1. Naive Bayes Classifier.

O Bayes classifier é o classificador ideal que escolhe uma classe observando qual das classes, tem é mais provável ocorrer para certos parâmetros, ou seja tem conhecimento à posteriori pois sabe a distribuição probabilística das classes em antemão (MAP). Este classificador baseia-se no teorema de Bayes para definir qual a classe e este é dado pela seguinte equação:

Onde corresponde à distribuição à posteriori das classes. É importante referir que no Bayes Classifier o é apenas um fator de escala pois é comum para todos os e por isso não necessita de ser calculado (a menos que se queiram comparar vários vetores de features). O Naive Bayes Classifier é em tudo igual ao Bayes Classifier mas com uma diferença importante, não são sabidas as distribuições de probabilidade de dado , e por isso é pressuposto que as features são condicionalmente independentes (Covariancia=0) ou seja c

Apartir desta hipótese, seria de esperar que o Naive Bayes Classifier não obtivesse bons resultados no entanto não é isso que ocorre como se pode observar ao longo deste laboratório

4.

Confirmar probabilidades iguais 50 em cada classe de 150 elementos

Como se pode observar o erro que foi obtido é inferior a 1.5% ; Podemos confirmar que a aproximação que foi feita de as distribuições das classes se a próxima uma distribuição de gaussiana; Naive assume independência entre as features com esta suposição no calculo das matrizes da variância e pondo o valor da covariância a 0, resultando a suposição de serem independentes vão ser feitas 7 predições erradas, enquanto que quando foram calculados os valores da covariância os erro de predições foram apenas 2.

Análise preliminar das frases:

“**O curso dura cinco anos**”:

A linguagem reconhecido nesta frase é a correta embora o valor da margem de classificação seja baixo, sendo que a classe que obteu a classificação mais próxima foi a linguagem espanhola, esta semelhança tem a sua origem em que as duas línguas têm raízes em comum muito perto entre si, por isso para que seja possível distinguir entre si é preciso que existam ocorrências de palavras exclusivas de língua para língua para que se consigam “afastar” as classificações da frase, os valores de classificação também se vão distancia consoante o tamanho do texto a analisar.Neste caso a maioria dos trigramas são comuns nas duas linguagens, sendo difícil distinguir entre o português e o espanhol.

“**El mercado está muy lejos.**” :

Neste caso a linguagem reconhecida pelo classificardor corresponde à linguagem real, e o valor da margem de classificador é consideravelmente elevado ou seja com este texto o classificador tinha informação para “afastar a concorrência” sendo que o texto a analisar possui característica comuns na linguagem espanhol que são raros nas outras linguagens, fazendo diminuir a pontuação da outras linguagens quando analisaram trigramas característicos da linguagem espanhola:

Como caso de estudo, por exemplo os trigrama muy, nos ficheiros utilizados para o treino do classificador aparece 355233 vezes no espanhol, 3103 no francês, 1218 no português e 10 no inglês. Não tomar estes resutados como uma representação total do texto a ser analisado mas consegue-se encontrar uma preferência evidente com a língua espanhola.

“**Eu vou à loja.**”:

Aqui a linguagem reconhecida não corresponde à real sendo que o classificador considerou que estava perante um texto na língua francesa quando na realidade se encontrava perante a portuguesa. No entanto podemos ver que a língua com melhor cotação a seguir à francesa foi a portuguesa, no entanto o valor da margem de classificação não é baixo. Sendo que neste caso estas duas línguas também tem origem no latim as suas raízes também são comuns sendo que existem palavras/(silabas). Parte da possibilidade de o classificardor tenho errado pode vir do caso do facto de o texto a ser analisado ser relativamente pequeno, este conjunto de condições causa que o classificador não irá conseguir encontrar trigramas comuns numa linguagem que seja raro noutra, sendo que após várias ocorrências da situação agora definida o classificador irá isolar o candidato mais adequado ao texto a ser analisado.

“**the word é is very short**”:

Certo

Se não pusermos os ponto final considera português dizer que se este for o caso, não é comum palavras em portgues que acabem e t e que o trigramam “ é ” é muito comum no português 1966239 e não existe no inglês. Sendo que o score do português sem deste trigrama “é i” = 55737 e

I’m bored gonna stop now

SRY WILL COMEBACK LATER