Instituto Superior Técnico

Machine Learning
Lab 4 - Bayes Classifier

André Pereira	90016
Anton Erikson	98037

Group 16

2 Pm, Friday

Teacher: Catarina Barata

1

The Bayes classifier is a classification boundary that will tell us the label of a given observation. It can be written has:

$$C(x) = \underset{r=1,...,M}{\text{argmax}}\, P(Y = r|X = x),\ where\, X = observation\, and\, Y = label \tag{1}$$

Meaning, the Bayes classifier will give us if a given point x in the observation space has a given label r, and it will be the label with the greatest $a\, posteriori$ probability. The difference from the Bayes classifier to the Naive Bayes classifier is:

- The usage of the Bayes classifier assumes that we know or can accurately compute the $a\, posteriori$ probability of all classes, which makes it "easy" to solve and find out all the decision boundaries for all the labels.

- The Naive Bayes classifier is used when we cannot compute the $a\, posteriori$ probability of all classes, because the estimation of the conditional distribution P(x|ω_k) is a difficult problem, and the conditional distribution of each class might be dependent on other classes, which makes it´s estimations unfeasable. Therefore, the Naive Bayes classifier assumes that all the observations are conditionally independent.

$$P(x_1, ..., x_n|\omega_k) = \prod_{i=1}^{n} P(x_i|\omega_k) \tag{2}$$

2

As mentioned the Naive Bayes Classifier is a variant of the Bayes Classifier where we assume that all observations are conditionally independent. This is often not the case and the assumption makes for a less accurate classifier. This is what we observe when we classify the test data with a Naive Bayes Classifier and a Bayes Classifier respectively. The Naive Bayes Classifier gives an error percentage of approximately 4% while the Bayes Classifier gives an error percentage of approximately 2.7%. A visualization of the Bayes Classifier is shown in Figure 1.

A perfect classifier is practically impossible. This becomes clear looking at the plots of the test and the train data. Some of the points from different classes lay in almost exactly the same place. There is an overlap.

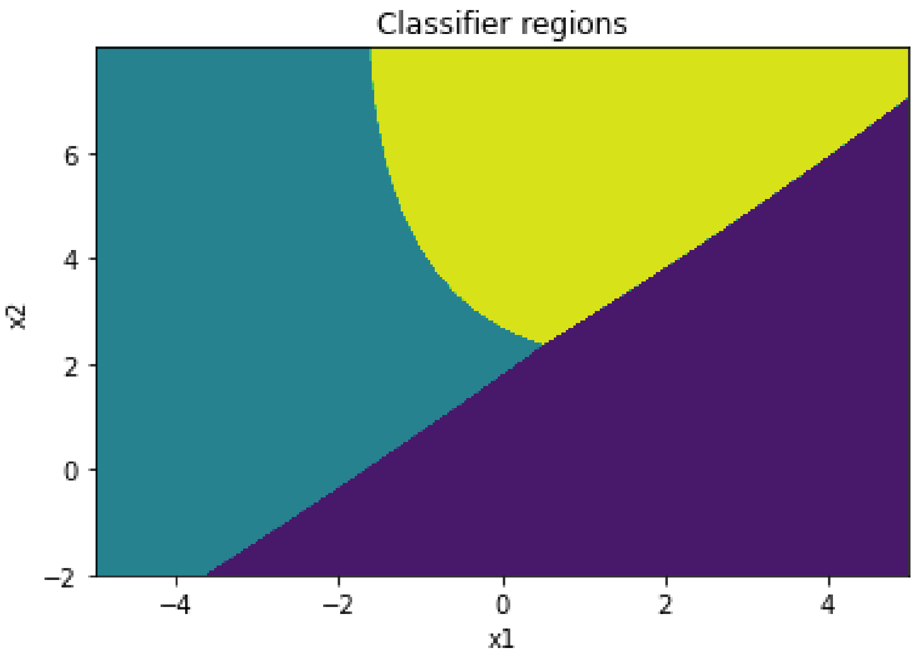

Figure 1: Bayes Classifier regions.

These were the achieved decisions boundaries.Purple is the decision boundary for 1, Yellow for 2 and Blue for 3

3- Language Recognition

Table 1: Obtained results for the language recognizer

Text	Real Language	Recognized Language	Score	Class. Margin
Que fácil es comer peras.	es	es	67.0347%	34.069%
Que fácil é comer peras.	pt	pt	99.999%	99.999%
Today is a great day for sightseeing.	en	en	100%	100%
Je vais au cinéma demain soir.	fr	fr	100%	100%
Ana es inteligente y simpática.	es	es	99.997%	99.994%
Tu vais à escola hoje.	pt	fr	79.305%	58.61%

In order to do the language recognition, we took the dataset of all the trigrams provided to us, and applied the Naive Bayes classifier, making us look at each trigram independently. So, in order to provided an accurate statement on the language of each sentence, we computed:

$$lang_{predicted} = \underset{\omega}{\text{argmax}} \prod_{i=1}^{n} P(x_i|\omega_k) \tag{3}$$

where n is the amount of trigrams in each sentence. Furthermore, we had to use aswell Laplace Smoothing for the computation of the conditional distributions for each class, to make sure that every trigram had an associated probability value different than 0:

$$P(x_i|\omega_k) = \frac{\#x, \omega + \lambda}{\sum_j \omega + \lambda D} \tag{4}$$

In this case, λ will be 1, since we are dealing with Laplace smoothing, and D is the number of different diagrams provided.

So what is happening is that each trigram has a probability of being associated with a given language. So, in our case, every single trigram (even the ones that aren´t contemplated in the dataset) have a probability of being of the language PT, ES, FR and EN. With that we can then predict a language for each sentence.

- In the first sentence, we got the language right, even though the score of the certainty of the NB classifier isn´t high. In this case, we got a classification margin of roughly 34.1%, because the 2 highest probability language (PT) has a lot of trigrams with probabilities somewhat close to the ES trigrams. Even if the recognized language was the corect one, the classifier wasn´t 100% confident it was the correct answer.

- In the next 4 sentences, we achieve great results, with the NB classifier being correct and sure of the answer it was giving, having almost 100% scores and classification margins.

- In the last sentence, the classifier actually gets the recognized language wrong. Yet again, we have a somewhat good score (around 79.3%) but we don´t have a great classification margin (around 58.6%). The classifier was "in doubt" between FR and PT, which is actually the correct language. Even with the classifier being more confident on it´s answer this time around than the 1º sentence it analized, it doesn´t necesarily mean he got it right. This might have happened because the amount of trigrams analized in the last sentence that weren´t provided in the dataset was high, and the result depicted is gonna be less accurate since a lot of the conditional distribution of the probabilities of the trigrams was purely done by Laplace smoothing guessing the probabilities of each language for each of those trigrams.