Asignación Latente de Dirichlet (LDA)

| ¿Qué es?
| Un modelo de aprendizaje no supervisado para el modelado de temas. Imagina que tienes una gran colección de textos (documentos) y quieres descubrir automáticamente los temas que se tratan en ellos, sin que nadie te los haya dicho de antemano. LDA hace precisamente eso.
|
| Objetivo
| Descubrir la estructura de temas oculta en una colección de documentos. Responde a dos preguntas clave:
1. ¿De qué temas está hecho cada documento?
2. ¿De qué palabras está hecho cada tema?
|
|
| Analogía Simplificada |
Piensa en un bibliotecario que tiene miles de libros sin catalogar. LDA es como un asistente inteligente que revisa los libros y dice: "Este libro tiene un 70% de temática sobre ‘Ciencia’ y un 30% sobre ‘Matemáticas’. Por otro lado, la temática ‘Ciencia’ se compone principalmente de palabras como ‘experimento’, ‘teoría’ y ‘física’".
|
|
| Funcionamiento
| Proceso generativo probabilístico: LDA asume que los documentos se crearon de la siguiente manera:
1. Se eligen temas de una distribución de temas global (Dirichlet).
2. Para cada tema elegido, se eligen palabras de una distribución de palabras específica para ese tema (también Dirichlet). El algoritmo de LDA invierte este proceso: dado que tenemos las palabras de los documentos, trata de inferir cuáles son los temas y qué palabras pertenecen a cada tema.
|
| Componentes Clave
| Distribución de Dirichlet: Es una distribución de probabilidad sobre distribuciones de probabilidad. Se usa para modelar la "mezcla" de temas en un documento y la "mezcla" de palabras en un tema. Es la base matemática que le da nombre al modelo.
Documentos: La unidad de texto de entrada.
Temas: Conceptos abstractos que descubrimos. Cada tema es una distribución de probabilidad sobre el vocabulario.
Palabras: Los tokens individuales del texto.
|
| Entrada (Input)
| Un corpus de documentos (por ejemplo, una lista de artículos, tweets, correos electrónicos). A menudo se requiere un preprocesamiento (tokenización, lematización, eliminación de stopwords).
|
| Salida (Output)
| Dos tipos de distribuciones:
1. Distribución Documento-Tema: Para cada documento, un vector de probabilidades que indica la proporción de cada tema en ese documento (ej: Documento A = 80% Tema 1, 20% Tema 2).
2. Distribución Tema-Palabra: Para cada tema, un vector de probabilidades sobre el vocabulario que indica qué palabras son más representativas de ese tema (ej: Tema 1 = "gobierno" (0.15), "elección" (0.12), "ley" (0.10)...).
|

| Hiperparámetros Importantes
| ***K (Número de Temas)***: Es un valor que debemos decidir de antemano. No hay una forma perfecta de elegirlo, pero hay métodos heurísticos (como la perplejidad o la coherencia del tema) para encontrar un buen valor.
alpha y beta: Parámetros de la distribución de Dirichlet que influyen en la dispersión de las distribuciones. Controlan qué tan "mezclados" están los temas en los documentos y las palabras en los temas.
|
| Aplicaciones Prácticas
| - *Organización de datos:* Clasificar grandes volúmenes de documentos automáticamente.
- *Sistemas de recomendación:* Recomendar contenido similar a los temas que un usuario ha consumido.
- *Análisis de sentimiento:* Identificar temas subyacentes en reseñas de productos o redes sociales.
- *Resumen de texto:* Entender rápidamente los temas principales de un corpus grande. |