Reinforcement Learning

Norah Jones

Invalid Date

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

```
1 + 1
```

```
[1] 2
```

1 Introduction

2 Activate the Core Packages

```
library(bibtex)
library(tidyverse) ## Brings in a core of useful functions
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

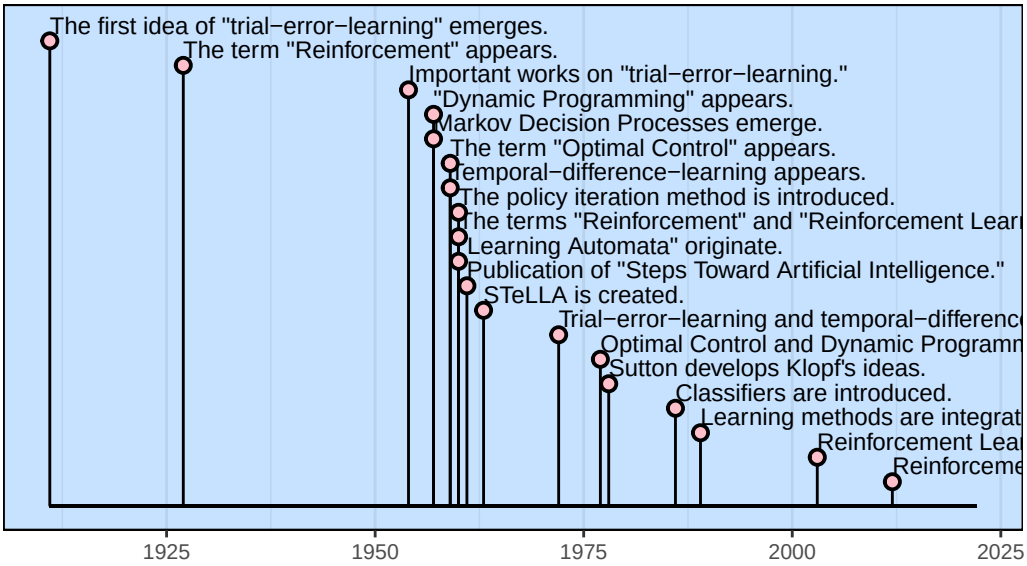
```
library(gt)        ## Tables
## Specific packages
library(milestones)
## Initialize defaults
## Initialize defaults
column <- lolli_styles()

data <- read_csv(col_names=TRUE, show_col_types=FALSE, file='rl_time_line.csv')
```

```
## Sort the table by date
data <- data |>
  arrange(date)

## Build a table
gt(data) |>
  #cols_hide(columns = event) |>
  tab_style(cell_text(v_align = "top"),
            locations = cells_body(columns = date)) |>
  tab_source_note(source_note = "Source: Sutton and Barto (2018)")
```

date	event	reference
1911	The first idea of "trial-error-learning" emerges.	Thorndike, E. L. (1911).
1927	The term "Reinforcement" appears.	Pavlov, P. I. (1927). Co
1954	Important works on "trial-error-learning."	Minsky, M. L. (1954). T
1957	"Dynamic Programming" appears.	Bellman, R. E. (1957).?I
1957	Markov Decision Processes emerge.	Bellman, R. (1957). A M
1959	The term "Optimal Control" appears.	NA
1959	Temporal-difference-learning appears.	Samuel, A. L. (1959). Sc
1960	The policy iteration method is introduced.	Howard, R. A. (1960). D
1960	The terms "Reinforcement" and "Reinforcement Learning" are used.	Waltz, M. D., Fu, K. S.
1960	"Learning Automata" originate.	Tsetlin, M. L. (1973). A
1961	Publication of "Steps Toward Artificial Intelligence."	Minsky, M. L. (1961). St
1963	STeLLA is created.	Andreae, J. H. (1963). S
1972	Trial-error-learning and temporal-difference-learning are combined.	Klopf, A. H. (1972). Bra
1977	Optimal Control and Dynamic Programming are connected.	Werbos, P. J. (1977). Ac
1978	Sutton develops Klopf's ideas.	Sutton, R. S. (1978a). L
1986	Classifiers are introduced.	Holland, J. H. (1986). E
1989	Learning methods are integrated.	Watkins, C. J. C. H. (19
2003	Reinforcement Learning in economics.	Camerer, C. (2003). Beh
2012	Reinforcement Learning and Games.	Nowe, A., Vrancx, P., De

Source: Sutton and Barto (2018)

```
## Adjust some defaults
column$color <- "pink"
column$size  <- 15
column$source_info <- "Source: Sutton and Barto (2018)"

## Milestones timeline
milestones(datatable = data, styles = column)
```



Source: Sutton and Barto (2018)

2.1 EJERCICIO 1

En el aprendizaje reforzado un agente aprende a tomar decisiones (acciones) a través de la interacción con su entorno y recibiendo recompensas o castigos en función de las mismas, a diferencia del aprendizaje supervisado, ya que en este tipo de aprendizaje automático, un modelo se entrena utilizando un conjunto de datos que incluye tanto las entradas como las salidas correspondientes (etiquetas). es decir, consiste en aprender a partir
de un conjunto de ejemplos ya etiquetados y proporcionados por un supervisor externo con conocimientos. Por lo que en este tipo de aprendizaje Cada ejemplo describe una situación especifica, y además existe una etiqueta que indica la acción adecuada que el sistema debe tomar en esa situación, El objetivo de este tipo de aprendizaje es que el
sistema generalice sus respuestas para que actúe correctamente en situaciones que no están presentes en el conjunto de entrenamiento. Por otra parte El aprendizaje por refuerzo también es diferente de lo que los investigadores del aprendizaje automático llaman aprendizaje no supervisado, que generalmente consiste en
encontrar estructuras ocultas en conjuntos de datos no etiquetados y Aunque en parte el aprendizaje por refuerzo es un tipo de aprendizaje no supervisado, en realidad este se centra mas que nada en maximizar una recompensa en lugar de buscar patrones ocultos en los datos.

2.2 EJERCICIO 2

es posible pensar que dicha expresión es una función con la cual se mide el desempeño del sistema bajo diferentes políticas de control dado el estado inicial, es decir, nos ayuda a identificar que acciones fueron buenas y cuales fueron malas, además dicha expresión nos da el valor esperado de cuanta recompensa obtendremos en un futuro al elegir dicha politica dado un estado inicial. Por otra parte, el factor de descuento en la expresión, nos ayuda a comparar las recompensas futuras con las recompensas inmediatas, basicamente nos dice que tan a favor estamos de obtener una recompensa en el estado actual frente a un futuro lejano

2.3 APD

del algoritmo de la programación dinámica se sigue que para este caso particular _ $J_N(x) = \beta^N(x_N)^{1-\gamma}$

luego, para $k = N-1$

$$J_{N-1} = \min_{a \in A(x)} \{\beta^{N-1}(a)^{1-\gamma} + \beta^N(1+r)^{1-\gamma}(x-a)^{1-\gamma}\}$$

derivando con respecto a a obtenemos

$$(1-\gamma)\beta^{N-1}a^{-\gamma} - \beta^N(1+r)^{1-\gamma}(x-a)^{-\gamma}$$

depués igualando a cero

$$(1-\gamma)\beta^{N-1}[a^{-\gamma} - \beta(1+r)^{1-\gamma}(x-a)^{-\gamma}] = 0$$

entonces

$$(\frac{x-a}{a})^{\gamma} = \beta(1+r)^{1-\gamma}$$

$$\frac{x-a}{a} = [\beta(1+r)^{1-\gamma}]^{\frac{1}{\gamma}}$$

3 Summary

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```