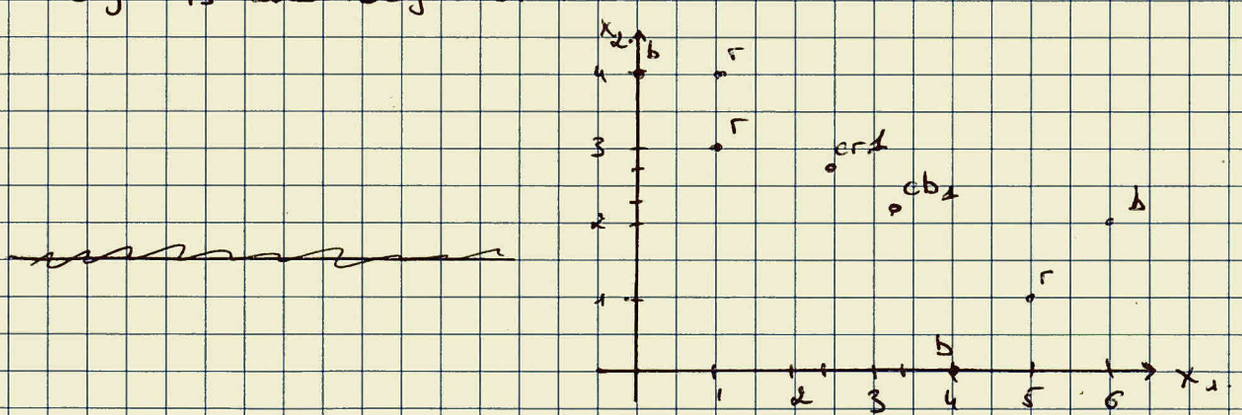
15,5

MATIÈRE Machine Learning (1/2)

1. KNN regression is a method used to predict values based on nearest neighboors. ($\frac{1}{k}\sum_{i=1}^{k} y_i$) meanwhile KNN classification is a method to ~~show~~ separate the dataset in to groups & ~~to~~ based on distance to the neighboors. It's output will be a label (red, blue or green color).

2. 3. With logistic regression, we have a hugher ~~one~~ error rate than in KNN classification in both training and testing set. ~~Moreover, we have different errors in logistic method which means that~~ So, KNN classification is has a better accuracy. But, logistic has a hugher interpretability, it is easier to understand it. Anyway, I would use KNN classification because of its ~~accuracy~~ error rates.

5.



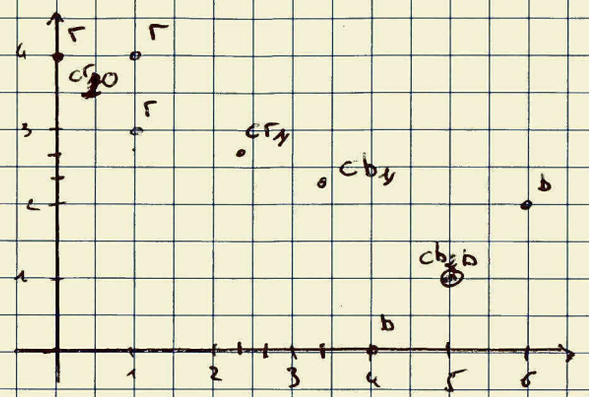
we randomly assign ~~color~~ label to the points (b or r)

we calculate the center of the two groups:

$cr_1 = \left(\frac{1+1+5}{3}; \frac{4+3+1}{3}\right) = \left(\frac{7}{3}; \frac{8}{3}\right)$ and $cb_1 = \left(\frac{10}{3}; \frac{2}{3}\right)$.

Now we will reassign label of points based on closest center point.

If ~~the~~ you are closer to cr_1 we label you r (same with cb_1).



now we calculate new centers based on new labels:

$cr_2 = (\frac{2}{3}; \frac{11}{3})$ and $cb_2 = (5, 1)$

now we reassign labels based on cr_2 and cb_2 and we find same groups so we can stop.

we will have as a result.

Obs	x1	x2	Label
1	1	4	r
2	1	3	r
3	0	4	r
4	5	1	b
5	6	2	b
6	4	0	b

6. That means that ~~if~~ ~~use~~ the value of the first component variable counts as 10% on the calculation of the estimated value. If we add 1 to the value of the variable, the estimated value will gain 10%.

7. k-fold cross validation is based on train - test sets. It will cut the set into k-folds and will perform train on k-1 fold and test on the one not used in train. It will do it k times in order to use ~~at least~~ one every fold as a ~~train~~ test set. We commonly use k=5 or 10.

~~a 2 value clustering only in two groups,~~

It will keep on mixed MSEs and choose best parameters.

a. ~~Dividing in two sets gives that your model will fit~~
~~k fold cross validation will not have variance problem as~~
~~it didn't use~~
k fold performs a lot of train test and picks the best so it will quite always have a better accuracy than just train-test once. But as it performs a lot of test it takes more time to execute than just dividing in two sets.

b. Leave one across means doing k fold with only one value in folder. So on large dataset, it will be much longer to perform than a k-fold. Moreover it uses almost all data as train so model can be have high variance we can't see because of small test set.

8. ~~As~~ It seems that we used a linear regression for a non linear model. ~~As~~ because we have residual that are ~~forming~~ sometimes low, sometime high but "equally" spread on true (this is nul) I would change it to polynomial model ~~or use KNN method~~

9. ~~As we have constant points~~ For a low Y, we have low residuals so the model is fitting well for low Y. But for the bigger ones, we have ~~under~~ a bigger residual ~~(abs)~~ (absolute). I think that this is also due to the choice of linear regression. I would try to use a KNN method and compare it to this model.

~~or a polynomial~~

Or I would test a polynomial as residual gets bigger when Y grows like x compared to x^2.

10 Logistic regression is a classification method. It assigns labels and separate dataset in two groups according to a threshold.

For a given threshold, we will obtain a matrix ~~to show~~:

	in reality: Positive	Negative
predicted Positive	TP	FP
Negative	FN	TN

Based on this we can calculate. Sensitivity = $\frac{TP}{TP+FN}$.
that shows part of positive well predicted
and Specifity : $\frac{TN}{TN+FP}$ that shows part of negative well predicted.

For each threshold you can calculate both sensitivity and specifity. So, for each threshold you have one point of the ROC. The best case is ~~sensitivity~~ sensitivity = specifity = 1.

In the example, for specifity = 0,9 we have sensitivity = 0,35 so we have only 35% of positive ~~predicted~~ predicted as positive. and 90% negative as negative. ~~In the~~

For specificity = 0,1, we have sensitivity = ~~0,97~~ 0,97 so 97% of positive predicted as positive and 10% of negative as negative.

None of them is ~~the best~~ better, it depends on the case we always study. If detecting positive, like detecting ebola is more important we would go for a high sensitivity.

2. To calculate β_0, we use $\beta_0 = \bar{y} - \beta_1 \bar{x}$ (in 1 dimension)
so we have $\bar{y} = \beta_0 + \beta_1 \bar{x}$ with β_0 and β_1 coefficient of the linear regression. For any linear regression, the line will ~~always~~ always pass by $(\bar{x} = \frac{1}{n}\sum x_i ; \bar{y} = \frac{1}{n}\sum y_i)$

NOM BOUKHELOUA

Prénom Rémi

Promo 2020

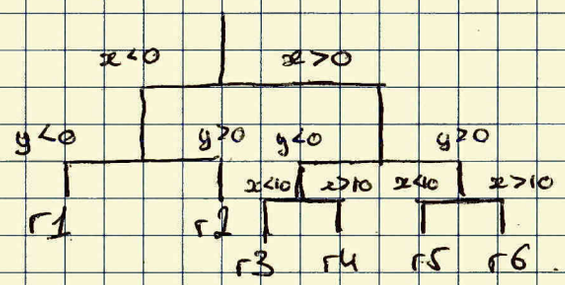
Date

MATIÈRE Machine Learning (2/2)

11. Elbow method will help us to choose number of cluster. We will ~~de~~plot the graph of ~~number of cluster~~ MSE by number of cluster.
   Thanks to the slope we will pick a ~~nb~~ number of cluster with low MSE but not that much which would result in to much clusters.
   ~~If we use~~

4.

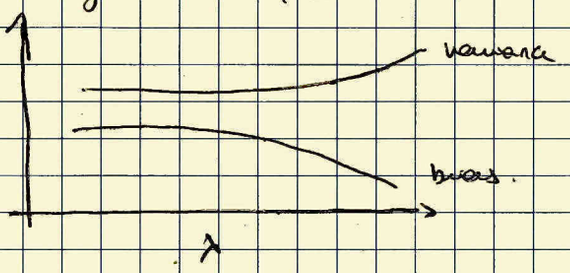
x	y	label
-5	-3	r1
-3	2	r2
5	-2	r3
12	-4	r4
3	3	r5
22	12	r6



If we had a column having 3 ~~different values "red", "blue"~~ and "green" we could

12. If your bias is low, that means your model fits well the reality ~~the train set~~. If your variance is low that means your are not flexible to test set, you are to dependent on your train test. They are anticorrelated, if one is low, the other is high.

In ridge method, we want to minimize $\lambda \sum(\beta_i)^2$ too



we will pick the point where variance and bias are closest.