Práctica 2:

Clasificación de textos

Ramon Ruiz Dolz
Salvador Marti Roman
3CO21

Error del clasificador multinomial en función del hiperparámetro épsilon

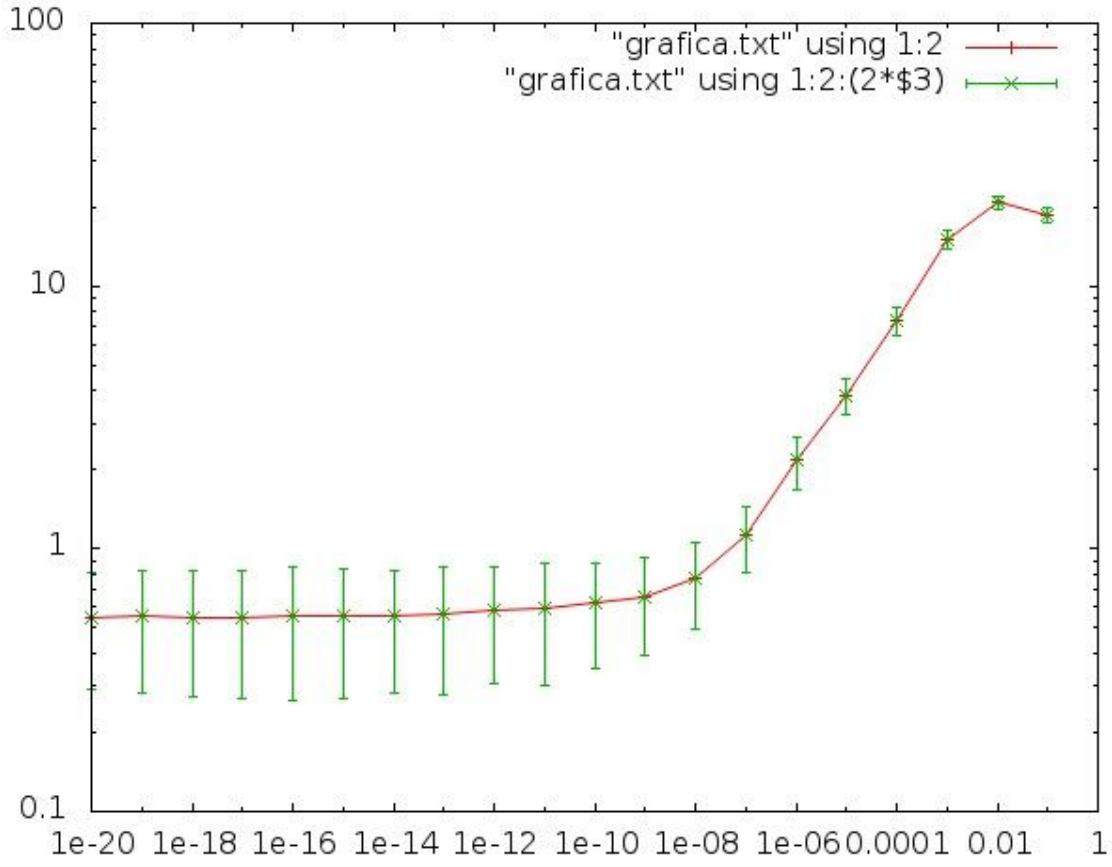

Tabla de datos:

Épsilon	%error	desviación típica
0.10000	18.625	0.60233
0.010000	20.803	0.64592
0.0010000	15.109	0.62753
1.0000e-04	7.3733	0.46511
1.0000e-05	3.8321	0.30271
1.0000e-06	2.1676	0.24661
1.0000e-07	1.1279	0.16002
1.0000e-08	0.77276	0.14051
1.0000e-09	0.65909	0.13388

1.0000e-10	0.61856	0.13292
1.0000e-11	0.59036	0.14401
1.0000e-12	0.57891	0.13680
1.0000e-13	0.56833	0.14557
1.0000e-14	0.55511	0.13603
1.0000e-15	0.55247	0.14233
1.0000e-16	0.55511	0.14561
1.0000e-17	0.54718	0.13852
1.0000e-18	0.55071	0.13816
1.0000e-19	0.55247	0.13539
1.0000e-20	0.54983	0.12976

Como podemos observar, tras la experimentación hemos obtenido el % de error y la desviación típica de la media de 30 mezclas diferentes del set de muestras trec06p con tal de obtener así valores más precisos. Primero tenemos un bucle más externo que se encarga de ir modificando el valor de epsilon, empezando en 0.1 y terminando en 10^-20 reduciendo su valor en 10^-1 en cada iteración. Para cada valor de epsilon se realizan otras 30 iteraciones para el mismo set de datos barajando los correos cada vez consiguiendo así obtener diferentes muestras de entrenamiento y test. Con cada barajado tomamos el error de clasificación tras realizar un suavizado de Laplace y vamos anotando los valores para calcular así la media y la desviación típica. El valor de épsilon para realizar el suavizado, como se puede observar en la gráfica, a medida que es más pequeño menor error alcanzamos hasta tender a 0.5. El intervalo de confianza, sin embargo a medida que nos aproximamos a 0.5 tiene más impacto proporcional.