Instituto Superior Técnico

Departamento de Engenharia Electrotécnica e de Computadores

Machine Learning

4^{th} Lab Assignment

Shift Tuesday 17:00 Group number 19

Number 78308 Name Rui Daniel Ribeiro Dias

Number 75494 Name Nuno Pereira Azevedo Wallenstein Teixeira

Naive Bayes classifiers

Naive Bayes Classifier

Naive Bayes classifiers normally are rather simple, and are very effective in many practical situations. Describe in your own words how the Naive Bayes classifier works. Be precise. Use equations when appropriate.

R: Naive Bayes is based on the Bayes Theorem and it's works based on Conditional Probability.
The different is that Naive Bayes assumes that the features are independent and so much less computational work is required to solve the problem. With Naive Bayes we can calculate the probability of a certain feature to belong to a certain class in the following way:

$$p(x_1, \ldots, x_p|\omega_k) = \prod_{i=1}^{p} p(x_i|x_1, \ldots, x_{i-1}, \omega_k) = \prod_{i=1}^{p} p(x_i|\omega_k)$$

Knowing the conditional distribution of one feature , if multiply that with the conditional distribution of the other features and then we multiply the total result with the probability of that class, we obtain the probability of a certain input X(X1,X2...Xp) be classified in the class we are evaluating.

$$P(C = c) \prod_{i=1}^{k} P(F_i = f_i \mid C = c).$$

If we do this procedure for all classes we can obtain the probability of a certain feature be classified as being in that class. With that information we choose the one with highest probability.

So:

$$\operatorname{argmax}_c \; P(C = c) \prod_{i=1}^{k} P(F_i = f_i \mid C = c).$$

A simple example

1. Obtain a scatter plot of the training and test data, using different colors, or symbols, for the different classes. Don't forget to use equal scales for both axes, so that the scatter plot is not distorted.

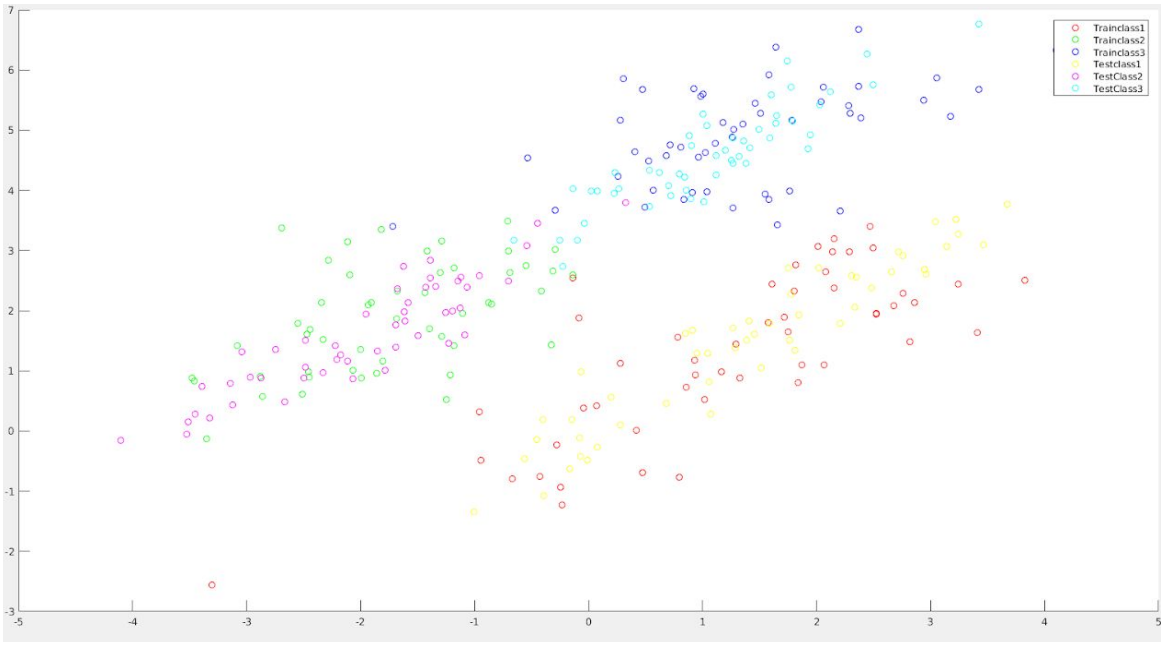

Fig1. Plot of the Train and Test Data

It is possible to see that there is a correlation with the data, as X1 increases X2 also increases. This is a good sign when we are training our data because if the data is not related, it will be difficult for the model to predict the class that a certain input belongs.

We can see that there are some sets of data that are more dispersed than others. In general, the training data is more dispersed than the test data. This is an advantage to our training because it makes it possible to increase the range of the mode since the values are not close to each other, reducing the extrapolations when in the boundaries. Overall we can see that each training and test data is clustered partly around its class.

3. Plot the classifications of the test data as a function of the test pattern index.

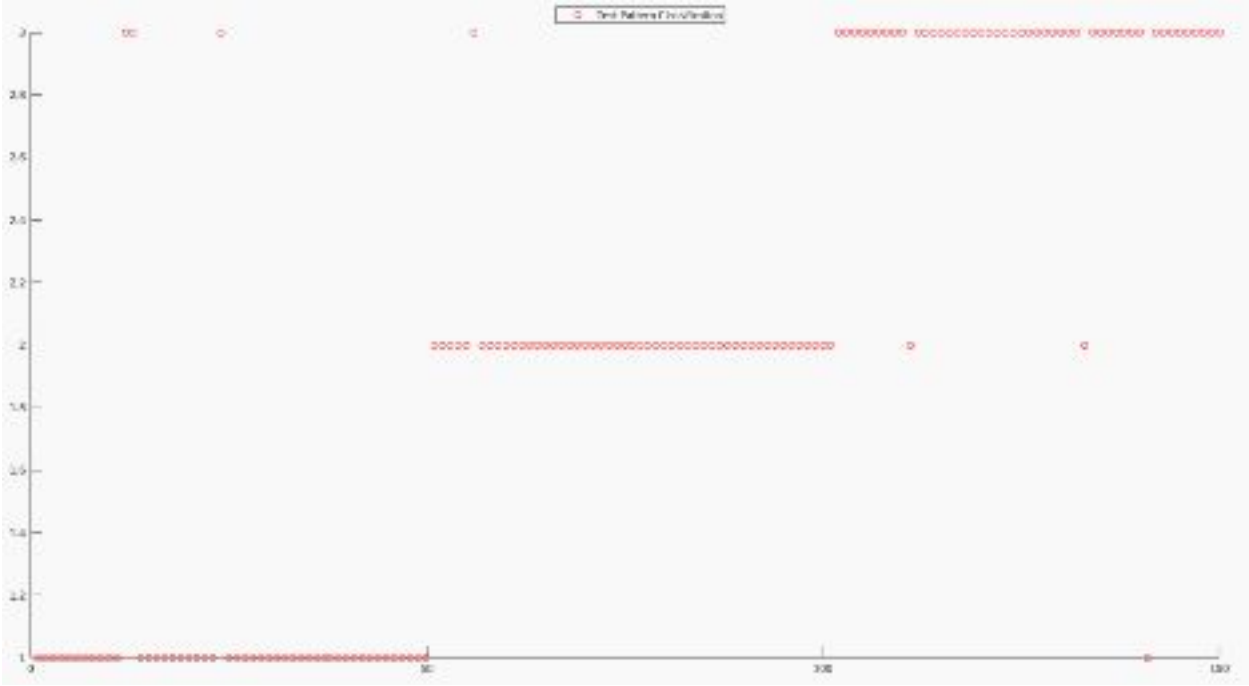

Fig2. Plot of the Test Data as Test Pattern Index

We can see from this plot that only a few inputs were no correctly classified, it our case, we identified 8 input values that were wrong classified.

4. Indicate the percentage of errors that you obtained in the test set.
Test set error rate: We obtained a test error or 5.333%

As we mentioned, we observed 8 inputs that were miss classified.

5. Comment the results.

The results obtained show that the model works well, since we observed an error of only 5.333%. The variance results shows that there are some set of data that are less dispersed than others. One of the remarks about this aspect is the fact the training data is more dispersed than the test data. This has some advantages because we can predict results since the extrapolations are smaller making it possible to increase the range of prediction of the model.

With an error of 5.333% this model has great accuracy to predict in which class does the data belong.

Language recognizer

3.1.1 Practical assignment

1. Complete the code given in the file language recognizer.m. Transcribe here the code that you have added to the program. Clearly separate and identify Sections 1, 2 and 3 of the added code. Include comments.

```
%-------------------------------------------------------------------------
%Code Section 1
%In this section we initialize a variable called SumPCond_trig to 0 which
%has the meaning of a cumulative log conditional probability function
        SumPCond_trig =0;
%-------------------------------------------------------------------------
%Code Section 2
%In this section we accumulate the sum of the conditional proba
        %bility regarding the Naive Bayes Classifier
        %We applied the Laplace smoothing by adding +1 to the variable
        %trigramcount and in the denominator by adding the total number
        %trigrams with 60 possible characters.
                SumPCond_trig=
SumPCond_trig+log10((trigramcount+1)/(total_counts(languageindex)+60.^3));

%-------------------------------------------------------------------------
%Code Section 3
  % In this section we calculate the probability à priori of the
        % class, in this case the language. Laplace smoothing was applied


PLang=log10((total_counts(languageindex)+60.^3)/(total_counts(1)+total_counts(2)+total_count
s(3)+total_counts(4)+4.*60.^3));

        %Since we are working with logs the product of the argument results in
        %a sum of logs, therefore P(C_k|x_1,x_2...,x_p)=P(C_k)*Productory_{i=0}^p(P(x_i|C_k))
        %turns into log(P(C_k|x_1,x_2...,x_p))=log(P(C_k)+sum_{i=0}^p
        %(log(P(x_i|C_k))

        scores(languageindex)=PLang+SumPCond_trig;
```

%----------

2. Once you have completed the code and verified that the recognizer is operating properly, complete the table given below, by writing down the results that you obtained for the pieces of text that are given in the first column.
The last piece of text is intended to check whether your recognizer is able to properly classify relatively long pieces of text. It is formed by the sentence “I go to the beach. ” repeated ten times (in the table, the piece of text is abbreviated). Note that the given sentence has a blank space after the period, so that the repeated sentences are grammatically correct. You may use copy and paste operations to ease the input of this piece of text.

Text	Real Language	Recognized Language	Score	Classification Margin
O curso dura cinco anos	Pt	pt	-71.1105	0.28695396
El mercado está muy lejos	Es	es	-80.239483	8.3772982
Tu vais à loja	Pt	fr	-50.226571	3.929884
The word é is very short	En	en	-85.835754	0.13807796
I go to the beach….I go to the beach	En	en	-576.64454	113.2946

3. Give a detailed comment on the results that you have obtained for each sentence.

A matlab script was made regarding the information of section 3 -Language Recognizer. With this script it was possible to fill the table above tackling the problem of language recognition.

The first phrase "O curso dura cinco anos.", is a portuguese phrase and our model correctly predicts the language associated. Due to the latin similarities with spanish the classification margin, i.e. the difference between the scores of the most recognized and the second most

recognized language is 0.3. This is not much. This difference can be amplified if put a longer sentence since there will be more trigrams regarding the language that we are using.

The second phrase "El mercado está muy lejos.", is clearly a very spanish phrase and our program recognizes this since we have a classification margin of 8.3772982 which is resonably big difference since we are talking about differences of logarithms. One of the factors for this difference that distinguishes this phrase with a spanish idiom are the words "El" and "muy" which have no analog in portuguese, or french or english.

The third phrase "Tu vais à loja." was recognized by our script as a french, even though it is a portuguese phrase. This is a clear error made by our script. The reasons for this error are the words "Tu", "vais" and "à", are words that are communly used by both languages often making that it is harder to distinguish portuguese from french in this case. One of the ways to increase the accuracy of our script is, since we are taking each trigram as independent, even though they aren't, is to increase the number of total trigrams in the text. Adding more words related to the idiom in question, to the text really improves the performance of the script. For example if the phrase was "Tu vais à loja amanhã." our model would have predicted as a portuguese phrase becouse we increased the total number of trigrams with a unique portuguese word "amanhã".

The forth phrase "The word é is very short." was recognized as english and in this phrase the word "é" has a large contribution on the recognition of the language since it has nothing to do with english and everything to do with portuguese. It is possible to obtain a value since we applied laplace smoothing to make it so that our script doesnt go to -infinity when we encounter no trigrams. With a classification margin of 0.13807796 this phrase is very sensible since if we try to recognize this same phrase without the last dot, i.e. "The word é is very short" the program will recognize this phrase as portuguese with a classification margin of 0.17162468.

The fifth portion of text, "I go to the beach. I go to the beach. I go to the beach. I go to the beach. I go to the beach. I go to the beach. I go to the beach. I go to the beach. I go to the beach. I go to the beach." has a really long length and it exemplifies the usefullness of logarithm in the calculation of the probabilities regarding the Naive Bayes Classifier. The bigger the argument, the lower the growth of the log function, since the probability is not 0. Due the big length of the text, it is recognized as english by a large margin with a classification 113.2946 with not much margin for error