Bayesiano ingenuo aplicado al TALN

Cuando mencionamos la clasificación de texto, el algoritmo imprescindible que nos permite realizar esta tarea es el bayesiano ingenuo, que se basa en la frecuencia de aparición de las palabras. Para los distintos ejemplos que siguen le invitamos a dejar su ordenador a un lado durante algunos instantes y a equiparse con una hoja de papel, un lápiz y una calculadora.

1. El teorema

Como ya hemos mencionado e ilustrado en el capítulo Principales algoritmos de Machine Learning, el algoritmo bayesiano ingenuo se basa en el teorema de Bayes. Este teorema se fundamenta en las probabilidades condicionales, es decir, la determinación de la probabilidad de que se produzca un acontecimiento en función de otro que ya se ha producido. Los acontecimientos deben ser independientes entre sí.

El teorema de Bayes se explica con la siguiente fórmula:

Texto

Descripción generada automáticamente con confianza media

2. Elegir palabras para un ejemplo

Paro ilustrar el funcionamiento del algoritmo con un ejemplo concreto, hemos elegido el caso de un sitio que ofrece talleres de programación para niños y adolescentes. Vamos a intentar determinar si el hecho de utilizar las palabras clave «Taller niños» y «Curso IA» en la publicidad de las redes sociales, incita a los internautas a los que va dirigida esta publicidad a inscribir más a sus hijos en los cursos.

Para ello, disponemos de los datos de 10 internautas que han hecho o no una inscripción en los cursos después de haber visto una publicidad que solo contenía la palabra clave «Taller niños», otra que solo tenía la palabra clave «Curso IA», una tercera que incluía las dos palabras clave y una cuarta que no contenía ninguna de las palabras clave.

| **internauta** | **taller\_niños** | **curso\_ia** | **inscripción** |
| --- | --- | --- | --- |
| 1 | No | Sí | No |
| 2 | No | Sí | Sí |
| 3 | Sí | No | No |
| 4 | Sí | No | Sí |
| 5 | No | Sí | No |
| 6 | No | Sí | No |
| 7 | Sí | No | Sí |
| 8 | No | Sí | Sí |
| 9 | No | No | Sí |
| 10 | Sí | No | Sí |

Podemos extraer algunas conclusiones de esta tabla:

* El primer internauta ha visto la publicidad que contenía la palabra clave «Curso IA» pero no ha hecho la inscripción.
* El internauta n.º 4 ha visto la publicidad que contenía la palabra clave «Taller niños» y ha hecho una inscripción.
* El internauta n.º 9 ha visto la publicidad que no contenía ninguna de las palabras clave, «Taller niños» y «Curso IA», y ha hecho una inscripción.

**a. Determinación de las probabilidades**

Con objeto de determinar las distintas probabilidades necesarias para la aplicación del teorema bayesiano ingenuo, respondemos a estas preguntas:

* *¿Cuántas inscripciones hay?* 6 (6 «sí» en la columna inscripción).
* *¿Cuál es la probabilidad de que haya una inscripción? Designaremos la probabilidad como* P(inscripción =sí)= 6/10 = 0,6, donde 6 es la cantidad de inscripciones, y 10, la cantidad de observaciones.
* *¿Cuántos visitantes hay sin inscripción?* 4 (4 «no» en la columna inscripción).
* *¿Cuál es la probabilidad de que haya una inscripción? Designaremos la probabilidad como*P(inscripción =no)= 4/10 = 0,4, donde 4 es la cantidad de no inscripciones, y 10, la cantidad de observaciones.
* *¿Cuántas inscripciones hay con la palabra clave «Taller niños»?* 3 (los internautas n.° 4, 7 y 10 porque tienen «sí» en la columna taller\_niños y «sí» en la columna inscripción).
* *¿Cuál es la probabilidad de que haya una inscripción con la palabra clave «Taller niños»?Designaremos esta probabilidad como* P(taller\_niños| inscripción =sí)= 3/6 = 0,5, donde 3 es la cantidad de inscripciones realizadas con la palabra clave, como hemos calculado antes, y 6 es la cantidad de inscripciones.
* *¿Cuántas inscripciones hay con la palabra clave «Curso IA»?* 2 (los internautas n.° 2 y 8 porque tienen un «sí» en la columna curso\_ia y «sí» en la columna inscripción).
* *¿Cuál es la probabilidad de que haya una inscripción con la palabra clave «Curso IA»?Designaremos esta probabilidad como* P(curso\_ia| inscripción =sí)= 2/6 = 0,3333, donde 2 es la cantidad de inscripciones realizadas con la palabra clave que hemos calculado antes, y 6 es la cantidad de inscripciones.
* *¿Cuántas* ***no inscripciones*** *hay con la palabra clave «Taller niños»?* 1 (el internauta n.° 3 porque tiene «sí» en la columna taller\_niños y no en la columna inscripción).
* *¿Cuál es la probabilidad de que no haya inscripción con la palabra clave «Taller niños»?Designaremos esta probabilidad* P(taller\_niños| inscripción =no)= 1/4 = 0,25, donde 1 es la cantidad de inscripciones no realizadas con la palabra clave que hemos calculado antes, y 4 es la cantidad de no inscripciones.
* *¿Cuántas* ***no inscripciones*** *hay con la palabra clave «Curso IA»?* 3 (los internautas n.° 1, 8 y 9 porque tienen «sí» en la columna curso\_ia y «no» en la columna inscripción).
* *¿Cuál es la probabilidad de que no haya inscripción con la palabra clave «Curso IA»?Designaremos esta probabilidad* P(curso\_ia| inscripción =no)= 3/4 = 0,75, donde 3 es la cantidad de inscripciones no realizadas con la palabra clave que hemos calculado antes, y 4 la cantidad de no inscripciones.
* *¿Cuál es la probabilidad de uso de la palabra clave "Taller niños "? Designaremos esta probabilidadP(taller\_niños)= 4/10 = 0,4*.
* *¿Cuál es la probabilidad del uso de la palabra clave «Curso IA»? Designaremos esta probabilidadP(curso\_ia)= 5/10 = 0,5.*

Ahora tenemos distintas probabilidades reunidas en la tabla que aparece aquí debajo:

| **Probabilidad** | **Valor** |
| --- | --- |
| P(inscripción=sí) | 0,6 |
| P(inscripción=no) | 0,4 |
| P(taller\_niños|inscripción=sí) | 0,5 |
| P(curso\_ia|inscripción =sí) | 0,3333 |
| P(taller\_niños|inscripción=no) | 0,25 |
| P(curso\_ia|inscripción=no) | 0,75 |
| P(taller\_niños) | 0,4 |
| P(curso\_ia) | 0,5 |

Vamos a poder utilizar estas probabilidades con la finalidad de hacer las predicciones siguientes ayudados por el teorema bayesiano ingenuo:

* *¿Cuál es la probabilidad de que una persona inscriba a su hijo/a después de haber visto una publicidad que solo contiene la palabra clave «Taller niños»? Designaremos esta probabilidad* P(inscripción|taller\_niños).

Texto

Descripción generada automáticamente con confianza media

| **Probabilidad** | **Valor** |
| --- | --- |
| P(inscripción|taller\_niños) | 0,75 |
| P(inscripción|curso\_ia) | 0,39 |

**b. Conclusión**

A la vista de los distintos resultados, comprobamos que hay un buen rendimiento de la publicidad que solo lleva la palabra clave «Taller niños».