Proyecto Final de Machine Learning

Freddy Abad, Edwin Cabrera, Daniel Campoverde

**Resumen.-** Empleamos los algoritmos *Q-Learning* y *Deep Q-Learning* en la resolución del problema *“La Final”*, un pequeño juego donde un arquero de fútbol aprende, mediante aprendizaje reforzado, a tapar tiros de penal. Descubrimos un éxito tal… tales resultados, de tal forma.. Blah blah….

## Introducción

Se pretende emplear el algoritmo *Q-Learning* y su variación *Deep Q-Learning* para solucionar el problema *“La final”*.

### Definición del problema

*“La Final”* es un pequeño juego donde un arquero de fútbol aprende a tapar tiros de penal. Tanto el arquero como el balón se mueven en unidades espaciales discretas dentro de una rejilla de *9x8* recuadros. El problema se rige al siguiente conjunto de reglas:

1. Cuando el arquero ataja, se ganan dos puntos (+2)
2. Cuando el arquero no ataja, se pierden dos puntos (−2)
3. Cuando la pelota es disparada fuera del arco y el arquero se mueve a ese lugar, se pierde un punto (−1). Tenga en cuenta que el arquero sí podría moverse, pero no fuera del arco (vea el punto 4)
4. Cuando la pelota es disparada fuera del arco y el arquero no se sale del arco, se gana un punto (+1)
5. Los disparos del balón son realizados enseguida uno después del otro. Esto quiere decir que el arquero, luego de un disparo, no se posiciona al centro del arco sino que trata de atajar el siguiente disparo desde el lugar en el que quedó luego del primer disparo.

## Marco Teórico

El *aprendizaje reforzado* es una área de *machine learning* inspirada en la psicología del comportamiento, en esta se modela el entorno como un *proceso de decisión de Markov* (MDP), donde ciertas acciones tomadas por el *agente* modifican el estado y suponen una cierta recompensa. El aprendizaje supervisado es particularmente apropiado para problemas que involucran beneficios a corto y largo plazo donde el beneficio total obtenido de ambos a de ser maximizado.

*Q-Learning* es una técnica de aprendizaje reforzado de tipo *model-free*, es decir, no requiere que el entorno sea previamente formulado en un modelo. Puede ser empleado para encontrar *reglas de acción* óptimas para un proceso de decisión de Markov aprendiendo una función *Q(s,a)* que representa el beneficio de tomar una acción *a* en un estado *s*.

El uso de técnicas de aprendizaje reforzado permite crear soluciones para tareas específicas, sin embargo, es altamente deseable encontrar soluciones generalizadas que sean aplicables en varias tareas diversas. Con este objetivo, *Deep Q-Learning* combina *Q-Learning* con aprendizaje profundo, que usa varias capas de unidades de procesamiento y representaciones que corresponden con distintos niveles de abstracción.

## Materiales y Métodos

Desarrollamos la solución al problema planteado usando el lenguaje *python* y la librería *pygame* para el control del aspecto gráfico del juego.

El juego emplea la división de cuadrículas para el desplazamiento de los elementos, según la especificación, la interfaz resultante se observa en la *Figura 1*.

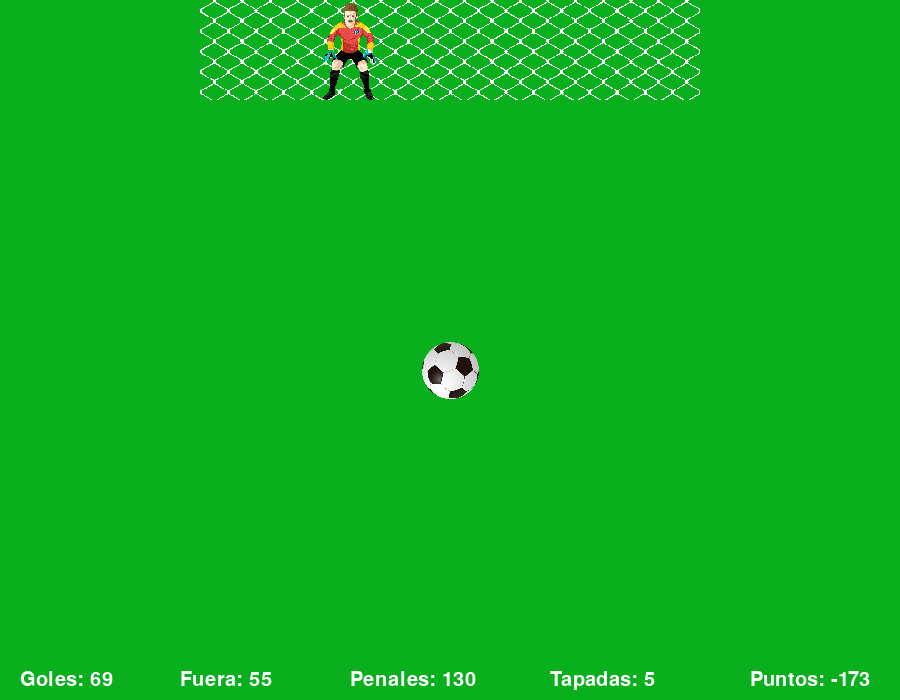


Figura 1: Interfaz del Juego

La solución se estructura en 3 clases, según se observa en la *Figura 2*. El estado de las instancias correspondientes soporta el algoritmo de aprendizaje y el control gráfico del juego, proceso que se ejecuta en un *bucle de episodios.*

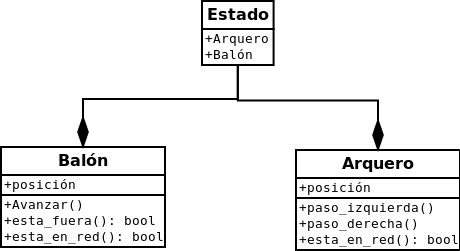


Figura 2: Diagrama de clases

El funcionamiento general a alto nivel de la solución se aprecia en la *Figura 2*; Existe una matriz ***Q*** de dimensiones |S| *x* |A|, donde |S| es el número de estados posibles en el juego y |A| el número de acciones posibles. Cada elemento de la matriz ***Q*** representa el *beneficio* de tomar una acción ***a*** desde cierto estado ***s*.** En cada *episodio* del juego, esto es, en cada tiro de penal, se define el estado actual ***e*** como la combinación de la posición del arquero y la posición del balón, luego se elije de la matriz ***Q*** la mejor acción posible para dicho estado, la misma que permite saltar a un nuevo estado definido ***e’***. Tras patear el balón, se aplican el conjunto de reglas previamente definidas para evaluar el *reward* producto de la acción del arquero, con el cual se actualiza la matriz ***Q*** según la *Ecuación 1,* donde ***lr*** representa el *learning rate* y ***gamma***el factor de descuento para el beneficio de estados futuros.

(1)

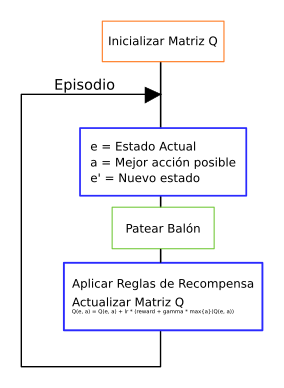


Figura 3: Funcionamiento de alto nivel

La matriz de aprendizaje ***Q*** se implementa como un diccionario que asocia instancias de *Estado* a vectores de acciones que representan el movimiento del *Arquero* sobre el eje X.

### Optimización de parámetros influyentes

Los parámetros *learning rate (lr)* y *Gamma*, definidos para , configuran y alteran el comportamiento del proceso de aprendizaje. Con propósitos de optimización, probamos multiples configuraciones y comparamos el *Puntaje* obtenido como resultado de las reglas planteadas contra un número de *Episodios* fijo de *2000*. Según se observa en la *Figura 4,* el puntaje más alto para el número definido de Episodios es de *26306*  y ocurre para un *Learning rate = 1* y  *Gamma = 0.*

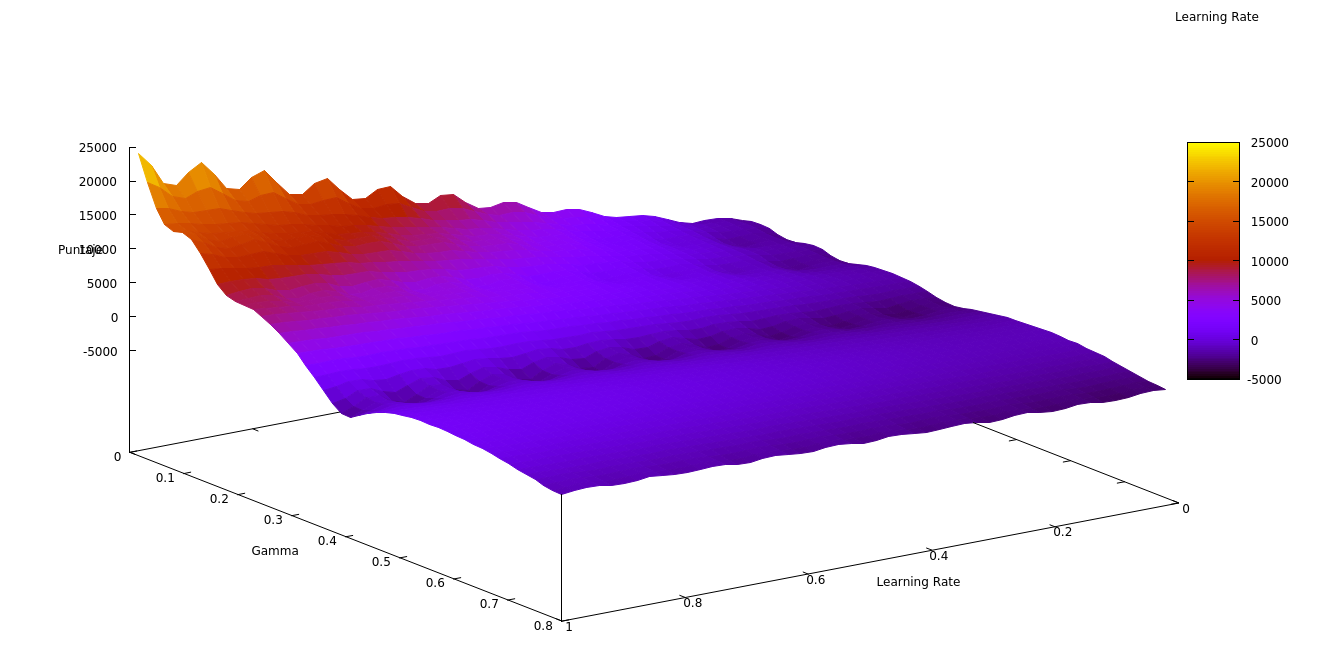


Figura 4: Optimización de parámetros

## 

## Resultados y Discusión

El puntaje máximo obtenido para la configuración explorada en la *Figura 3* concuerda con lo que cabría esperar dada la naturaleza del problema: El aprendizaje no obtiene ningún beneficio al analizar el impacto de acciones futuras en nuevos tiros de penal, sino únicamente del beneficio de la acción inmediata, situación que se maximiza con un *Gamma = 1.*

## Conclusiones

## Bibliografía

1. Entender lo que es apredizaje reforzado
2. Entender el algoritmo Q-learning
3. Entender deep Q-learning

Escribir el reporte escrito y presentación sobre los 3 puntos y sobre las experiencias al realizar la solución