Dynamic Pricing con Aprendizaje por Refuerzo (Multi-Armed Bandits)

Arnau Sastre
linkedin.com/in/arnausastre

August 9, 2025

Abstract

Este documento presenta un sistema de pricing dinámico utilizando algoritmos de **aprendizaje por refuerzo** (Reinforcement Learning), enfocado en el problema clásico de *Multi-Armed Bandits*. El objetivo es maximizar ingresos en un entorno de ventas simuladas, donde un agente aprende a seleccionar precios óptimos en función de la conversión observada. Se comparan dos estrategias: Epsilon-Greedy y Thompson Sampling.

1 Introducción

En entornos de negocio dinámicos, la elección del precio óptimo no es trivial. Factores como la elasticidad de la demanda, la variabilidad de conversión o el comportamiento cambiante del usuario afectan directamente a los ingresos.

Este proyecto simula un sistema que aprende automáticamente qué precios generan más ingresos, sin requerir etiquetas ni supervisión directa. Utiliza aprendizaje por refuerzo para explorar diferentes opciones de precio y adaptarse en tiempo real.

2 Simulación del entorno de ventas

Se define un entorno con 5 precios posibles:

$$P = \{10, 20, 30, 40, 50\}$$

Cada precio tiene una tasa de conversión asociada:

$$\text{Conversión}(p) \Rightarrow \text{Probabilidad de compra}$$

El ingreso esperado se calcula como:

$$\text{Revenue Esperado} = p \cdot \text{Conversión}(p)$$

Por ejemplo:

- \$10 → 35.0% → \$3.50
- \$30 → 20.0% → \$6.00 (mejor precio real)
- \$50 → 5.0% → \$2.50

3 Algoritmos implementados

Epsilon-Greedy

- Con probabilidad ε, se explora un precio aleatorio.
- Con probabilidad $1 - \varepsilon$, se elige el precio con mejor media de reward.
- Sencillo, ideal para MVPs y entornos donde se requiere poca complejidad.

Thompson Sampling (Bayesian Bandit)

- Cada precio tiene una distribución Beta que representa su tasa de conversión.
- En cada ronda se muestrea un valor de cada distribución y se elige el precio con mayor valor.
- Convergencia rápida, rendimiento superior.
- Ampliamente usado en plataformas como Booking, Amazon o YouTube.

4 Evaluación de resultados

Métricas principales

- **Ingreso acumulado** por algoritmo
- **Porcentaje de decisiones óptimas** en cada iteración

Visualizaciones

- Evolución del reward total por algoritmo
- Comparativa de ingresos por ronda
- Comparación entre Epsilon-Greedy y Thompson Sampling

5 Aplicaciones reales

Este sistema es directamente aplicable a:

- **Marketplaces:** precios dinámicos según stock, demanda o tipo de cliente
- **Ecommerce:** ajuste inteligente de precios sin perder conversión
- **SaaS:** pruebas de pricing por canal o segmento
- **Publicidad online:** optimización de oferta frente a tasa de clics (CTR)

6 Ventajas del enfoque Bandits

- No necesita etiquetas ni datasets históricos
- Aprende de forma continua y autónoma
- Requiere poca supervisión
- Se adapta a cambios en el comportamiento del usuario

7 Conclusiones

Se ha desarrollado un sistema funcional de pricing dinámico basado en Multi-Armed Bandits. La comparación entre algoritmos demuestra que Thompson Sampling converge más rápido y obtiene mejores ingresos acumulados.

Este tipo de solución es especialmente útil para entornos cambiantes, donde ajustar precios manualmente no es escalable. El sistema puede integrarse en plataformas reales y evolucionar con datos en vivo.

Contacto

Si te interesa implementar un sistema de precios inteligentes en tu negocio o quieres explorar soluciones con aprendizaje por refuerzo, puedes escribirme por **LinkedIn** o **Malt**. También puedes consultar otros proyectos técnicos en mi `GitHub`.