Trabajo Práctico 6 – Inteligencia Artificial

Filtro de Spam

Pedro Tamargo Allué (758267)

Comparación Naive Bayes en función de la distribución

Para la realización de esta comparación se han realizado pruebas con las distribuciones Bernoulli y Multinomial, ambas aplicando suavizado de Laplace.

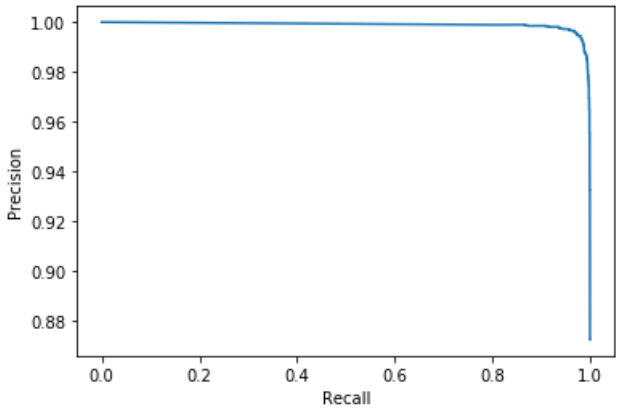
Para la implementación de las pruebas con la distribución Bernoulli se ha utilizado la clase BernoulliNB, implementada en el paquete sklearn.naive_bayes. Los resultados de la misma en tasa de acierto y f1 son:

```
Bernoulli Naive Bayes
BEST SIZE (Laplace): 0.1
BEST ACCURACY: 0.9868667799616745
BEST F1 SCORE: 0.9857640932603907
```

Captura de pantalla de los resultados del entrenamiento de la red bayesiana con la distribución Bernoulli.

```
[[1423   77]
 [  24 4476]]
```

Matriz de confusión para la Red Bayesiana con distribución Bernoulli.



Curva Precisión – Recall de la Red Bayesiana con distribución Bernoulli.

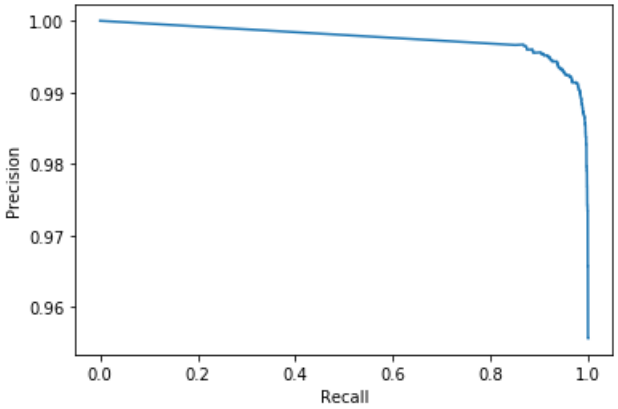
Para las pruebas con la distribución Multinomial se ha realizado un proceso análogo al de la distribución Bernoulli, utilizando la clase MultinomialNB del paquete nombrado anteriormente. Los resultados de la ejecución de la misma son:

```
Multinomial Naive Bayes
BEST SIZE (Laplace): 0.1
BEST ACCURACY: 0.9884903973890082
BEST F1 SCORE: 0.9874496502960289
```

Captura de pantalla de los resultados del entrenamiento de la Red Bayesiana con la distribución Multinomial.

```
[[1438   62]
 [  31 4469]]
```

Matriz de confusión para la Red Bayesiana con distribución Multinomial.



Curva Precisión – Recall para la Red Bayesiana con una distribución Multinomial.

Podemos concluir que, para ambas distribuciones se ha utilizado el valor 0.1 para el parámetro del suavizado de Laplace, y que ambas proveen buen resultado, un 98% de tasa de acierto en ambos casos, siendo la de la distribución Multinomial mínimamente superior.

Comparación Naive Bayes en función del parámetro del suavizado de Laplace

El suavizado de Laplace es un parámetro utilizado para dar cierta estabilidad estadística al comportamiento del estimador. En este caso, se utiliza para que cuando aparezcan palabras que no han sido vistas anteriormente, es decir, no tienen una probabilidad asociada, su probabilidad no sea 0 y pueda ocasionar fallos en la predicción.

En este caso, para la realización de las comparaciones se van a muestrear, con las dos distribuciones estudiadas anteriormente, los distintos resultados para distintos valores del parámetro del suavizado de Laplace.

	Bernoulli		Multinomial	
Valor Alpha	Accuracy	F1	Accuracy	F1
0	0,9849	0,9836	0,9853	0,984
0,01	0,9872	0,9861	0,9882	0,9871
0,1	0,9866	0,9855	0,9883	0,9872
0,3	0,986	0,9849	0,9878	0,9866
0,6	0,9858	0,9845	0,9871	0,986
0,9	0,9856	0,9843	0,9869	0,9857
1	0,9851	0,9838	0,987	0,9859
3	0,9842	0,9826	0,9856	0,9842
5	0,9675	0,9633	0,9837	0,9821
7	0,9342	0,9225	0,9805	0,9785
9	0,891	0,8648	0,9775	0,9728
15	0,7722	0,6686	0,9538	0,9472
20	0,6937	0,4964	0,9356	0,9248
35	0,5867	0,1753	0,8894	0,8629
50	0,5633	0,086	0,8558	0,8133
75	0,5468	0,0173	0,816	0,7485
100	0,5429	0,0006	0,7859	0,6948

Tabla comparativa variando el valor del parámetro de suavizado de Laplace.

A la vista de los resultados de la tabla anterior, podemos observar que, con $\alpha \leq 1$ se obtienen valores de accuracy y de F1 bastante altos, mientras que con $\alpha > 1$ obtenemos valores peores, empeorando más cuanto mayor es α. En el caso de usar la distribución Bernoulli los resultados empeoran hasta obtener un valor casi nulo en F1.

```
[[1500    0]
 [4463   37]]
```

Matriz de confusión de una Red Bayesiana con distribución Bernoulli con $\alpha = 100$.

Podemos observar que, la matriz de confusión anterior muestra una gran cantidad de falsos negativos.

```
[[1492    8]
 [1943 2557]]
```

Matriz de confusión de una Red Bayesiana con distribución Multinomial con $\alpha = 100$.

En este caso observamos que también existen falsos negativos, pero, a diferencia del caso anterior, se han identificado correctamente más elementos.

Conclusiones

Para la elección del mejor clasificador se ha utilizado la métrica F1 score debido a que se ha buscado el mejor equilibrio entre precisión y recall.

Para la evaluación, se han considerado los siguientes clasificadores: MultinomialNB($\alpha = 0.1$), BernoulliNB($\alpha = 0.01$).

```
Bernoulli Naive Bayes
Laplace value: 0.01
ACCURACY: 0.987227556355285
F1 SCORE: 0.986153637471564
```

```
Multinomial Naive Bayes
Laplace value: 0.1
ACCURACY: 0.9882379046775582
F1 SCORE: 0.9871857857773534
```

Comparación de los resultados de los modelos a evaluar.

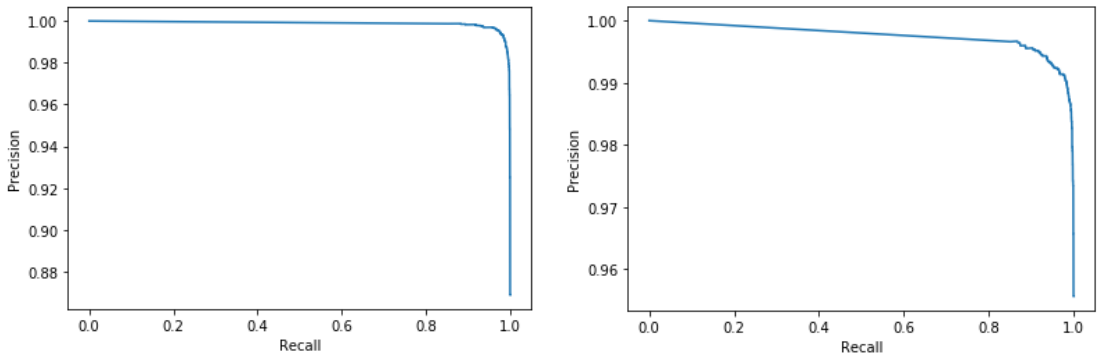
Se puede observar que ambos modelos muestran unos valores bastante altos en F1 score, siendo los resultados del modelo MultinomialNB($\alpha = 0.1$) mínimamente superiores.

```
[[1428   72]
 [  26 4474]]
```

```
[[1438   62]
 [  31 4469]]
```

Matrices de confusión de ambos modelos, a la izquierda BernoulliNB($\alpha = 0.01$), derecha MultinomialNB($\alpha = 0.1$).

Comparando las matrices de confusión de ambos modelos, podemos observar que los resultados son similares, pero que el modelo de Bernoulli muestra más falsos positivos que el modelo Multinomial, el cual muestra más falsos negativos.

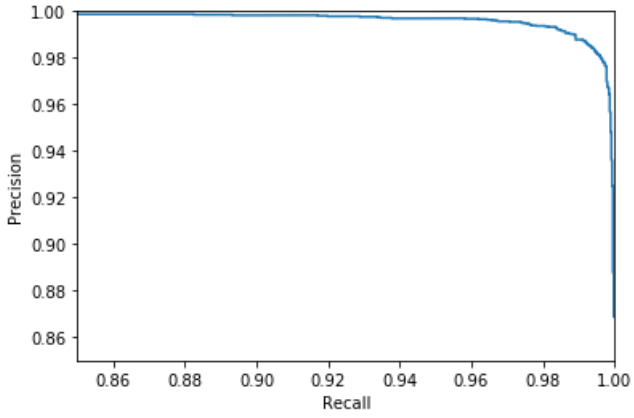


Curvas Precisión-Recall de ambos modelos, a la izquierda BernoulliNB($\alpha = 0.01$), derecha MultinomialNB($\alpha = 0.1$).

Podemos observar que la curva del modelo Bernoulli se aproxima más a la esquina superior derecha, lo que significa que tiene una precisión y un recall altos.

Por lo tanto, en vista de los resultados anteriores, se ha elegido como el mejor clasificador el Bernoulli con suavizado $\alpha = 0.1$.

En términos del umbral de decisión adecuado, se ha seleccionado el intervalo $[0.94, 0.975]$ debido a que en la siguiente gráfica se puede apreciar que existe un equilibrio entre la precisión y el recall, priorizando la precisión, ya que, con más precisión menos falsos positivos aparecerán.



Curva Ampliada del modelo Bernoulli en el intervalo [0.85,1] en el eje de Recall.

Se puede observar que en el intervalo elegido anteriormente existe una reducción de la precisión conforme aumenta el recall, pero siendo el valor de la precisión superior al 98% en todo el intervalo.