

Cadenas de Markov y Modelos Ocultos de Markov (HMM)

Andrew Esteban Henao Becerra - Universidad Nacional de Colombia

Resumen

Este trabajo aborda los fundamentos y aplicaciones de las Cadenas y Modelos Ocultos de Markov (HMM), con implementaciones en Python: un generador de texto basado en Don Quijote y un HMM entrenado con frases emocionales. Ambos muestran la dinámica de transición y la inferencia de estados ocultos en procesos secuenciales.

Introducción

Muchos fenómenos reales, como el lenguaje, el clima o el ADN, dependen de su estado previo.

Las Cadenas de Markov modelan estas dependencias mediante probabilidades de transición, y los Modelos Ocultos de Markov amplían este enfoque al incluir estados no observables que generan las observaciones visibles.

Estos modelos se aplican ampliamente en reconocimiento de voz, predicción del clima, biología y generación de texto.

Markov: Estado → Estado

HMM: Estado Oculto → Emisión visible

Marco Teórico

Cadena de Markov:

$$P(X_{t+1} = j \mid X_t = i, X_{t-1}, \dots, X_0) = P(X_{t+1} = j \mid X_t = i) = \pi_{ij}$$

$$\pi A = \pi$$

El estado futuro depende solo del presente. Con suficiente tiempo, las probabilidades de los estados tienden a una distribución estacionaria π .

Modelo Oculto de Markov (HMM):

$$P(O \mid \lambda) = \sum_Q P(O \mid Q, \lambda) P(Q \mid \lambda)$$

$$A = P(S_{t+1} \mid S_t), B = P(O_t \mid S_t), \pi = P(S_1) \quad \lambda = (A, B, \pi)$$

Introduce una capa oculta que genera observaciones visibles con cierta probabilidad.

Metodología e implementación

Se desarrollaron dos proyectos complementarios en Python:

1. Cadenas de Markov.

Generador de texto basado en Don Quijote de la Mancha, que analiza secuencias de 3-4 tokens (Unidad básica del proceso o secuencia modelada) para predecir la palabra siguiente según probabilidades de transición.

texto generado por la markov chain en base a un fragmento del quijote y con tokenizacion 4

“el tambien el nombre y le cobrase famoso y de estruendo como convenia a la nueva orden y al nuevo ejercicio que ya profesaba y asi despues de muchos nombres que formo borro y quito anadio deshizo”

Fragmento de texto original

“Que trata de la condición y ejercicio del famoso hidalgo D. Quijote de la Mancha

En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocin flaco y galgo corredor. “

2. Modelo supervisado HMM.

Implementado con la librería hmmlearn.

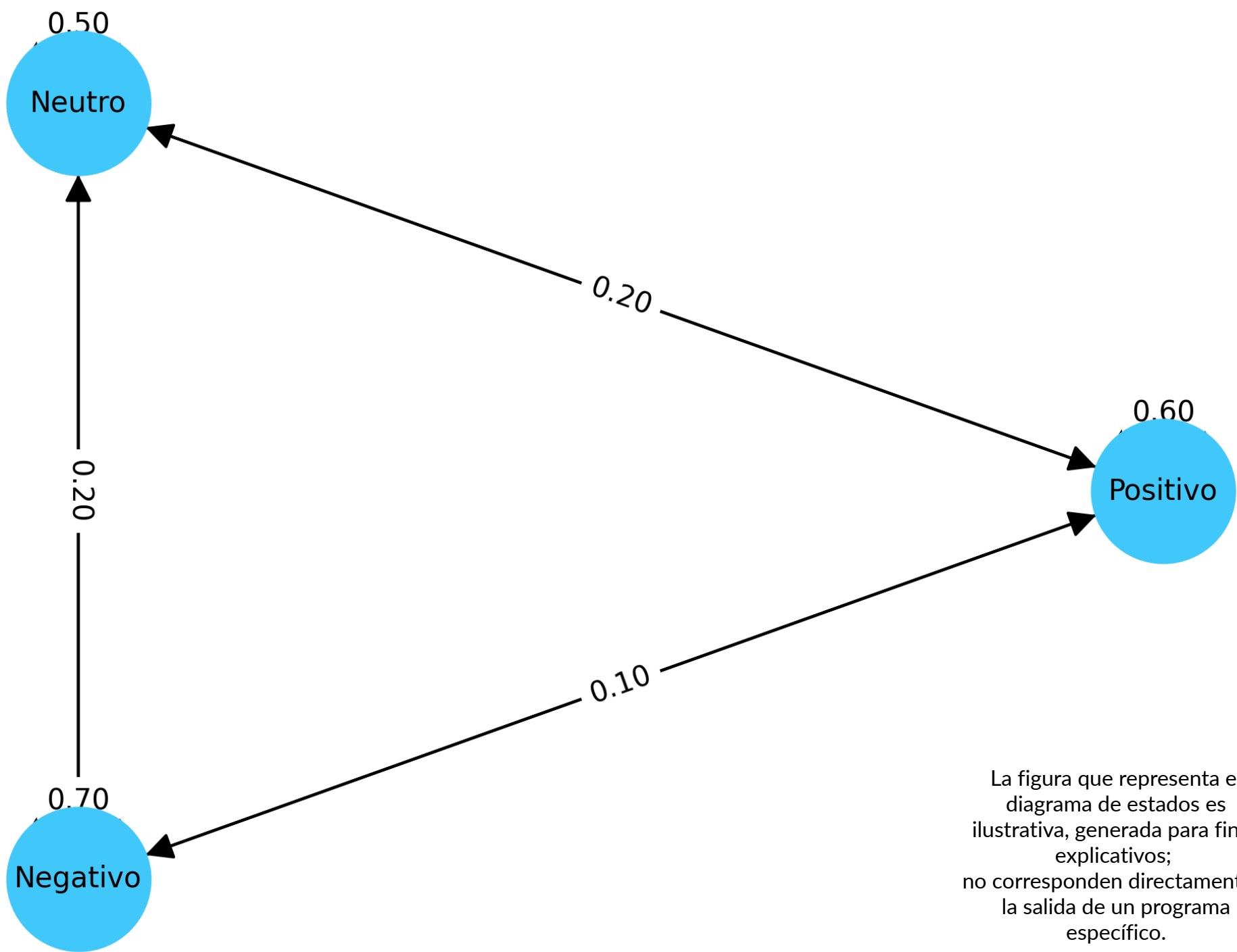
Se emplearon los algoritmos Forward y Backward para estimar las probabilidades de secuencias observadas y estados emocionales ocultos (feliz, triste, neutro).

La figura adjunta muestra un ejemplo de la inferencia realizada por el modelo, donde se representa la estimación de los estados emocionales ocultos a partir de frases de entrada.

```
-----
Texto: I hate this, it was terrible
→ Estado inferido: Positivo
-----
Texto: The experience was awful and bad
→ Estado inferido: Negativo
-----
Texto: It was okay, not great but not bad either
→ Estado inferido: Neutro
-----
```

Figura X. Ejemplo de ejecución del modelo de Modelo Oculto de Markov (HMM) implementado en Python. Se configuraron las matrices de transición (A), emisión (B) e inicialización (π), y se aplicaron los algoritmos Forward y Backward para estimar las probabilidades de las secuencias observadas y los estados emocionales ocultos (feliz, triste, neutro). El bloque muestra la inferencia del estado más probable para una frase determinada.

Diagrama de Transiciones de Estados (HMM de Sentimientos)



La figura que representa el diagrama de estados es ilustrativa, generada para fines explicativos; no corresponden directamente a la salida de un programa específico.

Matriz de Transición de Estados (HMM de Sentimiento)

	Positivo	Neutro	Negativo
Positivo	0.60	0.30	0.10
Neutro	0.20	0.50	0.30
Negativo	0.10	0.20	0.70
	Positivo	Neutro	Negativo

Resultados

Las Cadenas de Markov mostraron coherencia local pero limitada escalabilidad, debido al crecimiento exponencial de estados (el programa solicitó 231GB de RAM para ejecutarse cuando se le pidió analizar todo El Quijote). El modelo HMM permitió identificar patrones parcialmente coherentes entre las secuencias de entrada y las emociones etiquetadas, mostrando correspondencia entre ciertas características del texto y los estados ocultos inferidos.