UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
-Universidad del Perú, La Decana de América-
Facultad de Ingeniería de Sistemas e Informática

TRABAJO COMPUTACIONAL

IMPLEMENTACIÓN DE UN FILTRO SPAM MEDIANTE TÉCNICA BAYES INGENUO

Bryan Gama Solórzano
Facultad de Ingeniería de Sistemas e Informática
Universidad Nacional Mayor de San Marcos
Lima, Perú
bryan.gama@unmsm.com

Robert Silva
Facultad de Ingeniería de Sistemas e Informática
Universidad Nacional Mayor de San Marcos
Lima, Perú
robert.silva @unmsm.com

RESUMEN

Los contenidos de spam en correos electrónicos son considerados como distractores y molestos para los usuarios. Este problema es foco de atención para grandes empresas de servicios de correos, es por ello que como objetivo de este estudio, tenemos la implementación de una técnica basada en el teorema de Bayes, la técnica Bayes Ingenuo; un tipo de red Bayesiana, usado normalmente para problemas de clasificación, en la que se asume que las variables consideradas para la clasificación son independientes. Además realizaremos la evaluación de la funcionalidad del sistema, así como un análisis de resultados que nos dirán qué tan adecuada es esta técnica para la solución de este problema. Empezaremos definiendo el dataset, un conjunto de más de 5000 mensajes en idioma inglés, el cual dividiremos en dos: un grupo de mensajes (75%) serán de entrenamiento y el resto de mensajes (25%) serán de prueba. Realizamos modificaciones sobre el teorema de Bayes original, teniendo así una distribución multinomial para las variables aleatorias consideradas (en este caso, que en un mensaje sea spam o ham, dada una palabra) la cual, al ser llevada al espacio logarítmico, alcanza aún mejores resultados. Veremos que el clasificador implementado alcanzó un 98% de aciertos para la data de prueba, mostrando así que esta técnica es adecuada para la solución de este problema.

INDICE

Contenido

INTRODUCCIÓN....3
Planteamiento del problema....3
Justificación....3
Limitaciones....3
Objetivos....3
MARCO TEÓRICO....4
PROBABILIDAD CONDICIONAL....4
INDEPENDENCIA CONDICIONAL....4
TEOREMA DE BAYES....4
Aplicaciones....5
REDES BAYESIANAS....5
CLASIFICADOR BAYESIANO INGENUO....6
Concepto probabilístico....7
TIPOS DE CLASIFICADORES BAYES INGENUOS SEGÚN LA DISTRIBUCIÓN DE LA PROBABILIDAD DE SUS VARIABLES CLASIFICADORAS....9
MÉTODOS....10
UN PEQUEÑO EJEMPLO....10
CLASIFICADOR SPAM....13
EJEMPLO DE FUNCIONAMIENTO....14
Entrenamiento....15
Obtendremos así la siguiente tabla:....16
Analizando la entrada....16
EXPERIMENTOS COMPUTACIONALES....18
Conclusiones....18
REFERENCIAS....19

INTRODUCCIÓN

Este estudio tiene como objetivo resolver uno de los problemas más comunes para las empresas que brindan servicios de correos electrónicos, nos referimos a la detección de spam. Muchos algoritmos tratan de responder a la siguiente pregunta: Dado un mensaje, ¿es este spam o ham(spam)? Presentaremos la ténica Bayes ingenuo y sus fundamentos para resolver este problema en específico. Esta técnica además muestra una ventaja en la detección de contenido spam frente a otros, como el método de los K vecinos más cercanos (K-nearest neighbors, k-nn), con un 81% de precisión, frente a un 71% del método k-nn.

Planteamiento del problema

El comunicarnos de manera efectiva es algo esencial y para ello usamos frecuentemente los correos electrónicos, pero algunas veces recibimos mensajes no deseados ya sean ofertas, publicidad, entre otros, que causan distracción o molestia debido a que los correos esperados pueden perderse entre los mensajes no deseados. Además de esto, el spam también puede ser enviado para realizar un ataque a una organización, pues pueden saturar la red de esta, haciéndola vulnerable.

Justificación

El presente trabajo se enfocará en mostrar los fundamentos de la técnica Bayes Ingenuo, la comparación de esta con otras redes bayesianas, mostrar un ejemplo de su funcionamiento y la implementación de un sistema que permita conocer al usuario con alta confianza si un mensaje dado es spam o ham (no spam).

Limitaciones

- El sistema implementado contó con un dataset compuesto por mensajes en inglés, por ende, este solo podrá responder con alta confianza si el mensaje que se le presente está también en este idioma.
- El sistema no detecta las palabras cuando una de las letras que la conforman es reemplazada a modo de ocultar el mensaje.
- El sistema dará la misma respuesta en caso ninguna de las palabras contenidas en el mensaje dado, se encuentren en el diccionario: ham.

Objetivos

- Desarrollar un clasificador de mensajes basado en el teorema de Bayes.
- Demostrar que la técnica Bayes ingenuo resuelve el problema de clasificación de mensajes con una alta confianza.

MARCO TEÓRICO

PROBABILIDAD CONDICIONAL

Es la probabilidad de que ocurra un evento A, sabiendo que también sucede otro evento B. La probabilidad condicional se escribe P(A|B) o P(A/B), y se lee «la probabilidad de A dado B».

INDEPENDENCIA CONDICIONAL

Se dice que dos variables aleatorias A y B son independientes condicionalmente, si existe una tercera C con la que se cumple:

$$p(A,B|C)=p(A|C)*p(B\vee C)$$

Adicionalmente:

$$p(A|C,B)=p(A|C)$$

Esto nos dice que las variables A y B son independientes siempre y cuando haya ocurrido el evento C.

Nota:

No confundir con la independencia general de variables; en esta última se habla de que las variables A y B son independientes entre sí; es decir, no existe la necesidad de que ocurra un evento C, pues la ocurrencia del evento B no afecta absolutamente la ocurrencia del evento A, ni viceversa. En la independencia general se cumple:

$$p(A\vee B)=p(A)$$

TEOREMA DE BAYES

El teorema de Bayes, en la teoría de la probabilidad, expresa la probabilidad condicional de un evento aleatorio A dado B en términos de la distribución de probabilidad condicional del evento B dado A y la distribución de probabilidad marginal de solo A.

En términos más generales y menos matemáticos, el teorema de Bayes es de enorme relevancia puesto que vincula la probabilidad de A dado B con la probabilidad de B dado A. Es decir, por ejemplo, que sabiendo la probabilidad de tener un dolor de cabeza dado que se tiene gripe, se podría saber (si se tiene algún dato más), la probabilidad de tener gripe si se tiene un dolor de cabeza. Muestra este sencillo ejemplo la alta relevancia del teorema en cuestión para la ciencia en todas sus ramas, puesto que tiene vinculación íntima con la comprensión de la probabilidad de aspectos causales dados los efectos observados.

Sea $\{A_1, A_2, .., A_i, \ldots, A_n\}$ un conjunto de sucesos mutuamente excluyentes y exhaustivos, y tales que la probabilidad de cada uno de ellos es distinta de cero (0). Sea B un suceso cualquiera del que se conocen las probabilidades condicionales $P(B\vee A_i)$. Entonces, la probabilidad $P(A_i\vee B)$viene dada por la expresión:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

Donde:

- $P\left(A_i\right)$Son las probabilidades a priori
- $P\left(B \vee A_i\right)$Es la probabilidad B en la hipótesis A_i
- $P\left(A_i \vee B\right)$ Son las probabilidades a posteriori

Aplicaciones

El teorema de Bayes es válido en todas las aplicaciones de la teoría de la probabilidad. Sin embargo, hay una controversia sobre el tipo de probabilidades que emplea. En esencia, los seguidores de la estadística tradicional solo admiten probabilidades basadas en experimentos repetibles y que tengan una confirmación empírica mientras que los llamados estadísticos bayesianos permiten probabilidades subjetivas. El teorema puede servir entonces para indicar cómo debemos modificar nuestras probabilidades subjetivas cuando recibimos información adicional de un experimento. La estadística bayesiana está demostrando su utilidad en ciertas estimaciones basadas en el conocimiento subjetivo a priori y el hecho de permitir revisar esas estimaciones en función de la evidencia empírica es lo que está abriendo nuevas formas de hacer conocimiento. Una aplicación de esto son los clasificadores bayesianos que son frecuentemente usados en implementaciones de filtros de correo basura o spam, que se adaptan con el uso. Otra aplicación se encuentra en la fusión de datos, combinando información expresada en términos de densidad de probabilidad proveniente de distintos sensores.

REDES BAYESIANAS

Una red bayesiana, red de Bayes, red de creencia, modelo bayesiano (de Bayes) o modelo probabilístico en un grafo acíclico dirigido es un modelo grafo probabilístico (un tipo de modelo estático) que representa un conjunto de variables aleatorias y sus dependencias condicionales a través de un grafo acíclico dirigido (DAG por sus siglas en inglés). Por ejemplo, una red bayesiana puede representar las relaciones probabilísticas entre enfermedades y síntomas. Dados los síntomas, la red puede ser usada para computar la probabilidad de la presencia de varias enfermedades.

Formalmente, las redes bayesianas son grafos dirigidos acíclicos cuyos nodos representan variables aleatorias en el sentido de Bayes: las mismas pueden ser cantidades observables, variables latentes, parámetros desconocidos o hipótesis. Las aristas representan dependencias condicionales; los nodos que no se encuentran conectados representan variables las cuales son condicionalmente independientes de las otras.

Cada nodo tiene asociado una función de probabilidad que toma como entrada un conjunto particular de valores de las variables padres del nodo y devuelve la probabilidad de la variable representada por el nodo. Por ejemplo, si por padres son m

variables booleanas entonces la función de probabilidad puede ser representada por una tabla de 2^m, una entrada para cada una de las 2^mposibles combinaciones de los padres siendo verdadero o falso. Ideas similares pueden ser aplicadas a grafos no dirigidos, y posiblemente cíclicos; como son las llamadas redes de Markov.

Existen algoritmos eficientes que llevan a cabo la inferencia y el aprendizaje en redes bayesianas. Las redes bayesianas que modelan secuencias de variables (ej. señales del habla o secuencias de proteínas) son llamadas redes bayesianas dinámicas. Las generalizaciones de las redes bayesianas que pueden representar y resolver problemas de decisión bajo incertidumbre son llamados diagramas de influencia.

CLASIFICADOR BAYESIANO INGENUO

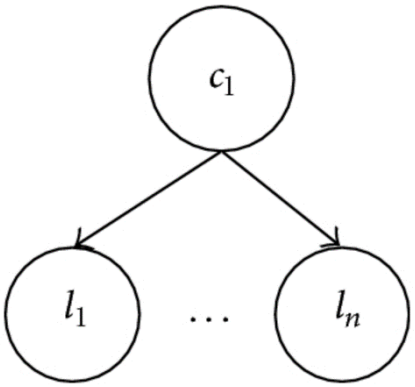
En teoría de la probabilidad y minería de datos, un clasificador Bayesiano *ingenuo* es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales. Es a causa de estas simplificaciones, que se suelen resumir en la hipótesis de independencia condicional entre las variables predictoras, que recibe el apelativo de *ingenuo*.

En términos simples, un clasificador de Bayes ingenuo asume que la presencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra característica, dada la clase variable. Por ejemplo, una fruta puede ser considerada como una manzana si es roja, redonda y de alrededor de 7 cm de diámetro. Un clasificador de Bayes ingenuo considera que cada una de estas características contribuye de manera independiente a la probabilidad de que esta fruta sea una manzana, independientemente de la presencia o ausencia de las otras características.

Para otros modelos de probabilidad, los clasificadores de Bayes ingenuo se pueden entrenar de manera muy eficiente en un entorno de aprendizaje supervisado. En muchas aplicaciones prácticas, la estimación de parámetros para los modelos Bayes ingenuo utiliza el método de máxima verosimilitud, en otras palabras, se puede trabajar con el modelo ingenuo de Bayes sin aceptar probabilidad bayesiana o cualquiera de los métodos bayesianos.

Una ventaja del clasificador de Bayes ingenuo es que solo se requiere una pequeña cantidad de datos de entrenamiento para estimar los parámetros (las medias y las varianzas de las variables) necesarias para la clasificación. Como las variables independientes se asumen, solo es necesario determinar las varianzas de las variables de cada clase y no toda la matriz de covarianza.

Representación Grafica



$Donde:$

$C_1 \in C = \{Conjunto\ de\ Clases\}$

$l_1 \in l = \{Conjunto\ de\ Atributos\}$

CONCEPTO PROBABILÍSTICO

En abstracto, el modelo de probabilidad para un clasificador es $p(C \vee F_{1,\dots}, F_n)$ sobre una variable dependiente$C$, con un pequeño número de resultados (o clases). Esta variable está condicionada por varias variables independientes desdeF_1a F_n(las frecuencias de las palabras en el diccionario presentes en el mensaje). El problema es que si el número *n* de variables independientes es grande (o cuando éstas pueden tomar muchos valores), entonces basar este modelo en tablas de probabilidad se vuelve imposible. Por lo tanto el modelo se reformula para hacerlo más manejable:

Usando el teorema de Bayes se escribe:

$$p(C|F_{1,\dots}, F_n) = \frac{p(C) * p(F_1, F_2, \dots, F_n|C)}{p(F_1, F_2, \dots, F_n)} \dots (a)$$

Lo anterior podría reescribirse en lenguaje común como:

$$Posterior = \frac{Anterior * Probabilidad}{Evidencia}$$

En la práctica sólo importa el numerador, ya que el denominador no depende de C y los valores de $F_i$son datos, por lo que el denominador es, en la práctica, constante.

El numerador puede ser reescrito como sigue, aplicando repetidamente la definición de probabilidad condicional:

$$p(C) * p(F_1, F_2, \dots, F_n|C)$$

$$\begin{aligned} &= p(C)\ p(F_1|C)\ p(F_2, \dots, F_n|C, F_1) \\ &= p(C)\ p(F_1|C)\ p(F_2|C, F_1)\ p(F_3, \dots, F_n|C, F_1, F_2) \\ &= p(C)\ p(F_1|C)\ p(F_2|C, F_1)\ p(F_3|C, F_1, F_2)\ p(F_4, \dots, F_n|C, F_1, F_2, F_3) \end{aligned}$$

Y así sucesivamente. Ahora es cuando el supuesto "naïve" de independencia condicional entra en juego: se asume que cada $F_i$es independiente de cualquier otra F_j para $j \neq i$ cuando están condicionadas a C. Esto significa que:

$$p(F_i \vee C, F_j) = p(F_i \vee C)$$

Por lo que el numerador puede expresarse como:

$$p(C)\prod_{i=1}^{n} p(F_i|C)\ldots(b)$$

Esto significa que, haciendo estos supuestos, la distribución condicional C sobre las variables clasificatorias puede expresarse de la siguiente manera:

$$p(C|F_{1,\ldots}, F_n) = \frac{1}{Z} p(C)\prod_{i=1}^{n} p(F_i|C)$$

Donde Z es un factor que depende sólo de $F_{1,\ldots}, F_n$, es decir, constante si los valores de F_i son conocidos. Así, nuestra función quedará de la siguiente manera:

$$p(C|F_{1,\ldots}, F_n) = p(C)\prod_{i=1}^{n} p(F_i|C)\ldots(c)$$

Si nos damos cuenta, veremos que, si una palabra se repite más de una vez en el mensaje a analizar, y llamamos a la cantidad de veces r_i, tendremos:

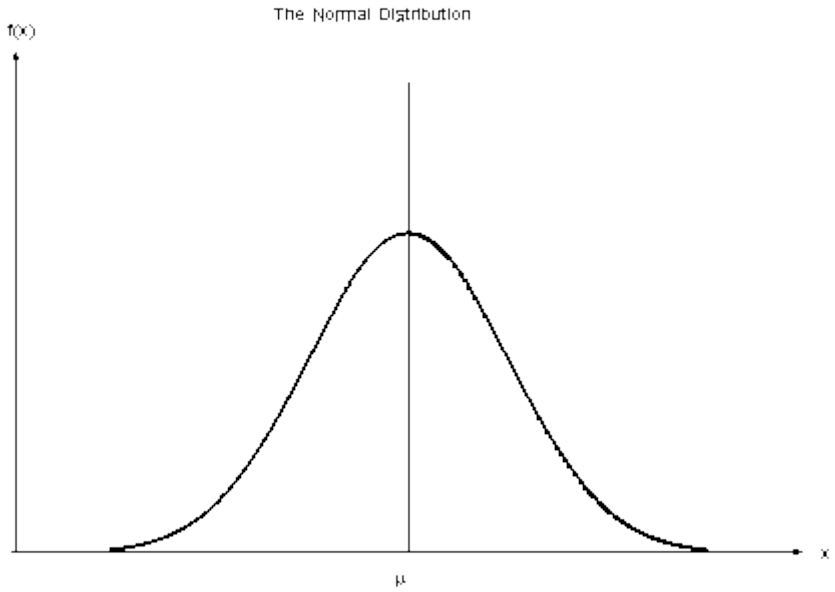
$$p(C|F_{1,\ldots}, F_n) = p(C)\prod_{i=1}^{n} p(F_i|C)^{r_i}\ldots(c)$$

Esta es la característica de una distribución multinomial. Veremos que existen otros tipos de clasificadores Bayes ingenuo que reciben el nombre de la distribución de probabilidad de las variables clasificadoras.

TIPOS DE CLASIFICADORES BAYES INGENUOS SEGÚN LA DISTRIBUCIÓN DE LA PROBABILIDAD DE SUS VARIABLES CLASIFICADORAS

Clasificador Gaussiano

Cuando se trabaja con datos continuos, una suposición usual es que los valores continuos asociados con cada clase están distribuidos de acuerdo con la distribución Gaussiana.



De la suposición previamente mencionada se llega a la conclusión siguiente:

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

Clasificador Multinomial

El modelo multinomial representa las frecuencias en las que ciertos eventos han sido generados por una distribución multinomial $(p¿¿1,\ldots,p_n)¿$donde p_i, es la probabilidad de que un evento ocurra. Donde un vector $x=(x¿¿1,\ldots,x_n)¿$ es un histograma, teniendo x_i como un contador del número de veces que el evento i es observado en un evento en particular.

$$p(\mathbf{x} \mid C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i {p_{ki}}^{x_i}$$

El clasificador se vuelve un clasificador lineal al ser expresado en log-space.

$$\log p(C_k \mid \mathbf{x}) \propto \log\left(p(C_k) \prod_{i=1}^{n} {p_{ki}}^{x_i}\right)$$
$$= \log p(C_k) + \sum_{i=1}^{n} x_i \cdot \log p_{ki}$$

Clasificador de Bernoulli

En el modelo de eventos con multivariables Bernoulli, las características son valores booleanos independientes que describen entradas.

Como el modelo multinomial este modelo es popular para la clasificación de texto donde un término booleano es usado para la ocurrencia de las caracterizas en vez de sus frecuencias.

Si x_i es un booleano que expresa la ocurrencia o ausencia del término i del vocabulario, entonces la formula inicial se reformula a:

$$p(\mathbf{x} \mid C_k) = \prod_{i=1}^{n} p_{ki}^{x_i}(1 - p_{ki})^{(1-x_i)}$$

MÉTODOS

UN PEQUEÑO EJEMPLO…

Para empezar, debemos darnos cuenta de que las clases presentes en nuestro caso son **spam** y **ham**. Además, cada una de las F_i será cada una de las palabras encontradas en el mensaje. Entonces debemos encontrar la probabilidad de que el mensaje sea ham o spam, dado el conjunto de palabras encontrados en este.

Asumamos que tenemos 100 mensajes, de los cuales 80 son ham y 20 son spam. Tendríamos los siguiente:

De los cuales, 20 de los mensajes **ham** contienen la palabra “barato”. Además, 10 de los mensajes **spam** también contienen la palabra “barato”. Tendríamos lo siguiente:

Luego, encontraremos la probabilidad de que un mensaje cualquiera sea **spam** si es que contiene la palabra “barato”. Según el teorema de Bayes, tenemos:

$$P(SPAM \vee BARATO) = \frac{P(BARATO \vee SPAM) * P(SPAM)}{P(BARATO)}$$

$$P(SPAM \vee BARATO) = \frac{\frac{10}{20} * 20}{\frac{30}{100}}$$

$$P(SPAM | BARATO) = \frac{1}{3}$$

Análogamente podemos encontrar la probabilidad de que el mensaje sea **ham** teniendo la palabra “barato”. Esto nos daría 2/3.

Ahora asumamos que 10 de los mensajes **ham** contienen la palabra “comprar”. Además, 15 de los mensajes **spam** también contienen la palabra “comprar”. Tendríamos ahora lo siguiente:

Luego, encontraremos la probabilidad de que un mensaje cualquiera sea **spam** si es que contiene la palabra “comprar”. Nuevamente, según el teorema de Bayes, tenemos:

$$P(SPAM \vee COMPRAR) = \frac{P(COMPRAR \vee SPAM) * P(SPAM)}{P(COMPRAR)}$$

$$P(SPAM \vee COMPRAR) = \frac{\frac{15}{20} * 20}{\frac{100}{25}}$$

$$P(SPAM | COMPRAR) = \frac{3}{5}$$

Ahora nuestra duda es ¿cuál es la probabilidad de que un mensaje sea **spam** si contiene ambas palabras? Pues bien, nuevamente veamos el teorema de Bayes:

$$P(SPAM | BARATO, COMPRAR ¿) = \frac{P(BARATO, COMPRAR \vee SPAM) * P(SPAM)}{P(BARATO, COMPRAR)} \ldots (d)$$

Para obtener el valor de $P(BARATO, COMPRAR \vee SPAM)$, veamos la que ecuación **(b)**. Esta nos dice que:

$$P(BARATO, COMPRAR | SPAM) = P(BARATO | SPAM) * P(COMPRAR | SPAM) \ldots (e)$$

Además, de forma general, tenemos que:

$$P(B) = P(B|S) * P(C|S) * P(S) + P(B|H) * P(C|H) * P(H) \ldots (f)$$

Donde:

B: BARATO
C: COMPRAR
S: SPAM
H: HAM

Reemplazando **(e)** y **(f)** en **(d)** tendremos:

$$P(SPAM | BARATO, COMPRAR ¿) = \frac{P(B|S) * P(C|S) * P(S)}{P(B|S) * P(C|S) * P(S) + P(B|H) * P(C|H) * P(H)}$$

$$P(SPAM \vee BARATO, COMPRAR) = \frac{\frac{\frac{10}{20} * 15}{20} * 20}{100}}{\frac{\frac{\frac{30}{20} * 25}{20} * 20}{100} + \frac{\frac{\frac{20}{80} * 10}{80} * 80}{100}}$$

$$P(SPAM | BARATO, COMPRAR) = 3/25$$

Hasta aquí hemos asumido que la probabilidad de que una palabra aparezca en un mensaje se calcula como una división simple entre la cantidad de mensajes en la que esta aparece y el total de mensajes del dataset. Pero ¿qué ocurre si una palabra encontrada en el mensaje dado

no ha aparecido en el dataset? Supongamos que la palabra es “abc”, veamos cómo afecta esto al mensaje total preguntándonos ¿cuál es la probabilidad de que un mensaje sea spam teniendo las palabras “barato” y “abc”? Tenemos:

$$P(SPAM|BARATO, ABC¿=\frac{P(B|S)*P(ABC|S)*P(S)}{P(ABC|S)*P(C|S)*P(S)+P(B|H)*P(ABC|H)*P(H)}$$

$$P(SPAM \vee BARATO, ABC)=\frac{\frac{\frac{10}{20}*0}{\frac{20}{100}}*20}{\frac{\frac{0}{20}*25}{\frac{20}{100}}*20+\frac{\frac{0}{80}*10}{\frac{80}{100}}*80}$$

$$P(SPAM|BARATO, ABC)=0/0$$

Este resultado nos indica que debemos tener una consideración para las palabras que no se encuentren, por lo que utilizaremos el suavizado de Laplace, un suavizado que nos sugiere usar una constante α pequeña que permita dar una probabilidad muy pequeña para las palabras no encontradas, pero que no anule por completo las probabilidades. Presentaremos más adelante este suavizado.

CLASIFICADOR SPAM

Lo que se pretende es implementar un sistema que pueda identificar si un texto es Spam o No Spam.

Se comenzara definendo el Data Set como un conjunto de mensajes con su respectiva clase. Se modificará el dataset teniendo algunas consideraciones. Este a su vez tendrá que ser subdividido en dos:

- Data Set de Entrenamiento
- Data Set de Testeo

El Data Set de Entrenamiento nos brindará la informacion necesaria para poder crear un diccionario o bolsa de palabras $V=\{p_1, p_2, p_3, \ldots, p_i\}$ en donde se hallaran todas las palabras, junto a su respectiva ocurrencia; que quieran ser utilizadas como caracteristicas del texto (Se recomienda que sean aquellas palabras que tengan el mayor numero de ocurrencias indifierentemente de la clase a al que pertenezcan).

En base a este diccionario se deberan definir otros dos:

$DicS=\{p_1, p_2, p_3, \ldots, p_i\}$ y $DicH=\{p_1, p_2, p_3, \ldots, p_i\}$ cada uno de estos para la clase Spam y No Spam respectivamente a su vez poseeran todos los elementos del diccionario V y la probabilidad de cada uno de estos elementos.

Un punto a aclarar es que las probabilidades que cada uno de los elementos de estos diccionarios sera proporcional a la frecuencia de su ocurrencia en la clase correspondiente. Es decir que la probabilidad de p_1 en el conjunto $DicS$ sera

directamente proporcional a su frecuencia en la clase Spam y la probabilidad de la misma palabra en el conjunto $DicNS$ sera directamente proporcional a su frecuencia en la clase No Spam; la probabilidad se vera afectada por la formula siguiente:

$$p(i \vee C) = \frac{f_{ic} + \propto}{f_C + \propto |V|} \ldots (g)$$

Donde:

$$p(i \vee C) : Probabilidad\ de\ que\ la\ palabra\ i\ aparezca\ en\ la\ clase\ C\ .$$

$$f_{ij} = Frecuencia\ de\ la\ palabra\ i\ en\ la\ clase\ C$$

$$\propto = Constante\ para\ el\ suavizado\ , usualmente\ 0.001$$

$$f_i = Suma\ de\ todas\ las\ frecuencias\ de\ las\ palabras\ dentro\ de\ la\ clase\ C$$

$$|V| = Numero\ de\ elementos\ del\ conjunto\ V$$

Teniendo en cuenta lo expuesto anteriormente y el texto o mensaje definido de la forma:

p } rsub { 1 } { p } rsub { 2 } { p } rsub { 3 } { p } rsub { 4 } … { p } rsub { i } ¿

Se debera obtener la probabilidad de su pertenencia a la clase Spam y a su vez a la clase Ham; para este fin y teniendo en cuenta el clasificador multinomial, ya que se trabajará con el numero de ocurrencias de cada una de las palabaras; se usará la fórmula:

$$p(C) * \prod_{i=1}^{|V|} p¿$$

Donde:

$$p(C) = Probabilidad\ de\ la\ Clase$$

$$|V| = Numero\ de\ elementos\ en\ el\ diccionario$$

$$p(i|j) = Probabilidad\ de\ la\ palabra\ i\ dada\ una\ clase\ j$$

$$f_i = Frecuencia\ de\ la\ palabra\ i\ en\ el\ mensaje\ dado$$

Que al ser expresado en log-space:

$$\log(p(C)) + \sum_{i=1}^{|V|} f_i * \log\ (p¿(i \vee j))¿$$

Un problema seria que si una palabra aparece más de una vez, la probabilidad calculada se eleva denasiado. Es por esto que se usa un suavizado teniendo en cuenta la funcion $\log()$.

Reformulando la ecuación tenemos:

$$\log(p(C))+\sum_{i=1}^{|V|}\log\,(1+f_i)*\log\,(p¿(i\vee j))\ldots(h)¿$$

Siendo esta la función principal de la clasificación.

El clasificador dará como respuesta la clase que obtenga el resultado mayor al evaluar la función anterior.

EJEMPLO DE FUNCIONAMIENTO

Asumiremos que tenemos un datase de 20 mensajes, de los cuales 5 son spam y el resto, ham. A partir de ello, se generó un diccionario de 3 palabras. Además, tendremos la ocurrencia de estas palabras dentro del total de mensajes considerados, además de su ocurrencia en los mensajes de cada una de las clases. Tenemos la siguiente tabla que resume estas ocurrencias.

Palabra	Ocurrencia (Total)	Ocurrencia (spam)	Ocurrencia (ham)
contiene	10	7	3
dos	4	1	3
palabra	6	4	2

Deduciremos si el mensaje **"Contiene dos veces la palabra dos"** es spam o no.

Entrenamiento

Para empezar, calcularemos la probabilidad de cada clase; esto es:

$$p(Spam)=\frac{Cantidad\,mensajes\,spam}{Cantidad\,mensajes\,total}$$

$$p(Spam)=\frac{5}{20}=\frac{1}{4}$$

Luego:

$$p(Ham)=1-\frac{1}{4}=\frac{3}{4}$$

Ahora encontraremos la probabilidad de encontrar cada una de las palabras del diccionario en cada clase, esto es $p(i\vee C), i: palabra, C: clase$. Para esto, debemos considerar la ecuación **(g)**:

$$p(i\vee C)=\frac{f_{iC}+\propto}{f_C+\propto|V|}$$

Además, utilizaremos la constante $\propto=0.001$.

- Calcularemos la probabilidad de encontrar la palabra **"contiene"** en la clase:
  - Spam

$$p(contiene\vee Spam)=\frac{f_{iC}+\propto}{f_C+\propto|V|}$$

$$p(contiene \vee Spam) = \frac{7 + 0.001}{12 + 0.001 * 3} = 0.583$$

- Ham

$$p(contiene \vee Spam) = \frac{f_{iC} + \propto}{f_C + \propto |V|}$$

$$p(contiene \vee Spam) = \frac{3 + 0.001}{8 + 0.001 * 3} = 0.375$$

- Calcularemos la probabilidad de encontrar la palabra **"dos"** en la clase:
  - Spam

$$p(dos \vee Spam) = \frac{f_{iC} + \propto}{f_C + \propto |V|}$$

$$p(dos \vee Spam) = \frac{1 + 0.001}{12 + 0.001 * 3} = 0.083$$

- Ham

$$p(dos \vee Spam) = \frac{f_{iC} + \propto}{f_C + \propto |V|}$$

$$p(dos \vee Spam) = \frac{3 + 0.001}{8 + 0.001 * 3} = 0.375$$

- Calcularemos la probabilidad de encontrar la palabra **"palabra"** en la clase:
  - Spam

$$p(palabra \vee Spam) = \frac{f_{iC} + \propto}{f_C + \propto |V|}$$

$$p(palabra \vee Spam) = \frac{4 + 0.001}{12 + 0.001 * 3} = 0.334$$

- Ham

$$p(palabra \vee Spam) = \frac{f_{iC} + \propto}{f_C + \propto |V|}$$

$$p(palabra \vee Spam) = \frac{2+0.001}{8+0.001*3} = 0.25$$

Obtendremos así la siguiente tabla:

Palabra	p(i,Spam)	p(i, Ham)
contiene	0.583	0.375
dos	0.083	0.375
palabra	0.334	0.25

Analizando la entrada

Obtendremos la frecuencia de palabras del mensaje **"Contiene dos veces la palabra dos"** que además se encuentren en el diccionario.

Palabra	f(i) en el mensaje
contiene	1
dos	2
palabra	1

Ahora haremos uso de la tabla obtenida en el entrenamiento y la ecuación **(h)**:

$$\log\left(p(C)\right)+\sum_{i=1}^{|V|}\log\,(1+f_i)*\log\,(p¿(i\vee C))¿$$

Veamos,

i) Evaluamos los datos del mensaje en la función **(h)** para la clase spam:

$$R_{spam}=\log\left(p(Spam)\right)+\sum_{i=1}^{|V|}\log\,(1+f_i)*\log\,(p¿(i\vee Spam))¿$$

$$¿\log\left(\frac{1}{4}\right)+\log(1+1)*\log(0.583)+\log(1+2)*\log(0.083)+\log(1+1)*\log(0.334)$$

$$R_{spam}=-5.255$$

i) Evaluamos en la función **(h)** para la clase spam:

$$R_{ham}=\log\left(p(Ham)\right)+\sum_{i=1}^{|V|}\log\,(1+f_i)*\log\,(p¿(i\vee Ham))¿$$

$$¿\log\left(\frac{1}{4}\right)+\log(1+1)*\log(0.375)+\log(1+2)*\log(0.375)+\log(1+1)*\log(0.25)$$

$$R_{ham}=-4.105$$

Luego, como $R_{spam}<R_{ham}\rightarrow El\ mensaje\ dado\ es\ HAM$.

Nota:

Si se quiere obtener la probabilidad de que el mensaje sea, por ejemplo, SPAM, se debe hacer una conversión sobre $R_{spam}\ y\ R_{ham}$. Podemos obtener los valores positivos de estos, para luego normalizar. Obtenemos los valores positivos:

$$V_{ham}=\frac{-1}{R_{ham}}=\frac{-1}{-4.105}=0.244$$

$$V_{spam}=\frac{-1}{R_{spam}}=\frac{-1}{-5.255}=0.19$$

Luego la probabilidad de que el mensaje dado sea **spam** es:

$$p(Spam|mensaje)=\frac{V_{spam}}{V_{spam}+V_{ham}}=\frac{0.19}{0.244+0.19}=0.438$$

Análogamente obtendremos que la probabilidad de que el mensaje dado sea **ham** es 0.562.

EXPERIMENTOS COMPUTACIONALES

Descripción del sistema (herramientas de h/w y s/w usadas, sistema operativo, etc.)

Hardward

- Notebook HP
- Core i3, segunda generación
- Sistema operativo Windows 10

Software

- Python 3.6
- Biblioteca Pandas de Python
- Biblioteca Numpy de Python
- Biblioteca Scikit Learn de Python
- Biblioteca notebook de Python. Jupyter Notebook.

INSTANCIAS DE PRUEBA

- Se realizaron pruebas con el dataset en bruto y con la función multinomial original, además con un diccionario de tamaño 3000, el resultado fue:
  - Aciertos: 94.69%
  - Error: 5.3%
- Se realizaron pruebas con el dataset en bruto y con la función multinomial lineal (llevada al espacio logarítmico) además con un diccionario de tamaño 3000, el resultado fue:
  - Aciertos: 94.83%
  - Error: 5.16%
- Luego de hacer los cambios en el dataset, probamos con cambiar el tamaño del diccionario (original: 3000). Con el tamaño del diccionario = 1000 se obtuvo:
  - Aciertos: 97.78%
  - Error: 2.22%
- Con el tamaño del diccionario = 2000 se obtuvo:
  - Aciertos: 98.2%
  - Error: 1.79%
- Con el tamaño del diccionario = 3000 se obtuvo:
  - Aciertos: 98.35%
  - Error: 1.65%
- Con el tamaño del diccionario = 4000 se obtuvo:
  - Aciertos: 98.06%
  - Error: 1.94%

Nota: Se redujo el porcentaje de aciertos cuando se aumentó el tamaño del diccionario. Se debe a palabras que pueden ser consideradas como basura que afectan a las probabilidades de las demás palabras.

Conclusiones

- Esta técnica resultó ser muy eficiente para resolver este problema, con 98,35% de aciertos al final de las pruebas respectivas.
- Los resultados encontrados sugieren que el método propuesto, al tomar en cuenta de forma conjunta la modificación del dataset, el suavizado de Laplace, la función multinomial lineal, además del suavizado final para controlar la alta ocurrencia de una misma palabra en un único mensaje, mejora el desempeño del clasificador.
- En nuestro ejemplo no se pudo apreciar el efecto del suavizado de Laplace, puesto que la cantidad de palabras en el diccionario fueron solo 3. En este caso se obtuvo un valor muy cercano al de calcular:

  $$p(i,C)=\frac{f_{iC}}{f_c}$$

- Que el tamaño del diccionario sea mayor, no implica que el resultado sea mejor, esto se debe a que las palabras que pueden llegar a aparecer pueden ya no ser palabras, siendo consideradas como basura. Esto afecta las probabilidades de las palabras que sí pueden ser sustanciales y por ende, al resultado que el sistema obtiene.

REFERENCIAS

[1] Introduction to Probability, Statistics and Random Processes. 1.4.4 Conditional Independence

https://www.probabilitycourse.com/chapter1/1_4_4_conditional_independence.php

[2] Syed Sadat Nazrul, «Multinomial Naive Bayes Classifier for Text Analysis», 2018

https://towardsdatascience.com/multinomial-naive-bayes-classifier-for-text-analysis-python-8dd6825ece67?fbclid=IwAR1OWof7Ftm96Zxuv9sOrFsWEvPehgf31OhwVHIsDDMx2WFAt3CP7OzPNp8

[3] Gavril Ognjanovski, «Building a Spam Filter from Scratch Using Machine Learning», 2018
https://medium.com/analytics-vidhya/building-a-spam-filter-from-scratch-using-machine-learning-fc58b178ea56

[4] Penn University. «BUILDING A SPAM FILTER USING NAÏVE BAYES»

https://towardsdatascience.com/spam-filtering-using-naive-bayes-98a341224038

[5] Atiar Rahman, «Filtering Spam Using Naive Bayes»

https://towardsdatascience.com/spam-filtering-using-naive-bayes-98a341224038