

Aprendizaje Probabilístico: *Naive Bayes*

Jorge Gallego

Facultad de Economía, Universidad del Rosario

Junio 21 de 2017

Introducción

- El algoritmo *Naive Bayes*, como su nombre lo indica, se basa en el teorema de Bayes
- Los clasificadores bayesianos utilizan datos de entrenamiento para calcular la probabilidad de un resultado
- Dicha probabilidad depende de las características de los objetos a clasificar
- Así, sirve para determinar a qué clase pertenece algo según sus características

Introducción

Este método ha sido aplicado a diversos campos, como son:

1. Clasificación de textos, como correos spam
2. Intrusión o detección de anomalías en redes de computadores
3. Diagnóstico de condiciones médicas dados ciertos síntomas

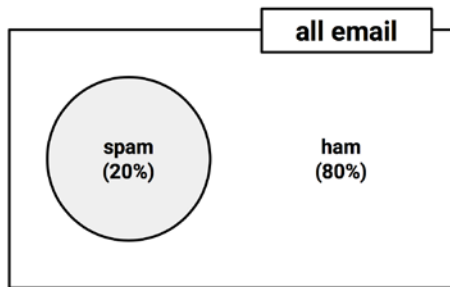
Buen método en casos en los que un conjunto de características deben considerarse simultáneamente para predecir la prob. de algo

Conceptos Básicos

- Calcularemos la probabilidad de un evento a partir de los datos observados
- Sea $P(A)$ la probabilidad de ocurrencia del evento A
- Llamaremos $\neg A$ al complemento de A . Así, $\neg A$ es la no ocurrencia de A
- Un evento y su complemento son mutuamente excluyentes y naturalmente $P(\neg A) = 1 - P(A)$

Conceptos Básicos

Por ejemplo, recibir correo Spam o correo “bueno”



Conceptos Básicos

- Pero muchos eventos no son mutuamente excluyentes
- Y de hecho la ocurrencia de uno nos da información sobre la probabilidad de ocurrencia del otro
- Lo cual nos debe servir para “actualizar nuestra creencia” de si un evento ocurrirá o no
- Por ejemplo, ¿qué nos dice sobre la prob. de estar ante un spam el hecho de que un correo tenga la palabra “Viagra”?

Conceptos Básicos

How to google right Spam | X

from [redacted] [hide details](#) 06:21 (13 hours ago) [Reply](#)

to [redacted]

date 11 October 2009 06:21

subject How to google right







This message contains graphics. If you do not see the graphics, [click here to view](#).


[Click here to forward to friend](#)




October 10, 2009

FREE SHIPPING * NO PRESCRIPTION * HIGHEST ANONYMITY * GUARANTEED DELIVERY

Huge discount

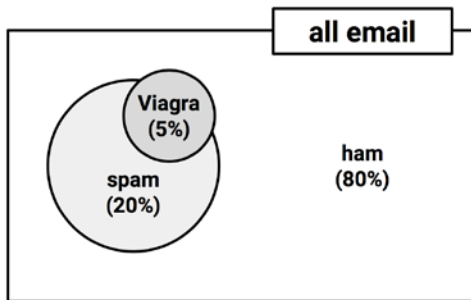
VIAGRA  \$115 PER PILL	CIALIS  \$199 PER PILL	VIAGRA PRO  \$157 PER PILL
CIALIS PRO  \$417 PER PILL	VIAGRA SUPER ACTIVE  \$282 PER PILL	VPXL  \$045 PER PILL



PayPal   

Conceptos Básicos

¿Qué tan probable es que sea spam si tiene la palabra “Viagra”?



Conceptos Básicos

- Dos eventos A y B son independientes si la ocurrencia de uno no nos dice nada sobre la probabilidad de ocurrencia del otro
- La probabilidad conjunta de A y B , si son independientes, es $P(A \cap B) = P(A) \cdot P(B)$
- Pero de hecho nos resultan más interesantes los eventos que son dependientes entre sí
- Porque entonces la ocurrencia de uno nos da información sobre la ocurrencia del otro
- ¿Cómo lo hacemos?

Conceptos Básicos

- Recordemos el teorema de Bayes para la probabilidad condicional:

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B | A) \cdot P(A)}{P(B)}$$

- La prob. de ocurrencia de A dado que B ocurrió es la frec. en la que A y B ocurren juntos sobre la frec. de B
- Si sabemos que B ocurrió, A es más probable cuánto más frecuente sea observar a A y B juntos
- Por ej, si B ocurre el 50% de las veces y A y B simultáneamente el 25%, la prob. de A cuando observamos B es del 50%

Ejemplo: Correos Spam y Viagra

- ¿Cuál es la probabilidad de que un correo sea spam dado que tiene la palabra Viagra?

$$P(\text{spam} \mid \text{Viagra}) = \frac{P(\text{Viagra} \mid \text{spam}) \cdot P(\text{spam})}{P(\text{Viagra})}$$

- $P(\text{spam} \mid \text{Viagra})$ es la *posterior probability*
- $P(\text{spam})$ es la *prior probability*
- $P(\text{Viagra} \mid \text{spam})$ es el *likelihood*
- $P(\text{Viagra})$ es el *marginal likelihood*
- ¡Perdón por el *Spanglish*!

Conceptos Básicos

Supongamos que tenemos los siguientes datos:

Frequency	Viagra		Total
	Yes	No	
spam	4	16	20
ham	1	79	80
Total	5	95	100

Likelihood	Viagra		Total
	Yes	No	
spam	4 / 20	16 / 20	20
ham	1 / 80	79 / 80	80
Total	5 / 100	95 / 100	100

Ejemplo: Correos Spam y Viagra

- De donde podemos deducir lo siguiente:
- $P(\text{spam}) = 20/100 = 0.2$
- $P(\text{Viagra} \mid \text{spam}) = 4/20 = 0.2$
- $P(\text{Viagra}) = 5/100 = 0.05$
- Por lo tanto:

$$P(\text{spam} \mid \text{Viagra}) = \frac{P(\text{Viagra} \mid \text{spam}) \cdot P(\text{spam})}{P(\text{Viagra})} = \frac{0.2 \cdot 0.2}{0.05} = 0.8$$

Ejemplo: Correos Spam y Viagra

- Así, la prob. de que el correo sea spam dado que tiene la palabra Viagra es del 80%
- Probablemente sea mejor filtrarlo
- Naturalmente, cualquier filtro usa muchas más palabras para tomar una decisión
- Pero esta es la esencia del método: predecir la prob. del nivel de una clase en función de información pasada

Naive Bayes

- ¿Por qué es ingénuo?
- Asume que todas las características en la base de datos son igual de importantes
- Además, asume que todas las características son independientes
- Uno esperaría que algunas características sean más importantes que otras
- Por ejemplo, quién es el remitente vs. el texto del mensaje en el caso de spam

Naive Bayes

- También es de esperar que ciertas características nos digan qué tan probable es que otras ocurran
- Por ejemplo, palabras como *Viagra*, *droga* o *prescripción* suelen ir de la mano
- Sin embargo, por más que estos supuestos son irreales, *Naive Bayes* suele hacer un buen trabajo
- Por lo que suele ser el primer candidato para labores de clasificación

Naive Bayes

- Naturalmente, la clasificación se basa en más características
- No solo en la ocurrencia de una palabra
- Es por esto que se hace necesario asumir independencia entre características
- De lo contrario, el problema sería muy difícil de resolver
- En el ejemplo de correos spam, supongamos que clasificamos en función de cuatro palabras (W_1 , W_2 , W_3 y W_4)

Ejemplo: más palabras para filtrar

Supongamos que tenemos los siguientes datos:

Likelihood	Viagra (W_1)		Money (W_2)		Groceries (W_3)		Unsubscribe (W_4)		Total
	Yes	No	Yes	No	Yes	No	Yes	No	
spam	4 / 20	16 / 20	10 / 20	10 / 20	0 / 20	20 / 20	12 / 20	8 / 20	20
ham	1 / 80	79 / 80	14 / 80	66 / 80	8 / 80	71 / 80	23 / 80	57 / 80	80
Total	5 / 100	95 / 100	24 / 100	76 / 100	8 / 100	91 / 100	35 / 100	65 / 100	100

Ejemplo: más palabras para filtrar

- Tenemos 100 mensajes que sabemos si son spam o no
- Para cada uno sabemos de la ocurrencia de estas cuatro palabras
- Debemos clasificar los mensajes entrantes como spam o no con base en la información de los 100 anteriores
- Es cuestión de adaptar el teorema de Bayes para múltiples características, independientes e igualmente importantes

Ejemplo: más palabras para filtrar

- Supongamos un mensaje que tiene las palabras *Viagra* y *Unsubscribe*. Entonces,

$$P(\text{spam} \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4 \mid \text{spam}) \cdot P(\text{spam})}{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4)}$$

- Computacionalmente es difícil resolver esta ecuación
- Pero se simplifica si asumimos independencia entre características:

$$P(\text{spam} \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{P(W_1 \mid \text{spam}) \cdot P(\neg W_2 \mid \text{spam}) \cdot P(\neg W_3 \mid \text{spam}) \cdot P(W_4 \mid \text{spam}) \cdot P(\text{spam})}{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4)}$$

Ejemplo: más palabras para filtrar

- Y con la tabla de frecuencia podemos deducir las probabilidades del numerador
- $P(\text{Viagra} \mid \text{spam}) = 4/20$; $P(\neg \text{Money} \mid \text{spam}) = 10/20$;
 $P(\neg \text{Groceries} \mid \text{spam}) = 20/20$;
 $P(\text{Unsubscribe} \mid \text{spam}) = 12/20$
- Por tanto, el numerador es $0.2 \cdot 0.5 \cdot 1 \cdot 0.6 \cdot 0.2 = 0.012$
- Que es la probabilidad de observar un correo con esas características dado que es spam

Ejemplo: más palabras para filtrar

- La probabilidad de observar un correo así dado que es bueno es $1/80 \cdot 66/80 \cdot 71/80 \cdot 23/80 \cdot 0.8 = 0.002$
- Por tanto, el denominador es

$$P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = 0.012 + 0.002 = 0.014$$

- Luego, la probabilidad de que este correo sea spam es

$$P(\text{spam} \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{0.012}{0.014} = 0.857$$

- Entonces es bastante probable que este correo sea spam

Características Numéricas

- Hasta el momento todas las características han sido dicotómicas
- Como si una palabra está presente o no en un correo
- Pero algunas características pueden ser numéricas
- Por ejemplo, la hora a la que se envía el correo
- ¿Qué hacer en esos casos? Discretizar la variable numérica
- Usar quintiles u otras divisiones que sean razonables
- Por ejemplo, franja del día en la que llega el correo