Ciencia de Datos y Lenguaje Natural

Regresión Logística

Grupo PLN - INCO

Universidad de la República

Agenda

- ➔ Regresión logística, definición
- Análisis de sentimientos
- Atributos o variables de entrada
- Función de clasificación
- Método de entrenamiento
- Métodos de evaluación

Regresión logística

La regresión logística es un método supervisado de aprendizaje automático, orientado a realizar clasificación binaria.

Modela directamente la probabilidad condicional de la variable de salida respecto a los datos de entrenamiento.

No impone ninguna restricción de independencia entre los distintos atributos, al estilo de Naïf-Bayes.

Regresión logística

Se puede aplicar para variables nominales, sin ningún tipo de relación de orden.

Implica la construcción manual de atributos que representan a los datos.

Antecedente más simple: la regresión lineal

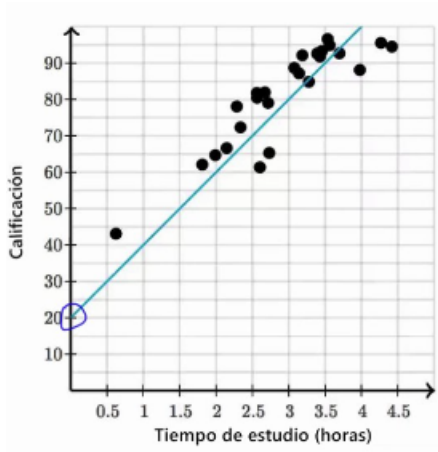
Regresión lineal

Es la tradicional aproximación de un conjunto de puntos por una recta.

Supongamos que tenemos un conjunto de puntos en el plano y deseamos construir una línea recta que aproxime la relación entre abcisas y ordenadas.

Una recta que aproxime nos puede servir para predecir el valor de la 'y' dado el valor de la 'x'.

Ejemplo de regresión lineal



https://www.youtube.com/watch?v=Eo9Yx-hVpLQ

Ejemplo de regresión lineal

- El ejemplo anterior nos permite predecir una variable en función de otra, aproximando por una recta.
- Los resultados tienen un margen de error, los métodos de resolución buscan minimizar el error de aproximación.

Ejemplo de regresión lineal

- Esto es generalizable a varias variables de entrada, nos daría una recta en un espacio n-dimensional.

$$y = w_1 x_1 + w_2 x_2 + ... + w_n x_n + b$$

- Un método habitual para resolver esto es el de mínimos cuadrados: se predice una recta tal que sea mínimo el error (suma de diferencias al cuadrado) entre el valor que predice la solución y el valor real de los puntos que se conocen.

Agenda

- Regresión logística, definición
- ➔ Análisis de sentimientos
- Atributos o variables de entrada
- Función de clasificación
- Método de entrenamiento
- Métodos de evaluación

Análisis de sentimientos

- Existen diversas aplicaciones que se resuelven como un problema de clasificación: determinar si un mensaje es o no spam, otorgar o no un crédito en un banco.

- Veremos como ejemplo un poco más en detalle el siguiente problema: dado un conjunto de opiniones sobre una película o un producto, entender, a partir del texto, si se trata de una opinión positiva o negativa.

- En algún sentido el problema es más simple que la regresión lineal: la variable de salida tiene solo 2 valores posibles, positivo y negativo.

Análisis de sentimientos

1. Buen estuche, excelente relación calidad-precio. 1
2. Genial para la mandíbula. 1
3. Atado al cargador para conversaciones de más de 45 minutos. PROBLEMAS IMPORTANTES!! 0

- Extraído y traducido de comentarios sobre productos de Amazon.
- 1 es positivo y 0 negativo.
- Algunas palabras parecen ser indicadores fuertes.
- También la mayúscula parece ser un indicador.

Análisis de sentimientos

1. Tengo que sacudir el enchufe para que se alinee correctamente y obtener un volumen decente. 0
2. Si tiene varias docenas o varios cientos de contactos, imagine la diversión de enviarlos uno por uno. 0
3. Si eres propietario de un Razr... ¡debes tener esto! 1

- Parece que no basta con encontrar palabras positivas y negativas
- Tanto 1 como 2 tienen palabras "positivas". (correctamente, diversión) y son negativos.
- En 2 hay ironía, dificil de detectar.
- 3 es positiva, pero no hay un indicador claro.

Análisis de sentimientos

- Vamos a aprender a partir del texto
- ¿Podremos aplicar regresión lineal?
- No, porque no tenemos variables numéricas, nuestras variables son las palabras que hay en el texto. Las vamos a transformar a números de algún modo, pero la salida luego de operar con los coeficientes de una combinación lineal va a ser una probabilidad.
- Vamos a utilizar un método adecuado para variables categoriales, con cierta similitud con la regresión lineal: regresión logística.

Agenda

- Regresión logística, definición
- Análisis de sentimientos
- ➔ Atributos o variables de entrada
- Función de clasificación
- Método de entrenamiento
- Métodos de evaluación

Atributos

- Cómo hacer para transformar texto, palabras en números ?
- Variadas propuestas, todas ellas son funciones numéricas de características del texto.

Ejemplos de atributos, J&M, chap 5

Var	Definition
x_1	count(positive lexicon words $\in$ doc)
x_2	count(negative lexicon words $\in$ doc)
x_3	$\begin{cases} 1 & \text{if "no"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$
x_4	count(1st and 2nd pronouns $\in$ doc)
x_5	$\begin{cases} 1 & \text{if "!"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$
x_6	log(word count of doc)

- Buen estuche, excelente relación calidad-precio. 1
- Genial para la mandíbula. 1
- Atado al cargador para conversaciones de más de 45 minutos. PROBLEMAS IMPORTANTES!! 0

Ejemplos de atributos, J&M, chap 5

Var	Definition
x_1	count(positive lexicon words $\in$ doc)
x_2	count(negative lexicon words $\in$ doc)
x_3	$\begin{cases} 1 & \text{if "no"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$
x_4	count(1st and 2nd pronouns $\in$ doc)
x_5	$\begin{cases} 1 & \text{if "!"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$
x_6	log(word count of doc)

- Buen estuche, excelente relación calidad-precio. 1 $x_1 = 2, x_6 = 5$
- Genial para la mandíbula. 1 $x_1 = 1, x_6 = 4$
- Atado al cargador para conversaciones de más de 45 minutos. PROBLEMAS IMPORTANTES !! 0 $x_2 = 1, x_5 = 1, x_6 = 11$

Ejemplos de atributos

Se puede representar palabras de interés con su frecuencia (absoluta o relativa).

Por ejemplo,

$$f_{malo}(x) = \begin{cases} K \text{ si } x = malo \text{ y } cant(x) = K \\ 0 \text{ en caso contrario} \end{cases}$$

Se puede utilizar *templates* para generar un conjunto de valores. Por ejemplo si deseamos determinar qué puntos son marca de fin de oración podemos generar atributos que representen a todas las palabras que preceden a un punto.

Agenda

- Regresión logística, definición
- Análisis de sentimientos
- Atributos o variables de entrada
- ➔ Función de clasificación
- Método de entrenamiento
- Métodos de evaluación