Universidad Zaragoza

Memoria Inteligencia Artificial

Jorge Sanz Alcaine 680182

Introducción

El objetivo del trabajo es desarrollar un filtro de spam en python mediante un clasificador de bayes ingenuo utilizando los datos de datos enron para ajustar el clasificador y probarlo.

La base de datos Enron es de dominio público y contiene miles de mail clasificados como spam o ham.

División de los datos

Los mails obtenidos de la base de datos se dividen en tres grupos, entrenamiento, validación y test. Los datos de entrenamiento, como el propio nombre dice, son los que se utilizan para entrenar el sistema, los datos de validación sirven para calcular el mejor valor del suavizado de Laplace y los datos de test sirven para comprobar la efectividad del clasificador con datos que no se han usado en el entrenamiento con el mejor valor del suavizado de Laplace.

Para que el sistema se ajuste bien a los datos es necesario seguir una proporción entre los tres grupos. Los datos de entrenamiento deben ser el 80% del total, los datos de validación el 10% y los de test el 10% restante.

Validación cruzada

Es posible que al escoger los datos de validación, los datos presenten propiedades distintas a los datos de entrenamiento. Para evitar estas diferencias se utiliza la validación cruzada.

La validación cruzada consiste en utilizar datos de validación en diferentes posiciones para minimizar las diferencias entre los datos de validación y entrenamiento.

V	Ent	Test

Ent	V	Ent	Test

Ent	V	Ent	Test

Ent	V	Ent	Test

Ent	V	Ent	Test

En este filtro el número de combinaciones en la validación cruzada que se ha utilizado es 5.

Bolsas de palabras

Para poder predecir si un mail es spam o ham es necesario saber cuales son las palabras mas comunes para cada uno. Por ello, el sistema debe aprender el vocabulario en los datos de entrenamiento para poder construir bolsas de palabras.

Las bolsas de palabras son matrices en las que cada fila representa una palabra del vocabulario previamente aprendido y cada columna un mail. El elemento correspondiente a un una fila y una columna representa el número de veces que ha aparecido esa palabra en ese mail. Esta matriz contiene un gran número de elementos iguales a 0, por lo que para ahorrar espacio se representa mediante una matriz escalonada.

Suavizado de Laplace

El suavizado de Laplace en un factor que sirve para que datos nuevos con palabras que no aparezcan en alguna de las categorías en los datos de entrenamiento no se clasifiquen directamente como la otra opción. También es posible que aparezcan palabras que no hallan aparecido en ninguna de las categorías y si no fuera por el suavizado podrían darse indeterminaciones.

El suavizado de Laplace se calcula con los datos de validación, sin embargo, debido a la gran cantidad de mails en el entrenamiento apenas aparecen palabras que no pertenezcan a ninguna de las categorías y el mejor suavizado de Laplace es 1, salvo en raras ocasiones en donde el mejor es 2.

Multinomial y Bernoulli

Se han utilizado dos clasificadores Multinomial y Bernoulli. El clasificador recibe como parámetro el suavizado de Laplace que se va a usar y se entrena con una bolsa de palabras de los datos de entrenamiento. A partir de una bolsa de palabras con un vocabulario idéntico a con el que ha sido entrenado es capaz de predecir la categoría de cada uno de sus mails.

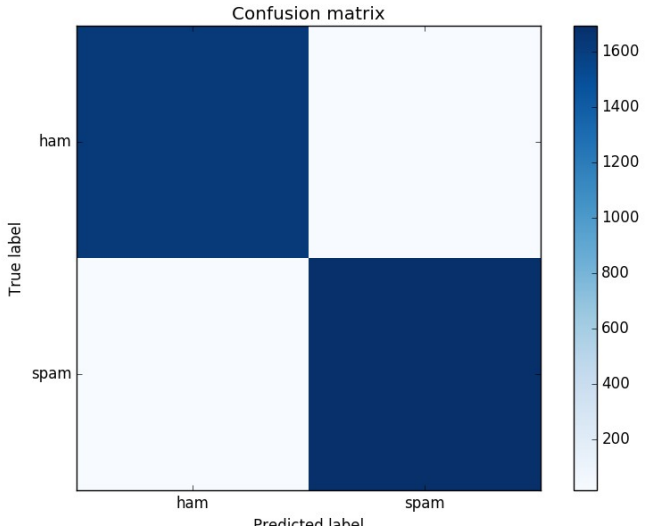
Medidas empleadas

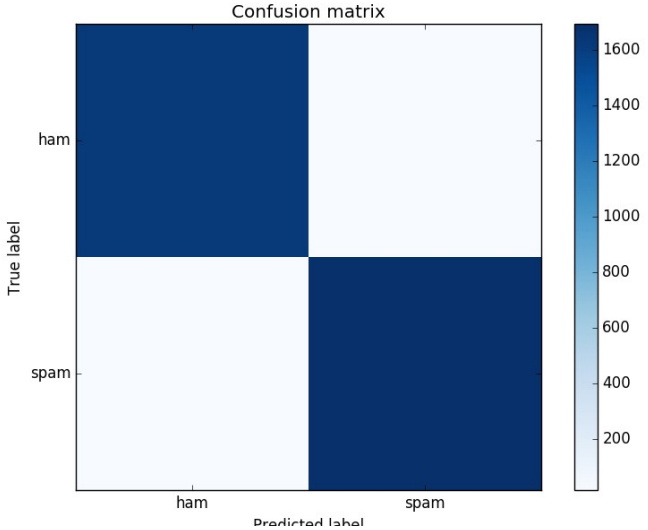
Para calcular el mejor valor en el suavizado de Laplace es necesario compara su predicción en los datos de validación con sus etiquetas verdaderas. Es posible sacar un porcentaje de éxitos al contar el número de aciertos y dividirlo por el total.

Multinomial	Bernoulli
0.9666	0.9681

Matriz de confusión

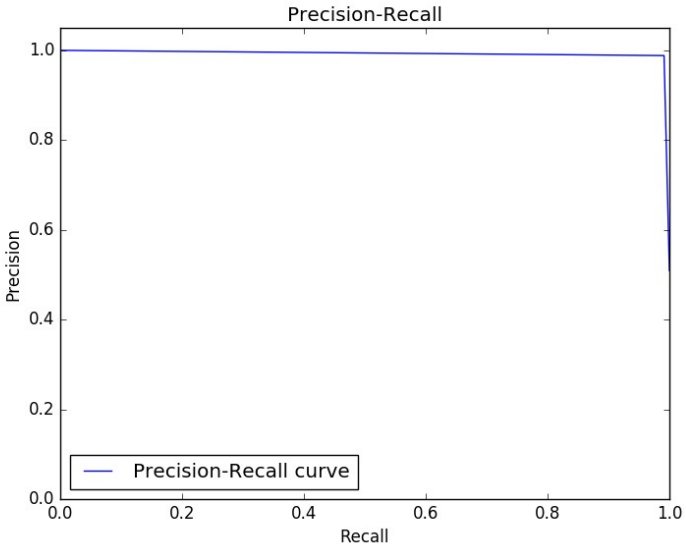
Multinomial	Bernoulli

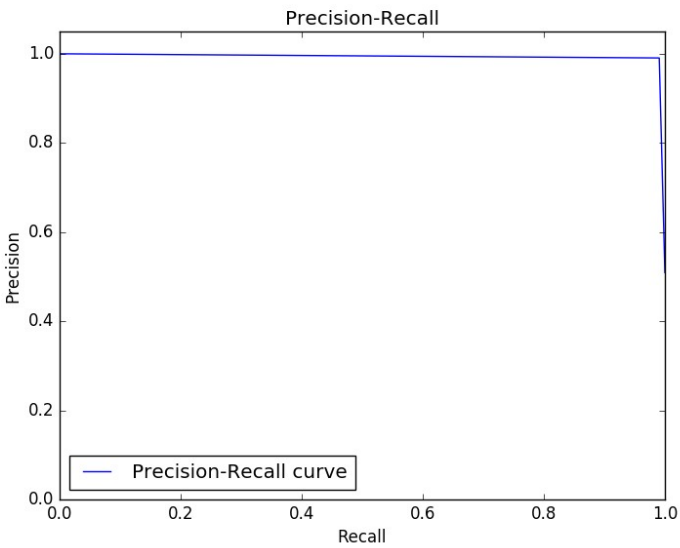




Curva Precision-Recall

Multinomial	Bernoulli





F1-score

Multinomial	Bernoulli
0.9898	0.9906