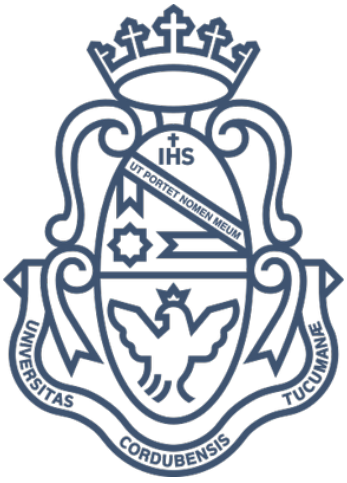
Universidad Nacional de Córdoba

FACULTAD DE CIENCIAS EXACTAS FÍSICAS Y NATURALES

CARRERA INGENIERÍA EN COMPUTACIÓN



ESTUDIO EXPLORATORIO PARA LA DETERMINACIÓN DE POLÍTICA EN SISTEMAS CONTROLADOS POR REDES DE PETRI

Práctica Profesional Supervisada

Alumnos:
Casas, Nicolás
40574806, nicolas.casas@mi.unc.edu.ar, 3584305868
Ciarrapico, Nicolás Valentin
39880567, nicolas.ciarrapico@mi.unc.edu.ar, 2994840005

Supervisor: Ing. Luis Ventre
Tutor: Dr. Ing. Orlando Micolini

3 de noviembre de 2022

Índice

1. Introducción **2**
1.1. Motivación . . . 2
1.2. Objetivos . . . 2
1.3. Requerimientos . . . 2
1.3.1. Lista de requerimientos . . . 3
1.4. Modelo de Desarrollo . . . 3
1.5. Gestión de riesgos . . . 3

2. Marco Teórico **5**
2.1. Redes de Petri . . . 5
2.1.1. Red de Petri Generalizada . . . 5
2.1.2. Red de Petri Marcada . . . 5
2.1.3. Tipos de Arcos . . . 5
2.1.4. Matrices . . . 6
2.1.5. Sensibilizado de una transición . . . 6
2.1.6. Disparo de una transición . . . 7
2.1.7. Ecuación de estado . . . 7
2.1.8. Extensión de la ecuación de estado . . . 7
2.1.9. Propiedades de las Redes de Petri . . . 8
2.1.10. Invariantes . . . 10
2.2. Reinforcement Learning . . . 10
2.2.1. Componentes del RL . . . 10
2.2.2. Tipos de algoritmos de RL . . . 11
2.3. Redes de Bayes . . . 13
2.3.1. La estadística de Bayes . . . 13
2.3.2. Probabilidad: definición axiomática . . . 14
2.3.3. Teorema de Bayes . . . 14
2.3.4. Fórmula de Bayes . . . 14
2.3.5. Redes Bayesianas . . . 14

3. Desarrollo **17**
3.1. Curso sklearn master . . . 17
3.2. Curso MIT: Introduccion to deep learning . . . 17
3.3. Componentes de software . . . 17
3.3.1. TensorFlow . . . 17
3.3.2. Matplotlib . . . 17
3.3.3. NumPy . . . 17
3.3.4. re — Operaciones con Expresiones Regulares . . . 18
3.3.5. pyAgrum . . . 18
3.4. Condiciones . . . 18
3.5. Micocubo . . . 19
3.5.1. Tamaño de la tabla Q . . . 19
3.5.2. Sesgo inicial . . . 19
3.5.3. Haciendo la tabla probabilística . . . 20

1. Introducción

1.1. Motivación

Un sistema de control es un arreglo de componentes conectados de tal manera que pueda comandar, dirigir o regular, asimismo o a otro sistema, con el fin de reducir las probabilidades de fallo y obtener los resultados deseados[6]. Los avances en la teoría y la práctica del control automático aportan los medios para obtener un desempeño óptimo de los sistemas dinámicos, mejorar la productividad, aligerar la carga de muchas operaciones manuales, repetitivas y rutinarias, así como de otras actividades. Cuando se desea mantener un objetivo de control determinado en un sistema son dos los esquemas de control que se pueden considerar: sistemas de control en bucle abierto y sistemas de control en bucle cerrado o realimentado.

En el presente trabajo nos centraremos en los sistemas de control en bucle cerrado. En estos sistemas existe una realimentación de la señal de salida o variable a controlar y se compara esta variable con la señal o variable de referencia de forma que, en función de esta diferencia entre una y otra, el controlador modifica la acción de control para obtener en la variable de salida un valor acorde al requerimiento o variable de entrada.

Estos sistemas entonces, tienen una interacción constante con el ambiente donde operan y la manera en que toman decisiones para interactuar con el mismo es a través de lo que llamamos política. Las diversas formas de implementar esta política en la actualidad (estáticas, aleatorias o round robin) resultan ser insatisfactorias cuando se aplican en sistemas reactivos, como los descriptos anteriormente. Es por esto que estudiaremos otras alternativas de realimentación que permitan obtener una política del sistema de manera dinámica, capaz de cumplir con requerimientos de usuario preestablecidos, y de esta forma optimizar los recursos empleados y el tiempo de respuesta del sistema.

1.2. Objetivos

El objetivo del presente trabajo es encontrar el compensador correcto para construir un sistema realimentado capaz de generar de manera dinámica una política que nos permita ordenar los disparos de una red de Petri de forma tal que se cumpla con requerimientos específicos definidos por el usuario. Además es importante poder realizar esto sin modificar el grafo de la red de Petri, teniendo en cuenta que la misma debe ser de tipo *S3PR* y no debe poseer deadlock. Para llevar a cabo esto estudiaremos las redes Bayesianas que usan la probabilidad para tratar la incertidumbre dentro de la inteligencia artificial y compararemos su rendimiento contra los compensadores más utilizados.

1.3. Requerimientos

Las definiciones de requerimientos del sistema especifican qué es lo que el sistema debe hacer (sus funciones) y sus propiedades esenciales y deseables[9]. Para crear definiciones de requerimientos del sistema requiere consultar con los clientes del sistema y con los usuarios finales, por lo que para obtener la información necesaria en esta etapa se llevaron adelante las siguientes tareas:

1. *Entrevistas*: Se realizaron una serie de entrevistas con el director de tesis Luis Ventre en conjunto con el co-director Orlando Micolini, en las cuales se definió de forma general los objetivos del proyecto. Además nos brindaron bibliografía de guía para estudiar posibles soluciones.

2. *Búsqueda y estudio de documentos*: Se analizó la bibliografía provista por los docentes, se buscó profundizar en material relacionado y se prepararon posibles soluciones que se discutieron en nuevas entrevistas.

A partir de esto se generó una lista de requerimientos para establecer las funcionalidades que tiene que cumplir el sistema.

1.3.1. Lista de requerimientos

1. El sistema debe ser capaz de controlar los disparos de una Red de Petri de forma tal que cumpla con requerimientos de invariantes de transición.
2. El sistema debe funcionar sin alterar o modificar el grafo de la red de Petri.
3. El sistema debe ser capaz de funcionar con cualquier Red de Petri S3PR.
4. Se debe implementar un modulo capaz de leer los archivos .html generados por Petrinator y generar las matrices de la red de petri en texto plano.
5. El sistema debe poder ejecutarse en Windows y Linux.
6. El usuario debe poder determinar los requerimientos por invariante de transición que debe seguir el sistema para una red en particular con el fin de generar las políticas de la misma.
7. El sistema debe ser capaz de ilustrar si el mismo convergió de forma correcta.

1.4. Modelo de Desarrollo

El modelo adoptado para el desarrollo del presente trabajo es el modelo iterativo. En este se realizan múltiples iteraciones en las que se desarrolla una versión del proyecto, se la expone a revisión por parte del cliente y se recolectan las correcciones del mismo. Estas se utilizan como entrada de la nueva iteración donde se genera una nueva versión repitiendo el proceso hasta obtener un sistema adecuado según los requerimientos planteados.

Las actividades de especificación, desarrollo y validación están entrelazadas en vez de separadas, con rápida retroalimentación a través de las actividades. Ian Sommervile[9].

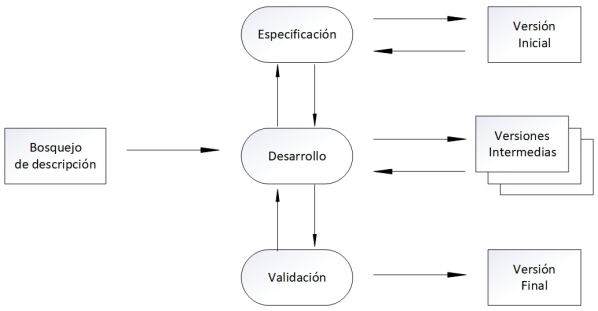


Figura 1: Desarrollo incremental.

En cada nueva versión se busca incorporar algunas de las funciones requeridas por el usuario. Es habitual comenzar por aquellas funcionalidades mas importantes o urgentes. Esta interacción constante con el usuario permite controlar, desde una etapa temprana del desarrollo, que el producto es lo que se requiere. Si esto no se cumple, solo la ultima iteración se debe modificar y, probablemente, definir una nueva funcionalidad para próximas iteraciones.

1.5. Gestión de riesgos

De forma simple, se puede concebir un riesgo como una probabilidad de que una circunstancia adversa ocurra. Los riesgos son una amenaza para el proyecto, para el software que se está desarrollando y para la organización[9]. Se los puede agrupar en tres categorías:

- *Riesgos del proyecto*: Repercuten sobre el cronograma o los recursos del proyecto.
- *Riesgos del producto*: Afectan la calidad o el rendimiento del software que se quiere desarrollar.
- *Riesgos del negocio*: Influyen a la organización que desarrolla o comercializa el software.

Dado que las consecuencias de estos riegos pueden provocar que el proyecto fracase, es importante generar un *plan de gestión de riesgos* una vez definidos los requerimientos y antes de iniciar el desarrollo del primero.

2. Marco Teórico

2.1. Redes de Petri

Las Redes de Petri son modelos matemáticos utilizados para la representación de sistemas con paralelismo, concurrencia, sincronización e intercambio de recursos[7]. La red de Petri esencial fue definida por Carl Adam Petri. Son una generalización de la teoría de autómatas que permite expresar un sistema como eventos concurrentes.

Las redes de Petri están fuertemente asociadas a la teoría de grafos, ya que las mismas pueden representarse como un grafo dirigido bipartito compuesto por cuatro elementos:

- *Plazas*: Representan estados del sistema. El estado de una plaza está dado por la cantidad de marcas o tokens que esta contiene.
- *Token*: Figuran como puntos negros dentro de las plazas. Estos representan el valor específico de una condición o estado y generalmente se traducen en la presencia o ausencia de algún recurso del sistema.
- *Transiciones*: Representan el conjunto de sucesos cuya ocurrencia produce la modificación de los estados de las plazas, y en consecuencia del estado global del sistema.
- *Arcos*: Indican las interconexiones entre las plazas y las transiciones, estableciendo el flujo de tokens que sigue el sentido de la flecha.

2.1.1. Red de Petri Generalizada

Una Red de Petri generalizada no marcada es una cuádrupla $< P, T, Pre, Post >$ donde:

- $P = P_1, P_2, \ldots, P_n$ es un conjunto finito, no vacío, de plazas.
- $T = T_1, T_2, \ldots, T_m$ es un conjunto finito, no vacío, de transiciones, donde $P \cap T = \emptyset$, i.e. los conjuntos P y T son inconexos.
- $Pre : PxT \to \mathbb{N}^P$ es la función de incidencia de entrada y
- $Post : PxT \to \mathbb{N}^P$ es la función de incidencia de salida.

$Pre(p_i, t_j)$ contiene el peso del arco que va de P_i a T_j.
$Post(p_i, t_j)$ contiene el peso del arco que va de T_j a P_i.

2.1.2. Red de Petri Marcada

Una Red de Petri Marcada está definida por el par (N, M), donde N es una Red de Petri y $M : P \to \mathbb{N}^P$ (donde $|P| = p$ es una aplicación llamada **marcado**. $m(N)$ define el marcado de la RdP y m_{pi} indica el marcado de la plaza p_i, es decir, el número de tokens contenido en la plaza i. La marca inicial se denota m_0 y da la cantidad inicial de tokens en todas las plazas de la red, por lo que especifica el estado inicial del sistema.

2.1.3. Tipos de Arcos

- *Común*: Consume o produce una cierta cantidad de tokens de una plaza, de acuerdo al peso del mismo.
- *Inhibidor*: Siempre en la dirección plaza a transición. En el caso que el marcado de la plaza sea mayor o igual al peso del arco, la transición no está sensibilizada. No consume tokens al producirse el disparo de la transición asociada.

- *Lector*: Siempre en la dirección plaza a transición. La transición está sensibilizada si la marca en la plaza es mayor al peso del arco. No consume tokens al producirse el disparo de la transición asociada.
- *Reset*: Siempre en la dirección plaza a transición. Consume todos los tokens de la plaza al dispararse la transición.

2.1.4. Matrices

Para una red con n plazas y m transiciones, mas matrices tienen un tamaño nxm. Cada fila representa una plaza, mientras que cada columna representa una transición. Se conforman de la siguiente manera:

- *Matriz de Incidencia*: Está compuesta por las matrices I^+ e I^-, las cuales son función de los arcos comunes.

  En la matriz I^+, denominada matriz de incidencia de entrada o *post*, cada elemento $post(P_i, T_j)$ contiene el peso del arco que va desde T_j a P_i. Indica la cantidad de tokens generados al disparar la transición.

  En la matriz I^-, denominada matriz de incidencia de salida o *pre*, cada elemento $pre(P_i, T_j)$ contiene el peso del arco que va desde P_i a T_j. Indica la cantidad de tokens consumidos por la transición al realizar el disparo.

  Finalmente, la matriz de incidencia I se forma de la siguiente forma:

$$I = I^+ - I^- \tag{1}$$

- *Matriz de inhibición*: Contiene en cada uno de sus elementos $inh(P_i, T_j)$, el peso del arco de inhibición que va desde $P_i a T_j$
- *Matriz de Reset*: Contiene en cada uno de sus elementos $res(P_i, T_j)$ un 1 si existe un arco de reset que va desde P_i a T_j
- *Matriz de lectura*: Contiene en cada elemento $lec(P_i, T_j)$ el peso del arco lector que va desde P_i a T_j.

2.1.5. Sensibilizado de una transición

Una transición está sensibilizada si todas las plazas de entrada a la transición tienen una marca igual o mayor al peso del arco que une cada plaza con la transición.

Previo a expresar la condición de sensibilizado de manera general, son necesarias las siguientes definiciones:

- $\bullet T_j$ es el conjunto compuesto por las plazas entrante a T_j
- $T_j \bullet$ es el conjunto compuesto por las plazas salientes a T_j.
- $M_k(P_i)$ es el marcado de la plaza P_i antes de disparar la transición T_j.
- $M_{k+1}(P_i)$ es el marcado de la plaza P_i después de dispara la transición T_j.
- w_{ij} es el peso del arco $P_i \rightarrow T_j$
- w_{ji} es el peso del arco $T_j \rightarrow P_i$

De esta forma, el sensibilizado de una transición T_j se expresa como:

$$T_j \text{ está sensibilizada } sii \ \forall P_i \in \ \bullet T_j \ \rightarrow M_k(P_i) \geq w_{ij} \tag{2}$$

Lo definido anteriormente se cumple para redes cuyos arcos son todos comunes. Para redes con arcos inhibidores y/o de lectura cambia la condición de sensibilizado. De esta forma, dependiendo del tipo de arco deben cumplirse ciertas condiciones:

- *Arco de inhibición*: El marcado de la plaza de la cual parte debe es menor que el peso del arco.
- *Arco de lectura*: El marcado de la plaza de la cual parte debe ser mayor o igual al peso del arco.

Por su parte los arcos de reset no alteran la condición de sensibilizado de una transición.

2.1.6. Disparo de una transición

Dada una marca $M_k(P)$, cualquier transición que se encuentre sensibilizada puede ser disparada, este disparo nos llevará a una nueva marca $M_{k+1}(P)$ dada por:

$$M_{k+1}(P) = M_k(P) + I^+(P_i, T_j) - I^-(P_i, T_j) \forall P_i \in P \tag{3}$$

Al disparar la transición T_j se extraen tantos tokens de $\bullet T_j$ como indiquen los arcos que unen estas plazas con T_j. Se añaden a $T_j\bullet$ la cantidad de tokens que indiquen los arcos que unen a T_j con estas plazas. El disparo de una transición T_j se denota $M_k \rightarrow T_j \rightarrow M_{k+1}$.

2.1.7. Ecuación de estado

La ecuación de estado representa matemáticamente el comportamiento dinámico del sistema [5]. Esta permite obtener el estado del sistema luego del disparo de una transición. Se utiliza el marcado en un instante k para calcular el marcado de la red en un instante de tiempo $k + 1$. Por lo tanto la ecuación de estado para una Red de Petri con n plazas y m transiciones se define de la siguiente forma:

$$M_{k+1} = M_k + I * \sigma \tag{4}$$

Siendo:

1. I la matriz de incidencia.
2. σ vector de disparo. Tiene dimensión m x 1 y contiene 1 en la posición de la transición que se quiere disparar.
3. M_k: vector de marcado actual.
4. M_{k+1}: vector de marcado del estado siguiente.

2.1.8. Extensión de la ecuación de estado

La ecuación de estado descrita previamente es útil únicamente en presencia de arcos comunes. Para representar matemáticamente la existencia de los demás arcos nombrados anteriormente se requiere de una matriz para indicar la conexión plaza transición para cada uno de estos.

- *Vector de transiciones des-sensibilizadas por arco inhibidor*: Es un vector de valores binarios de dimensión $m \times 1$, que indica con un cero cuáles transiciones están inhibidas por el arco y con un uno cuales no.
- *Vector de transiciones des-sensibilizadas por arco lector*: Es un vector de valores binarios de dimensión m x 1, que indica con un cero cuáles transiciones están inhibidas por el arco con un uno las que no.
- *Vector transiciones des-sensibilizadas por tiempo*: Es un vector de valores binarios de dimensión m x 1, que indica con un cero cuáles transiciones están inhibidas porque no se ha alcanzado o se ha superado el intervalo de tiempo transcurrido desde que la transición fue sensibilizada.
- *Vector de transiciones reset*: Es un vector de valores enteros de dimensión m x 1, que tiene el valor de la marca de la plaza que se quiere poner a cero, mientras que las otras componentes son uno.

Finalmente utilizando los vectores descriptos anteriormente se puede obtener el vector de sensibilizado extendido E_x:

$$E_x = E \; and \; B \; and \; L \; and \; Z \tag{5}$$

Para poder introducir el brazo reset hay que multiplicar elemento a elemento (#) al vector que resulte de la conjunción por A. Obteniendo así la ecuación de estado extendida:

$$M_{k+1} = M_k + I * ((\sigma \; and \; E_x)\#A) \tag{6}$$

2.1.9. Propiedades de las Redes de Petri

2.1.9.1 Propiedad de limitación

Dada una Red de Petri PN, se dice que una plaza P_i está **k-limitada** por una marcado inicial M_0 si hay un entero natural k tal que, para todas las marcas alcanzables desde M_0, el número de tokens en P_i no es mayor que k. Es decir:

$$\exists \; k \in \mathbb{N} \; / \; \forall M \in marcados(PN) \rightarrow M(P) \leq k \tag{7}$$

Una Red de Petri PN está limitada para un marcado inicial M_0 si todos los lugares están limitados para M_0. Es decir, PN está limitada por k si todas sus plazas están limitadas por k.

A partir de esta definición surgen varios conceptos, entre los cuales se encuentran los siguientes:

- Una Red de Petri es **segura** si todas sus plazas son **1-limitadas**.
- Una Red de Petri es **cíclica** si siempre existe la posibilidad de alcanzar el marcado inicial desde cualquier otro marcado alcanzable.
- Una Red de Petri es **repetitiva** si existe una secuencia de disparos que contiene a todas las transiciones y que lleva a la red desde el marcado actual al mismo marcado.
- Una Red de Petri es **conservativa** si se cumple que el número de tokens en el marcado es siempre el mismo.

2.1.9.2 Propiedad de vivacidad

La **vivacidad** de una transición indica que, en todo instante de la evolución de la red, su disparo es posible. Este concepto es particularmente relevante ya que determina si la ejecución de la red puede o no detenerse en un estado determinado. A partir de esto se puede definir la vivacidad de una red de Petri. Esta propiedad indica que una red es viva para un marcado si todas sus transiciones lo son.

Por otro lado, la **cuasi-vivacidad** de una transición expresa la posibilidad de dispararla al menos una vez a partir de un marcado inicial M_0. De la misma manera que para el caso de la vivacidad, una red de Petri es cuasi-viva si todas sus transiciones lo son.

Gracias a esta última definición, se puede definir la vivacidad en función de la cuasivivacidad de la siguiente manera: una transición es viva si la misma es cuasi-viva en la red para todo marcado alcanzable desde M_0.

La vivacidad está directamente asociada con la ausencia de **deadlock** o **interbloqueo**. En términos generales, el deadlock es el bloqueo permanente de un conjunto de procesos o hilos de ejecución en un sistema concurrente que compiten por recursos del sistema o bien se comunican entre ellos. En el caso de una red de Petri, esto suele ocurrir cuando dos o más transiciones esperan mutuamente por el disparo de la otra, produciendo el bloqueo permanente de esa porción de la red. Una red de Petri viva garantiza la ausencia de interbloqueo sin importar la secuencia de disparos.

2.1.9.3 Propiedad de alcanzabilidad

La **alcanzabilidad** de una Red de Petri es fundamental para el análisis de las propiedades dinámicas de un sistema. A grandes rasgos, permite determinar si el sistema modelado puede alcanzar un determinado estado.

Un marcado M_i es alcanzable desde el marcado inicial M_0 si existe una secuencia finita de disparos que me haga llegar a esa marca. Un marcado M_i es alcanzable desde M_0 si existe una secuencia finita de disparos σ tal que $M_0 \xrightarrow{\sigma} M_i$.

El conjunto de marcados alcanzables puede ser representado mediante un grafo, en el cual cada nodo representa un marcado. El arco entre estos nodos representa la transición que se necesita disparar para llegar de un marcado a otro.

Ejemplo:

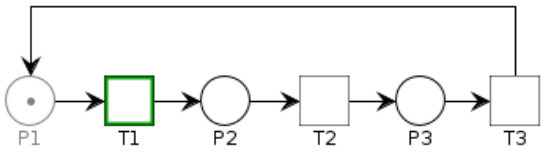


Figura 2: Red de Petri de tres estados.

Esta red cuenta con tres estados:

1. Marcado inicial o al disparar T3 desde m_2: $m_0 = [1, 0, 0]$
2. Marcado al disparar T1 desde m_0: $m_1 = [0, 1, 0]$
3. Marcado al disparar T2 desde m_1: $m_2 = [0, 0, 1]$

Con estos estados es posible generar el siguiente grafo:

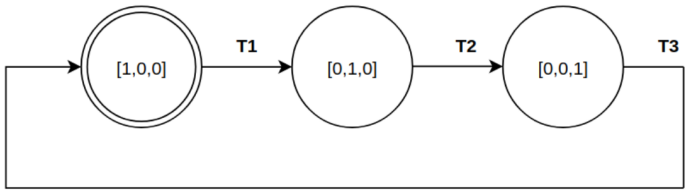


Figura 3: Grafo de alcanzabilidad.

2.1.9.4 Cobertura de una Red de Petri

Si la red de Petri no es acotada habrá plazas cuya cantidad de tokens aumentará indefinidamente, lo que también se repetirá en los nodos del grafo de alcanzabilidad, por lo que el algoritmo para obtener el árbol de alcanzabilidad no convergerá. Para estos casos existe otro tipo de análisis denominado **grafo de cobertura**.

El grafo de cobertura consiste en graficar todas las marcas posibles de la red, pero colapsando en un único marcado genérico, representado con el sımbolo ω, aquellas plazas cuya cantidad de tokens crecerá infinitamente.

2.1.10. Invariantes

Las invariantes de una red son propiedades independientes tanto del marcado inicial como de la secuencia de disparos, y pueden asociarse a ciertos subconjuntos de plazas o de transiciones. De esta forma surgen dos conceptos:

- **Invariante de plaza** o **p-invariante**: Se llama así a un conjunto de plazas si dado un determinado vector de ponderación con enteros positivos o cero se cumple que:

$$q_1 \text{ x } M(P_1) + q_2 \text{ x } M(P_2) + \cdots + q_n \text{ x } M(P_n) = K$$

  Siendo K constante para todo marcado. Es un componente conservativo independiente del marcado inicial. Sin embargo, el valor de la constante sí depende del marcado inicial.

- **Invariante de transición** o **t-invariante**: conjunto de transiciones que una vez disparadas secuencialmente generan el mismo marcado del cual habían partido, pudiendo disparar la secuencia indefinidamente.

2.2. Reinforcement Learning

Reinforcement Learning (RL) [11] es una rama del *machine learning* donde el aprendizaje se da interactuando con el ambiente. Es un aprendizaje orientado a objetivos, donde al algoritmo no se le enseña qué acciones debe tomar; sino que aprende de las consecuencias de sus acciones.

Este tipo de aprendizaje se complementa con los ya estudiados anteriormente como se ve en la siguiente Fig.4:

Figura 4: Tipos de aprendizajes.

El Reinforcement Learning entonces, intentará hacer aprender a la máquina basándose en un esquema de "premios y castigos" en un entorno en donde hay que tomar acciones y que está afectado por múltiples variables que cambian con el tiempo [3].

2.2.1. Componentes del RL

- *El agente:* Es el modelo que queremos entrenar para que aprenda a tomar decisiones
- *Ambiente:* Es el entorno en donde interactúa y "se mueve" el agente. El ambiente contiene las limitaciones y reglas posibles en cada momento.
- *Acción:* Son las posibles acciones que puede tomar en un momento determinado el Agente (cambiar de estado).

- *Estado:* Es el indicador de cómo se encuentran los diversos elementos que componen el ambiente en un momento dado.
- *Recompensas:* Cada acción tomada por el Agente tendrá como consecuencia un refuerzo positivo o negativo que orientará al Agente hacia la forma correcta de comportarse.
- *Política:* Define el comportamiento del agente en el ambiente.

Figura 5: Interacción entre el agente y el entorno.

El agente se encuentra en un primer momento en un "estado inicial" y luego realiza una acción, lo cual genera que esto influya en el ambiente, luego de esto el agente obtiene dos cosas: Un nuevo estado y la recompensa. Si esta ultima es negativa el agente actuara de forma distinta frente a dicha situación y si es positiva reforzara este comportamiento. Esta interacción se ve mas claramente en la Fig.5.

Al finalizar la instancia de aprendizaje todo el conocimiento aprendido es almacenado en la polıtica.

Figura 6: Interacción entre el agente y el entorno.

Se debe lograr un equilibrio entre la recompensa y el agente, dado que si el agente recibe una recompensa lo suficientemente alta el mismo siempre realizará la misma acción no consiguiendo generalizar el sistema. Fig.6.

2.2.2. Tipos de algoritmos de RL

Existen dentro del aprendizaje por refuerzo ciertos algoritmos a emplear, algunos de ellos son:

2.2.2.1 Q-Learning(Value Learning):

El objetivo principal al entrenar el modelo a través de las simulaciones es ir completando una matriz de Políticas de manera que las decisiones que tome nuestro agente obtengan "la mayor recompensa" evitando el sobreajuste.

- A la política se la denomina Q.
- Q(estado, acción) nos indica el valor de la política para un estado y una acción determinados.

Para completar la matriz de políticas se utiliza la ecuación de Bellman Fig.7.

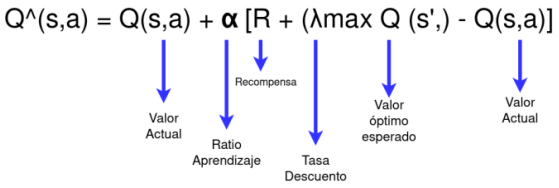


Figura 7: Ecuación de Bellman.

Esta ecuación determina cómo se irán actualizando las políticas Qˆ(s,a), en base a su valor actual más una recompensa recibida como consecuencia de dicha acción. Hay dos ratios que afectan a la manera en que influye esa recompensa: el ratio de aprendizaje, que regula "la velocidad" en la que se aprende, y la "tasa de descuento" que tendrá en cuenta la recompensa a corto o largo plazo.

2.2.2.2 Policy learning

Este método de aprendizaje por refuerzo tiene como principal diferencia a Value Learning (Q-Learning) que en esta última se busca tener una Red Neuronal/Política que aprenda a aproximar la función Q [1], para obtener un valor $Q_(s, a)$ de un estado dada una acción, y luego usamos este valor para inferir cuál es la mejor acción a tomar, esta es nuestra política.

Por otra parte **Policy Learning** busca directamente aprender la política pudiendo usar una NN o, en casos más simples, una matriz para luego poder alimentar la misma con una entrada y como salida se tendrá qué acción debemos tomar. Esto simplifica la situación, ya que para obtener la acción a tomar, es decir la acción que maximizar´a la recompensa dado un estado, simplemente se debe muestrear desde la función de política sin necesidad de realizar numerosas valuaciones.

En resumen en **Value Learning** se aproxima una función Q y se la usa para inferir la política óptima, mientras que en **Policy Learning** se optimiza la política de forma directa.

La política en esta metodología será alimentada con el estado actual, y como salida tendrá las probabilidades correspondientes a cada acción posible.

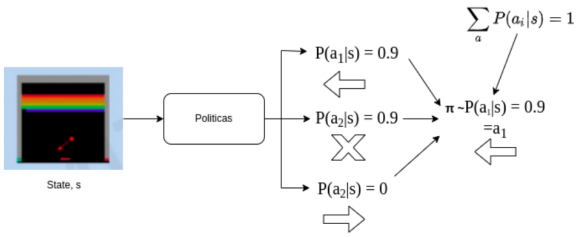


Figura 8: Policy gradient.

En la figura 8 se tiene como ejemplo las probabilidades de los distintos tipos de acciones que puede tomar el agente (la barra) con el fin de pegarle a la bola roja. Aquí la opción óptima es la de moverse a la izquierda, pero dado que esta es una distribución de probabilidad, la vez que se muestree podría

llegar a elegirse quedarse en el lugar (acción 2) dado que también cuenta con un valor no nulo de probabilidad.

Este método de RL puede trabajar con valores continuos. Siguiendo con el ejemplo previo, podría no solo obtenerse la dirección a la cual moverse, sino también la velocidad con la cual hacerlo. Esto puede realizarse visualizando la salida como una distribución probabilística dependiendo de cual se adapte más a la situación. En el ejemplo siguiente se toma una distribución gaussiana.

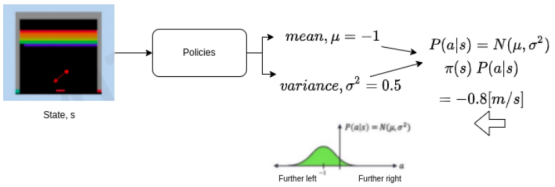


Figura 9: Policy gradient.

En la Fig.9 puede observarse que el agente debe moverse hacia la izquierda con una velocidad cercana a -1 m/s. Al muestrear en esta distribución, se puede notar que la velocidad en concreto que tomará el agente es de 0.8 m/s hacia la izquierda. De esta forma también se puede ver que a pesar de que la media de esta distribución es -1, no estamos limitados a ese número exacto.

Todo esto abre la posibilidad para escenarios donde puede llegar a tenerse un sin fin de acciones posibles a tomar.

2.3. Redes de Bayes

2.3.1. La estadística de Bayes

Es necesario definir ciertos términos que se utilizan con frecuencia en el manejo de la probabilidad Bayesiana. Estos son:

- **Probabilidad previa o a priori:** Es la probabilidad que se tiene de que cierto modelo sea cierto, antes de obtener datos a través de observaciones. Es necesario advertir que las observaciones cambian lo que sabemos del modelo.
- **Probabilidad posterior o a posteriori:** Es la probabilidad de que el modelo sea cierto después de las observaciones.
- **Variables aleatorias:** Representa una "parte"del mundo cuyo estado se desconoce. Se representan en mayúscula y sus posibles valores en minúscula. Las v.a pueden ser de distintos tipos:
  - Booleanas: El conjunto de valores posible esta formado solo por dos valores: $true$ (verdadero) y $false$ (falso).
    - Notación: a y $\neg a$ son equivalentes a $A = true$ y $A = false$ respectivamente.
  - Discretas: Sus posibles valores componen un conjunto discreto. Incluyen a las booleanas.
  - Continuas: El conjunto de valores posibles es un conjunto no numerable.
- **Proposiciones:** Usando las conectivas proposicionales y las variables, podemos expresar proposiciones.
  - Conectivas: $\vee(or)$, $\wedge(and)$, $\sim(nop)(\neg)$
  - Se asignan probabilidades a las proposiciones para expresar el grado de creencia que se tiene en las mismas.

2.3.2. Probabilidad: definición axiomática

Una función de probabilidad es una función definida en el conjunto de proposiciones (respecto de un conjunto dado de variables aleatorias), verificando las siguientes propiedades:

- $0 \leq P(a) \leq 1$ para toda proposición a.
- $P(true) = 1$ y $P(false) = 0$
  - donde true y false representan a cualquier proposición redundante o insatisfacible, respectivamente.
- $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$, para cualquier par de proposiciones a y b.

El calculo de probabilidades se construye sobre estos tres axiomas. Por ejemplo:

- $P(\neg a) = 1 - P(a)$
- $P(a \vee b) = P(a) + P(b)$, si a y b son independientes.
- $\sum_{i=0}^{n} P(D = d_i) = 1$, siendo D una variable aleatoria y $di, i = 1, 2, \ldots, n$ sus posibles valores.

2.3.3. Teorema de Bayes

Sea $\{A_1, A_2, \ldots, A_i, \ldots, A_n\}$ un conjunto de sucesos mutuamente excluyentes y exhaustivos tales que la probabilidad de cada uno de ellos es distinta de cero ($P[A_i] \neq 0$ para $i = 1, 2, \ldots, n$). Si B es n suceso cualquiera del que se conocen las probabilidades condicionales $P(B|A_i)$ entonces la probabilidad $P(A_i|B)$ viene dada por la expresión:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

donde:

- $P(A_i)$ son las probabilidades a priori.
- $P(B|A_i)$ son las probabilidades de B en la hipótesis A_i.
- $P(A_i|B)$ son las probabilidades a posteriori.

2.3.4. Fórmula de Bayes

Con base en la definición de probabilidad condicionada se obtiene la Fórmula de Bayes, también conocida como Regla de Bayes:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^{n} P(B|A_k)P(A_k)} \cdots$$

Esta fórmula nos permite calcular la probabilidad condicional $P(A_i|B)$ de cualquiera de los eventos A_i dado B.

2.3.5. Redes Bayesianas

Es un grafo que representa la relación entre las variables de un problema, permitiendo una representación compacta y son de ayuda en la toma de decisiones [4]. Están conformadas por:

- *Nodos:* círculos que representan una variable aleatoria.
- *Flechas:* relaciones causales entre las variables aleatorias (conexiones entre nodos).

Las redes Bayesianas nos permiten realizar inferencias[2], existen tres tipos:

- Razonamiento de predicción.
- Razonamiento de diagnostico.
- Razonamiento de justificación.

2.3.5.1 Razonamiento de predicción

Es explicado por la cadena de causalidad.

Se usa la causa (información dada) para inferir algo

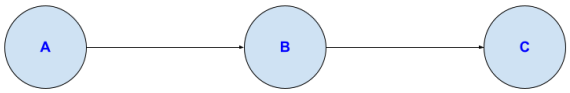


Figura 10: Cadena de Causalidad.
Red de Bayes con tres nodos.
Nodo A causa Nodo B y Nodo B causa Nodo C.

Aquí se utiliza la información dada en la causa para poder inferir algo. Si se tiene conocimiento sobre A, entonces se puede inferir le probabilidad de que ocurra B, y a su vez, la probabilidad de que ocurra algo sobre C.

El evento A causa el evento B y el evento B causa el evento C

Por lo que si se conoce el se puede inferir C y no es necesario conocer A. Esto se llama **bloqueo**, el evento B *bloquea* el evento A.

2.3.5.2 Razonamiento de diagnostico

Es explicado por una red de causa común.

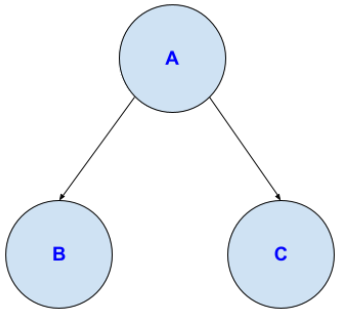


Figura 11: Red de causa común.
Nodo A causa Nodo B y también causa Nodo C.

2.3.5.3 Razonamiento de justificación

Dos o más eventos son causales de un tercero. Estas redes presentas características como:

- Independencia condicional
- Dependencia condicional

Entonces:

- Si A es verdadero es posible inferir la probabilidad de que sus padres A y B también sean verdaderos.
- Se puede explicar que ha causado C.

Cuando esto ocurre, si se conoce que uno de los eventos es verdadero, se puede inferir que la probabilidad que otro evento causante sea verdadero es menor. Esto es interesante dado que A y B son independientes.

En este tipo de redes también puede ocurrir un bloqueo. Existen casos donde un nodo puede bloquear varios nodos, esto es conocido con D-separación.

La figura 12, aclara estas ideas con un ejemplo, muestra si un vuelo puede estar retrasado; el retraso puede ser debido al mal tiempo o al exceso de trafico aéreo. La red esta compuesta por un nodo para cada variable y las relaciones entre estos (nodo Vr vuelo, nodo Mt mal clima, nodo Ta trafico aéreo). Las relaciones representada por flechas son causales donde los padres causan a los hijos.

En el ejemplo el retraso puede deberse al mal tiempo o al exceso de trafico aéreo y también por ambos.

Estas relaciones son probabilísticas y la fuerza de la causa depende del valor de la probabilidad que las relaciona. Por lo que en las aplicaciones, cuando sea posible conviene, que los eventos resultantes discretos sean mutuamente excluyentes, es decir tengamos solo un valor del grupo posible. Esto facilita el cálculo.

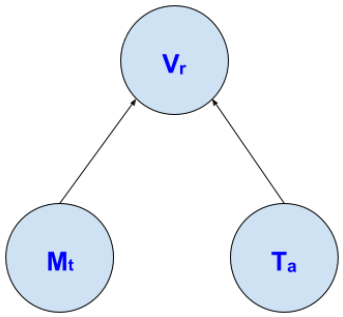


Figura 12: Red de efecto común.
Nodo A y Nodo B son causa de Nodo C.

Para el desarrollo de este ejemplo se supone a todas las variables aleatorias como binarias (V,F). Ahora, con esta información, es posible saber si el vuelo esta retrasado o no.

Para esto es necesario construir una tabla de probabilidades condicionales para saber que valores son necesarios utilizar en el calculo de la probabilidad final, dependiendo del mal tiempo y del exceso de trafico.

Valores de los padre		**Probabilidad para vuelos retrasados**	
Mal tiempo	Trafico aereo	Retrasado=Verdadero	Retrasado=Falso
v	v	$p(Vr \cap Mt \cap Ta)$	$p(\sim Vr \cap Mt \cap Ta)$
v	f	$p(Vr \cap Mt \cap \sim Ta)$	$p(\sim Vr \cap Mt \cap \sim Ta)$
f	v	$p(Vr \cap \sim Mt \cap Ta)$	$p(\sim Vr \cap \sim Mt \cap Ta)$
f	f	$p(Vr \cap \sim Mt \cap \sim Ta)$	$p(\sim Vr \cap \sim Mt \cap \sim Ta)$

Cuadro 1: Tabla de probabilidades condicionales para las variables Vr, Mt y Ta

La tabla de probabilidades condicionales debe construirse siempre para todos los nodos hijos de la red. Los nodos padres tendrán una tabla de probabilidad a priori.

3. Desarrollo

3.1. Curso sklearn master

En esta primera instancia, se llevó a cabo el tutorial sklearn-master, a modo de tener una primera aproximación al machine learning o aprendizaje automático.

El machine learning cubre principalmente tres áreas:

- *Aprendizaje supervisado*: A grandes rasgos, a partir de ciertos datos categorizables (de los cuales se conoce su categoría), se obtienen modelos matemáticos que permiten estimar, dada una entrada de categoría desconocida, a cuál pertenece.
- *Aprendizaje no supervisado*: En este caso, no se busca inferir a qué categoría pertenece algo, si no más bien encontrar características comunes entre los datos de entrada.
- *Aprendizaje por refuerzo*: Esta última área, tiene un enfoque distinto, ya que lo que busca es tomar decisiones de acciones a realizar a futuro, en base a una recompensa obtenida al haber realizado dicha acción en el pasado.

El curso, hace énfasis en las dos primeras áreas, con teoría, ejemplos y ejercicios. Los notebooks resueltos se encuentran en `https://github.com/los2nicos/PPS_y_PI/tree/main/tutorial-sklearn`.

3.2. Curso MIT: Introduccion to deep learning

De este curso[1] se prestó atención a las partes referidas a Reinforcement Learning o Aprendizaje por refuerzo, Deep Learning o aprendizaje profundo, para finalmente confluir en Deep Reinforcement Learning. También se realizaron algunos ejercicios del curso.

3.3. Componentes de software

A continuación se detallan los paquetes de software mas importantes que serán utilizados en el desarrollo del sistema del presente trabajo.

3.3.1. TensorFlow

TensorFlow es una es una plataforma de código abierto que ofrece herramientas para aprendizaje automático. Permite trabajar con tensores (matrices multidimensionales con un tipo uniforme), variables, desarrollar redes neuronales y bucles de aprendizaje para las mismas, etc.

Principalmente el manejo de tensores resultará de utilidad para el desarrollo del presente trabajo.

3.3.2. Matplotlib

Matplotlib es una biblioteca para la generación de gráficos en dos dimensiones, a partir de datos contenidos en listas o arrays en el lenguaje de programación Python.

Estos gráficos permitirán analizar el comportamiento del sistema.

3.3.3. NumPy

NumPy es el paquete más utilizado para la computación científica en Python[10]. Esta librería provee objetos de arreglos multidimensionales, de los que derivan las matrices, y una amplia variedad de rutinas para operar rápidamente con arreglos que incluyen operaciones matemáticas, lógicas, manipulación de forma y tamaño, ordenamiento, I/O, álgebra lineal básica, y más.

Como se detalló en el marco teórico, las Redes de Petri se pueden representar como un conjunto de matrices, por lo que es fundamental contar con una herramienta que permita manipularlas eficientemente.

3.3.4. re — Operaciones con Expresiones Regulares

Este módulo proporciona operaciones de coincidencia de expresiones regulares [8].

El manejo de expresiones regulares permitirá identificar cuando se complete un invariante de transición, al buscar en un archivo de log de actividades de la Red de Petri, la secuencia de disparo de transiciones que conforman el invariante.

El poder llevar un control de los invariantes completados permitirá comparar este numero con los requerimientos del usuario.

3.3.5. pyAgrum

PyAgrum es una librería desarrollada en C++ y Python dedicada al manejo de Redes Bayesianas y otros modelos gráficos de probabilidad. Permite crear estos modelos y realizar las operaciones necesarias para el cálculo de probabilidades de manera rápida y eficiente.

Esta librería nos permite realizar las gráficas de la Red de Bayes y de las tablas de probabilidad de cada nodo.

3.4. Condiciones

En esta sección detallaremos las condiciones que van a formar parte de la recompensa retornada por el ambiente al disparar una transición. Las mismas, son las siguientes:

- *Recompensa por poder dispararse*: Es la más simple de todas. Si dado un estado, puedo disparar una transición, va a tener esta componente. Sirve como caso base de prueba y equiponderar las transiciones disparables al deshabilitar el resto de las recompensas.
- *Recompensa por uso de recursos*: Si la transición disparada representa el uso de recurso, la recompensa otorgada por esta componente es directamente proporcional a la cantidad de recursos utilizados.
- *Recompensa por liberación de recursos*: Si la transición disparada representa la liberación de un recurso, la recompensa de esta componente es directamente proporcional a la cantidad de recursos liberados.
- *Recompensa por la liberación de recursos requeridos en la inmediatez*: Es un plus que se agrega a la recompensa anterior, en el caso en que el recurso liberado, sea necesario. i.e. permita que una nueva transición que tenga como plaza de entrada la que contenga el recurso liberado, se sensibilice por el disparo de la primer transición.
- *Recompensa por transiciones sensibilizadas*: Otorga recompensas si el disparo de la transición produce la sensibilización de otras que no lo estaban antes del disparo.
- *Recompensa por invariante completado*: La más importante de todas porque simboliza la finalización de un proceso o circuito. Otorga recompensa si la transición disparada, es la última necesaria para terminar un invariante. Notar el caso particular de que no sólo depende del disparo de una transición, si no de que se hayan disparado previamente el resto de transiciones pertenecientes a dicho invariante.

La recompensa obtenida será una suma ponderada de la siguiente forma:

$$R_t = \sum_{i=1}^{n} \alpha_i * R_i \tag{8}$$

donde:

- R_i es la recompensa iésima, y su valor representa la cantidad de veces que se cumplió esa condición,
- α_i es el peso o valor numérico de cumplir la recompensa iésima una vez y
- n es el número de condiciones que otorgan recompensas

3.5. Micocubo

Durante la primer etapa del trabajo, el objetivo fue maximizar el número de condiciones cumplidas. La primer aproximación al mismo, consistió en utilizar Q-learning.

A diferencia del trabajo final en el que se basa este trabajo, donde las tablas de Q-learning, se utilizaban en caso de conflictos, la idea fue extender el algoritmo para controlar todas las transiciones, independientemente del estado de la red.

Tras encarar este problema como un problema de Q learning tradicional, surgen un par de problemas:

- *Tamaño de la tabla Q*
- *Tiempo en converger*

3.5.1. Tamaño de la tabla Q

Una tabla Q tiene dimensiones n_states x $n\_actions$, donde $n\_actions$ es el número de transiciones a disparar, y n_states es la cardinalidad del conjunto de marcados posibles de la red de petri, dado un estado inicial.

Luego, para determinar los estados posibles, se desarrolló una búsqueda por amplitud (cuyo nodo raíz, es el estado inicial) disparando todas las transiciones y generando un nodo a explorar en caso de, poder disparar la transición i.e. environ.can_shoot(T_i) == $true$ y que el nodo no haya sido explorado previamente.

3.5.2. Sesgo inicial

Al haber redes grandes en las cuales, dado un estado, solo un subconjunto pequeño del conjunto de transiciones es factible de disparar, inicializar la tabla Q en ceros, lleva a que el sistema demore un tiempo considerable en poder aprender a disparar unas pocas transiciones.

Surge entonces la idea de crear una tabla sesgo, esto es, una tabla no inicializada en ceros.

Como el problema principal, es que, dado un estado, la mayoría de las transiciones no están sensibilizadas, se optó por aprovechar la búsqueda por amplitud del punto anterior, para completar una estructura que vinculara, estado - transiciones sensibilizadas (en dicho estado).

Si se superpusieran estos vectores, se obtendría una tabla en la cual $Q(estado, transicion) = False$ si la transición en ese estado no se puede disparar y $Q(estado, transicion) = True$ si la transición está sensibilizada.

El segundo paso consiste en darle un valor más útil a estos resultados. Para ello, se hizo un método $get\_recursive\_rewards()$ en el cual, bajo una profundidad parametrizable, evalúa la máxima cantidad de recompensa que se puede obtener desde un estado, disparando primero la transición t_i de forma que:

$$\text{Q(estado,t}_i) = \begin{cases} r > 0 & si\ t_i\ es\ disparable \\ r = 0 & en\ caso\ contrario \end{cases} \tag{9}$$

Superponiendo estos vectores se obtiene una Q_value_table, que no permite disparar transiciones no sensibilizadas y en la que al elegir la transición a disparar, elige la que mayor recompensa otorgue en el corto plazo.

3.5.3. Haciendo la tabla probabilística

Si bien es cierto que eligiendo siempre la transición cuya recompensa es la mayor, la recompensa global debería tender a ser la mayor, esto, en el escenario de representar un proceso productivo, podría llevar a realizar siempre el mismo proceso, condenando a los demás, a no completarse nunca.

Para ello, se generó una tabla similar a la anterior, pero en la que se dividió cada elemento de Q_{ij} por la suma de la fila perteneciente al estado, esto es $sum(Q_i)$. De esta forma:

$$Qprob_{ij} = \frac{Q_{ij}}{\sum_k Q_{ik}} \tag{10}$$

Con esta nueva tabla, se puede generar un número aleatorio, utilizando la fila Q_i como distribución de probabilidad y así elegir una transición con un peso ponderado a la recompensa de dispararla respecto de disparar las demás.

Referencias

[1] Alexander Amini. Deep reinforcement learning. mit, 6:s191, 2019.

[2] Nicolas Arrioja Landa Cosio. *Inteligencia Artificial, Sistemas Inteligentes con C#*. USERS, 2011.

[3] Juan Ignacio Bagnato. Aprendizaje por refuerzo. `http://aprendemachinelearning.com/aprendizaje-por-refuerzo/.`

[4] José Luis Iglesias Feria. Ia probabilidad - redes bayesianas. `https://www.youtube.com/playlist?list=PLYWD-VqrD5BCr-QeESSOvpEqkAcvQOrQ0`, 2018.

[5] Micolini Orlando, Cebollada Marcelo, Eschoyez Maximiliano, Ventre Luis O., and Schild Marcelo. Ecuación de estado generalizada para redes de petri no autónomas y con distintos tipos de arcos. *XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016)*, 2016.

[6] Mario A. Perez, Analía Perez Hidalgo, and Elisa Perez Berenguer. Introducción a los sistemas de control y modelo matemático para sistemas lineales invariantes en el tiempo. Universidad Nacional de San Juan. Facultad de Ingeniería. Departamento de Electrónica y Automática, 2007.

[7] J.L. Peterson. *Petri Net Theory and the Modeling of Systems*. Independently Published, 2019.

[8] Python.org. 3.11.0 documentation.la biblioteca estándar de python.servicios de procesamiento de texto.re — operaciones con expresiones regulares. `https://docs.python.org/es/3/library/re.html#id1`, 2022.

[9] Ian Sommerville. *Ingenieria del software*. Pearson Educacion, 7 edition, 2005.

[10] the NumPy community. *NumPy User Guide. Release 1.23.0*. NumPy, 2022.

[11] Simonini Thomas. Deep reinforcement learning course. `https://simoninithomas.github.io/deep-rl-course/`.