Title: State–action–reward–state–action

URL: https://en.wikipedia.org/wiki/State%E2%80%93action%E2%80%93reward%E2%80%93state%E2%80%93action

PageID: 10584297

Categories: Category:1994 in artificial intelligence, Category:Machine learning algorithms, Category:Reinforcement learning

-----

Supervised learning

Unsupervised learning

Semi-supervised learning

Self-supervised learning

Reinforcement learning

Meta-learning

Online learning

Batch learning

Curriculum learning

Rule-based learning

Neuro-symbolic AI

Neuromorphic engineering

Quantum machine learning

Classification

Generative modeling

Regression

Clustering

Dimensionality reduction

Density estimation

Anomaly detection

Data cleaning

AutoML

Association rules

Semantic analysis

Structured prediction

Feature engineering

Feature learning

Learning to rank

Grammar induction

Ontology learning

Multimodal learning
Apprenticeship learning
Decision trees
Ensembles Bagging Boosting Random forest
Bagging
Boosting
Random forest
k -NN
Linear regression
Naive Bayes
Artificial neural networks
Logistic regression
Perceptron
Relevance vector machine (RVM)
Support vector machine (SVM)
BIRCH
CURE
Hierarchical
k -means
Fuzzy
Expectation–maximization (EM)
DBSCAN
OPTICS
Mean shift
Factor analysis
CCA
ICA
LDA
NMF
PCA
PGD
t-SNE
SDL
Graphical models Bayes net Conditional random field Hidden Markov
Bayes net
Conditional random field
Hidden Markov
RANSAC

k -NN

Local outlier factor

Isolation forest

Autoencoder

Deep learning

Feedforward neural network

Recurrent neural network LSTM GRU ESN reservoir computing

LSTM

GRU

ESN

reservoir computing

Boltzmann machine Restricted

Restricted

GAN

Diffusion model

SOM

Convolutional neural network U-Net LeNet AlexNet DeepDream

U-Net

LeNet

AlexNet

DeepDream

Neural field Neural radiance field Physics-informed neural networks

Neural radiance field

Physics-informed neural networks

Transformer Vision

Vision

Mamba

Spiking neural network

Memtransistor

Electrochemical RAM (ECRAM)

Q-learning

Policy gradient

SARSA

Temporal difference (TD)

Multi-agent Self-play

Self-play

Active learning

Crowdsourcing

Human-in-the-loop

Mechanistic interpretability

RLHF

Coefficient of determination

Confusion matrix

Learning curve

ROC curve

Kernel machines

Bias–variance tradeoff

Computational learning theory

Empirical risk minimization

Occam learning

PAC learning

Statistical learning

VC theory

Topological deep learning

AAAI

ECML PKDD

NeurIPS

ICML

ICLR

IJCAI

ML

JMLR

Glossary of artificial intelligence

List of datasets for machine-learning research List of datasets in computer vision and image processing

List of datasets in computer vision and image processing

Outline of machine learning

v

t

e

State–action–reward–state–action (SARSA) is an algorithm for learning a Markov decision process policy, used in the reinforcement learning area of machine learning . It was proposed by Rummery and Niranjan in a technical note [1] with the name "Modified Connectionist Q-Learning" (MCQ-L). The alternative name SARSA, proposed by Rich Sutton , was only mentioned as a footnote.

This name reflects the fact that the main function for updating the Q-value depends on the current state of the agent " S 1 ", the action the agent chooses " A 1 ", the reward " R 2 " the agent gets for choosing this action, the state " S 2 " that the agent enters after taking that action, and finally the next action " A 2 " the agent chooses in its new state. The acronym for the quintuple (S t , A t , R t+1

, S_{t+1}, A_{t+1}) is SARSA. [2] Some authors use a slightly different convention and write the quintuple (S_t, A_t, R_t, S_{t+1}, A_{t+1}), depending on which time step the reward is formally assigned. The rest of the article uses the former convention.

Algorithm

A SARSA agent interacts with the environment and updates the policy based on actions taken, hence this is known as an on-policy learning algorithm . The Q value for a state-action is updated by an error, adjusted by the learning rate α. Q values represent the possible reward received in the next time step for taking action a in state s , plus the discounted future reward received from the next state-action observation.

Watkin's Q-learning updates an estimate of the optimal state-action value function Q^{*} {\displaystyle Q^{*}} based on the maximum reward of available actions. While SARSA learns the Q values associated with taking the policy it follows itself, Watkin's Q-learning learns the Q values associated with taking the optimal policy while following an exploration/exploitation policy.

Some optimizations of Watkin's Q-learning may be applied to SARSA. [3]

Hyperparameters

Learning rate (alpha)

The learning rate determines to what extent newly acquired information overrides old information. A factor of 0 will make the agent not learn anything, while a factor of 1 would make the agent consider only the most recent information.

Discount factor (gamma)

The discount factor determines the importance of future rewards. A discount factor of 0 makes the agent "opportunistic", or "myopic", e.g., [4] by only considering current rewards, while a factor approaching 1 will make it strive for a long-term high reward. If the discount factor meets or exceeds 1, the Q {\displaystyle Q} values may diverge.

Initial conditions ($Q(S_0, A_0)$)

Since SARSA is an iterative algorithm, it implicitly assumes an initial condition before the first update occurs. A high (infinite) initial value, also known as "optimistic initial conditions", [5] can encourage exploration: no matter what action takes place, the update rule causes it to have higher values than the other alternative, thus increasing their choice probability. In 2013 it was suggested that the first reward r {\displaystyle r} could be used to reset the initial conditions. According to this idea, the first time an action is taken the reward is used to set the value of Q {\displaystyle Q} . This allows immediate learning in case of fixed deterministic rewards. This resetting-of-initial-conditions (RIC) approach seems to be consistent with human behavior in repeated binary choice experiments. [6]

See also

Prefrontal cortex basal ganglia working memory

Sammon mapping

Constructing skill trees

Q-learning

Temporal difference learning

Reinforcement learning

References

v

t

e

History timeline
timeline
Companies
Projects
Parameter Hyperparameter
Hyperparameter
Loss functions
Regression Bias–variance tradeoff Double descent Overfitting
Bias–variance tradeoff
Double descent
Overfitting
Clustering
Gradient descent SGD Quasi-Newton method Conjugate gradient method
SGD
Quasi-Newton method
Conjugate gradient method
Backpropagation
Attention
Convolution
Normalization Batchnorm
Batchnorm
Activation Softmax Sigmoid Rectifier
Softmax
Sigmoid
Rectifier
Gating
Weight initialization
Regularization
Datasets Augmentation
Augmentation
Prompt engineering
Reinforcement learning Q-learning SARSA Imitation Policy gradient
Q-learning
SARSA
Imitation
Policy gradient
Diffusion
Latent diffusion model

Autoregression
Adversary
RAG
Uncanny valley
RLHF
Self-supervised learning
Reflection
Recursive self-improvement
Hallucination
Word embedding
Vibe coding
Machine learning In-context learning
In-context learning
Artificial neural network Deep learning
Deep learning
Language model Large language model NMT
Large language model
NMT
Reasoning language model
Model Context Protocol
Intelligent agent
Artificial human companion
Humanity's Last Exam
Artificial general intelligence (AGI)
AlexNet
WaveNet
Human image synthesis
HWR
OCR
Computer vision
Speech synthesis 15.ai ElevenLabs
15.ai
ElevenLabs
Speech recognition Whisper
Whisper
Facial recognition
AlphaFold

Text-to-image models Aurora DALL-E Firefly Flux Ideogram Imagen Midjourney Recraft Stable Diffusion

Aurora

DALL-E

Firefly

Flux

Ideogram

Imagen

Midjourney

Recraft

Stable Diffusion

Text-to-video models Dream Machine Runway Gen Hailuo AI Kling Sora Veo

Dream Machine

Runway Gen

Hailuo AI

Kling

Sora

Veo

Music generation Riffusion Suno AI Udio

Riffusion

Suno AI

Udio

Word2vec

Seq2seq

GloVe

BERT

T5

Llama

Chinchilla AI

PaLM

GPT 1 2 3 J ChatGPT 4 4o o1 o3 4.5 4.1 o4-mini 5

1

2

3

J

ChatGPT

4

4o

o1

o3

4.5

4.1

o4-mini

5

Claude

Gemini Gemini (language model) Gemma

Gemini (language model)

Gemma

Grok

LaMDA

BLOOM

DBRX

Project Debater

IBM Watson

IBM Watsonx

Granite

PanGu-Σ

DeepSeek

Qwen

AlphaGo

AlphaZero

OpenAI Five

Self-driving car

MuZero

Action selection AutoGPT

AutoGPT

Robot control

Alan Turing

Warren Sturgis McCulloch

Walter Pitts

John von Neumann

Claude Shannon

Shun'ichi Amari

Kunihiko Fukushima

Takeo Kanade

Marvin Minsky

John McCarthy
Nathaniel Rochester
Allen Newell
Cliff Shaw
Herbert A. Simon
Oliver Selfridge
Frank Rosenblatt
Bernard Widrow
Joseph Weizenbaum
Seymour Papert
Seppo Linnainmaa
Paul Werbos
Geoffrey Hinton
John Hopfield
Jürgen Schmidhuber
Yann LeCun
Yoshua Bengio
Lotfi A. Zadeh
Stephen Grossberg
Alex Graves
James Goodnight
Andrew Ng
Fei-Fei Li
Alex Krizhevsky
Ilya Sutskever
Oriol Vinyals
Quoc V. Le
Ian Goodfellow
Demis Hassabis
David Silver
Andrej Karpathy
Ashish Vaswani
Noam Shazeer
Aidan Gomez
John Schulman
Mustafa Suleyman
Jan Leike
Daniel Kokotajlo

François Chollet

Neural Turing machine

Differentiable neural computer

Transformer Vision transformer (ViT)

Vision transformer (ViT)

Recurrent neural network (RNN)

Long short-term memory (LSTM)

Gated recurrent unit (GRU)

Echo state network

Multilayer perceptron (MLP)

Convolutional neural network (CNN)

Residual neural network (RNN)

Highway network

Mamba

Autoencoder

Variational autoencoder (VAE)

Generative adversarial network (GAN)

Graph neural network (GNN)

Category