# Trading Bitcoin with Reinforcement Learning

Ricardo, Ismael, Jorge Antonio Hernández Casildo

26 de junio de 2018

**Resumen**

El comercio algorítmico ha existido durante décadas y, en su mayor parte, ha disfrutado de una buena cantidad de éxitos en sus variadas formas. Tradicionalmente, la negociación algorítmica implica la selección de reglas de negociación cuidadosamente diseñadas, optimizadas y probadas por humanos. Si bien estas estrategias tienen la ventaja de ser sistemáticas y de operar a velocidades y frecuencias que van más allá de los comerciantes humanos, son susceptibles a todo tipo de sesgos de selección y no pueden adaptarse a las cambiantes condiciones del mercado. El aprendizaje de refuerzo (RL) por otro lado, es mucho más "libre"de elegir múltiples decisiones. En RL, un .ªgente"simplemente busca maximizar su recompensa en cualquier entorno dado e intenta mejorar su toma de decisiones a través de prueba y error a medida que experimenta más escenarios. En general, RL puede descubrir acciones que los humanos normalmente no encontrarían. Como una prueba de concepto, diseñamos e implementamos un sistema de comercio para bitcoins como prueba del algoritmo, aunque se puede implementar con otros activos. La arquitectura del modelo consiste en maximizar las recompensas implementando Q-learning, basado en la conocida función $Q(s, a)$ que es el valor de una acción aen un estado s, aprendiendo de una red neuronal multicapa perceptron con tres capas ocultas con una capa de entrada del tamaño del estado, y funciones de activación ReLu.

## 1. Introducción

El aprendizaje de refuerzo es apropiado cuando el espacio de estado (la descripción cuantitativa del entorno) es grande o incluso continuo. Puede ser especialmente útil cuando no es práctico obtener etiquetas para el aprendizaje supervisado. El comercio es un buen ejemplo de esto donde las acciones correctas no se conocen e incluso si lo fueran, sería casi imposible de aplicar a cada situación en la que el agente tiene que actuar. RL también es apropiado cuando, como en el comercio, las acciones tienen consecuencias a largo plazo y las recompensas pueden retrasarse. Los ingredientes esenciales para el aprendizaje de refuerzo son estados, acciones, recompensas y una política de selección de acción. En un problema dado, se supone que un agente debe seleccionar la mejor acción dado su estado actual. Esta acción produce una observación del nuevo estado, así como una recompensa, y esto se repite en lo que se conoce como un proceso de decisión de Markov. Para que el agente aprenda su comportamiento o política, la retroalimentación de recompensa para esta secuencia de acciones se usa para ajustar los parámetros del

modelo. Hay dos formas principales de formular el problema: basado en valores y basado en políticas. En un enfoque basado en el valor, se estima el valor de cada estado o el par acción- estado. La política se genera al estimar con precisión estos valores y luego seleccionar la acción con el valor más alto. En un enfoque basado en políticas, que es nuestro método elegido, directamente parametrizamos la política y luego encontramos los parámetros que maximizan las recompensas esperadas.