Workshop
Nuevas Tendencias en Estadística

UNIVERSIDAD NACIONAL DE COLOMBIA

Cadenas de Markov y Modelos Ocultos de Markov (HMM)

Andrew Esteban Henao Becerra - Universidad Nacional de Colombia

Resumen

En este trabajo se presentan los fundamentos y aplicaciones de las Cadenas de Markov y los Modelos Ocultos de Markov (HMM) como herramientas para modelar procesos secuenciales con dependencias probabilísticas.
Se desarrollaron implementaciones en Python: un generador de texto basado en Don Quijote de la Mancha y un modelo HMM entrenado con frases etiquetadas por emoción (feliz, triste, neutro).
Ambos proyectos permiten visualizar la dinámica de transición entre estados y la inferencia de estados ocultos a partir de secuencias observadas, evidenciando las diferencias entre un modelo determinista y uno probabilístico oculto.

Introducción

Muchos fenómenos del mundo real —como el lenguaje, el clima o el ADN— dependen de su estado previo.
Las Cadenas de Markov permiten modelar esta dependencia a partir de la probabilidad condicional de transición entre estados, mientras que los Modelos Ocultos de Markov extienden este concepto al considerar variables no observables que generan observaciones visibles.
Estos modelos son ampliamente utilizados en reconocimiento de voz, predicción meteorológica, análisis de secuencias biológicas y generación de texto.

Markov: Estado → Estado
HMM: Estado Oculto → Emisión visible

Marco Teórico

Cadena de Markov:

$$P(X_{t+1}|X_t) \quad \pi A = \pi$$

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):

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

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 para predecir la palabra siguiente según probabilidades de transición. Debido al crecimiento exponencial del número de estados, el modelo se aplicó solo a fragmentos reducidos del texto.
2. Modelo Oculto de Markov (HMM).
Implementado con la librería hmmlearn, configurando las matrices de transición (A), emisión (B) e inicialización (π).
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.

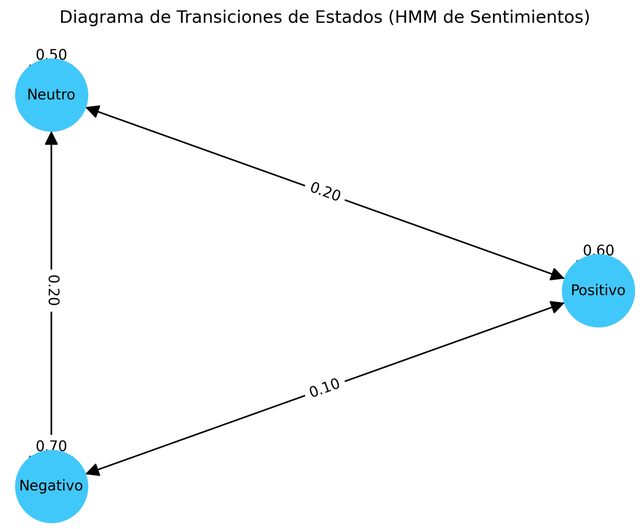
```
------
Texto: I hate this, it was terrible
→ Estado inferido: Positivo
------
Texto: The experience was awful and bad
→ Estado inferido: Negativo
------
Texto: It was okay, not great but not bad either
→ Estado inferido: Neutro
------
```

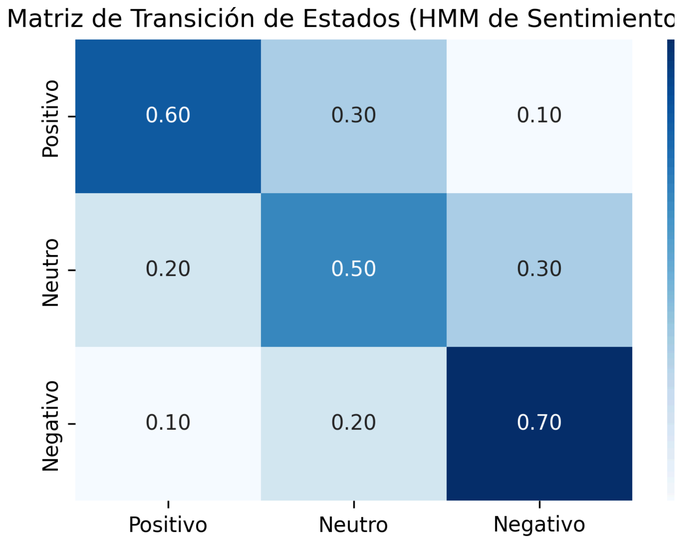
Figura X. Ejemplo de ejecución del modelo de Modelo Oculto de Markov (HMM) implementado en Python.
Se configuraron las matrices de transición (A), emisión (B) e inicialización (π), y se aplicaron los algoritmos Forward y Backward para estimar las probabilidades de las secuencias observadas y los estados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

Código y materiales disponibles en GitHub

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 y torno a hacer en su memoria e imaginacion al fin le vino a llamar rocinante nombre a su parecer alto sonoro y





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 pidio analizar todo El Quijote).
Los HMM, en cambio, integraron dinámicas internas y observaciones, logrando una inferencia robusta de estados no visibles y segmentación precisa en secuencias temporales.

Las figuras presentadas son representaciones ilustrativas generadas para fines explicativos; no corresponden directamente a la salida de un programa específico.

FCEN Sede Manizales | Facultad de Ciencias Exactas & Naturales