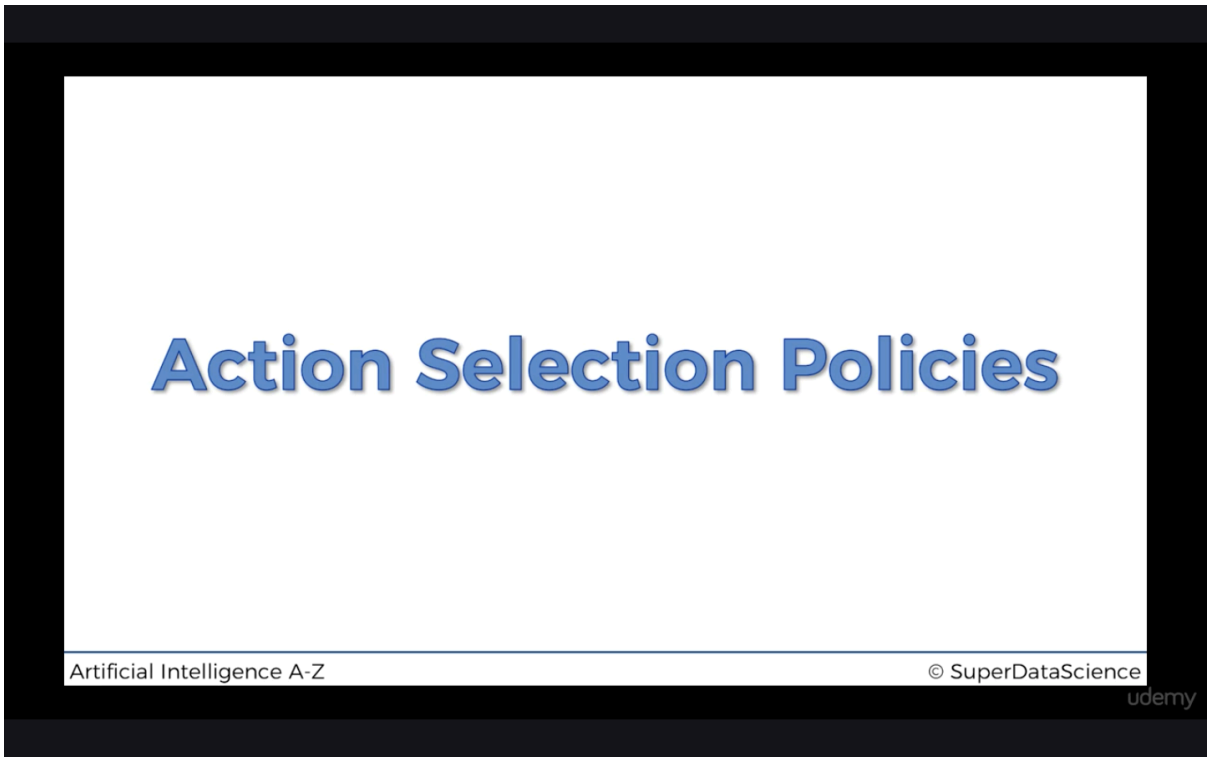
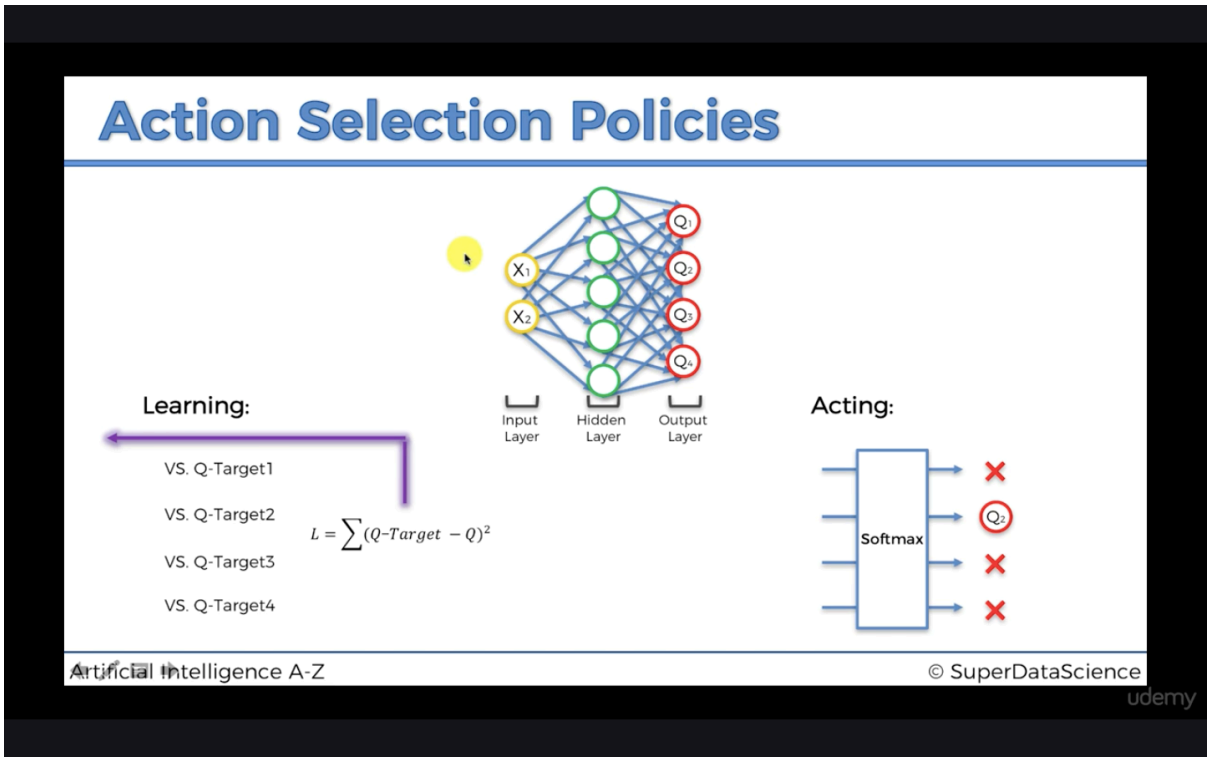
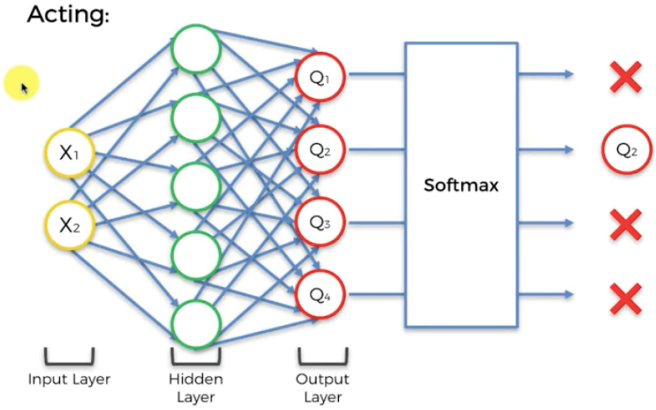
Action selection policies





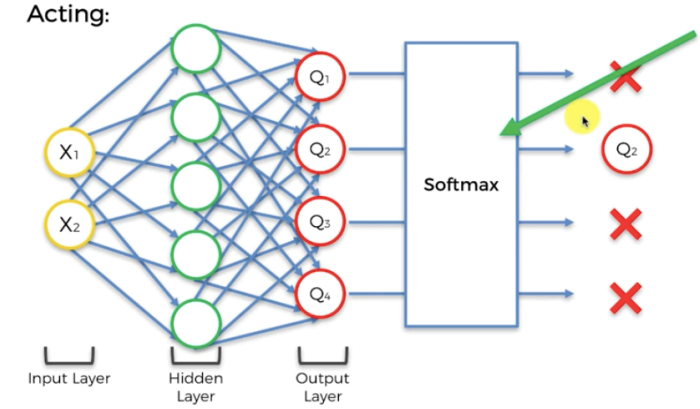
Action Selection Policies

Acting:



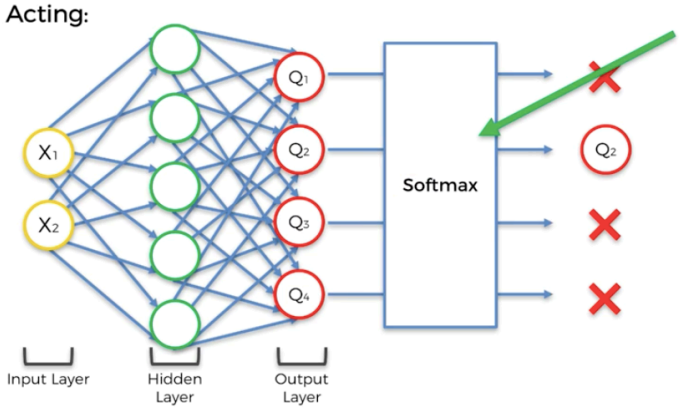
Action Selection Policies

Acting:



Action Selection Policies

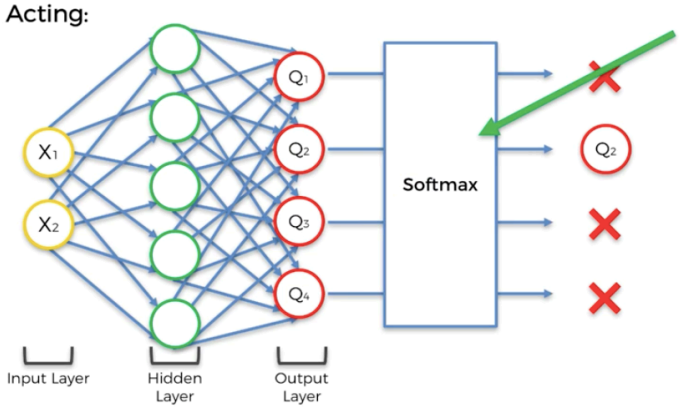
Acting:



Action Selection:

Action Selection Policies

Acting:

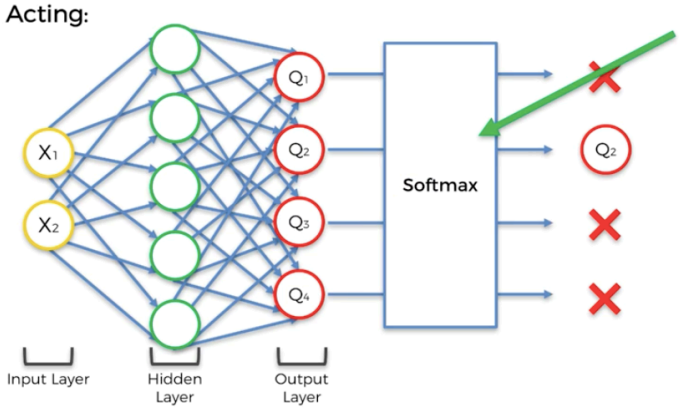


Action Selection:

- ε-greedy

Action Selection Policies

Acting:

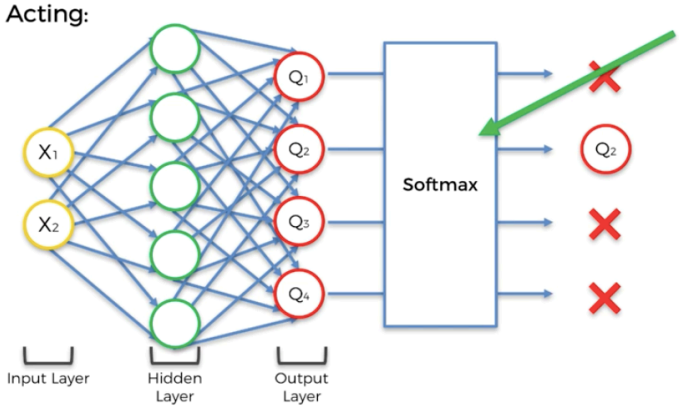


Action Selection:

- ε-greedy
- ε-soft (1-ε)

Action Selection Policies

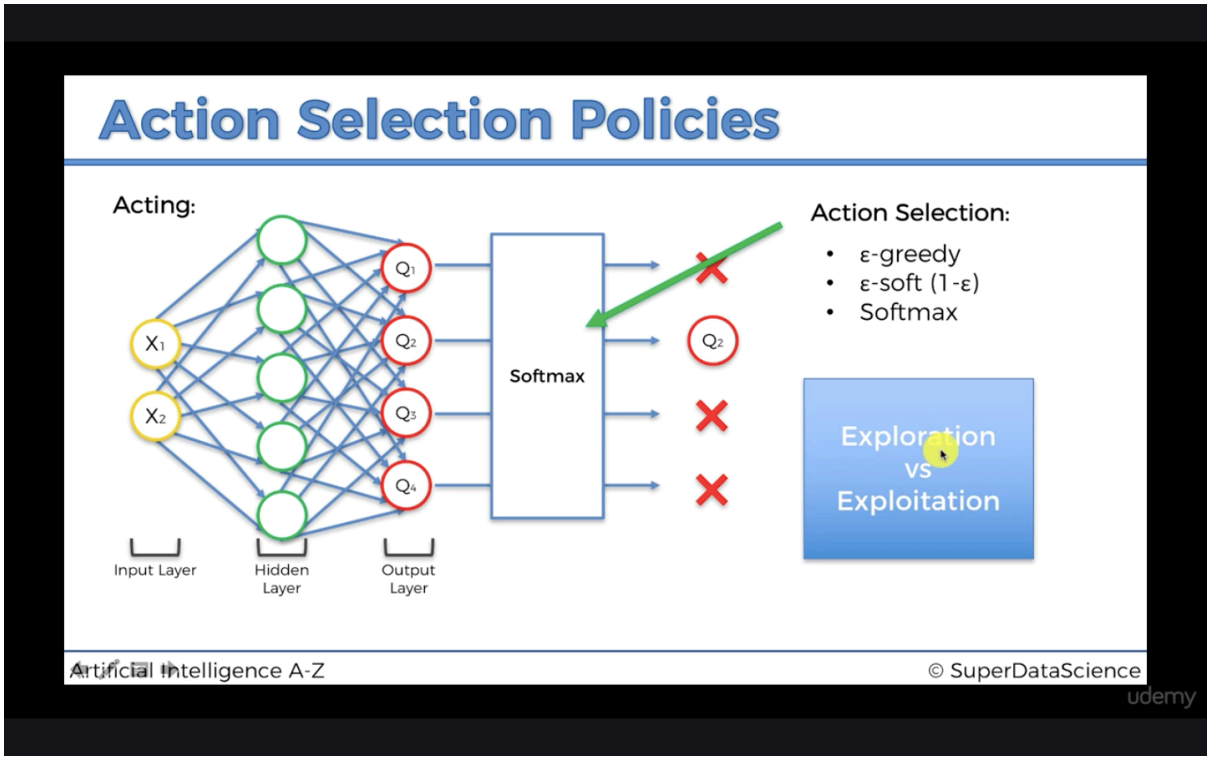
Acting:



Action Selection:

- ε-greedy
- ε-soft (1-ε)
- Softmax

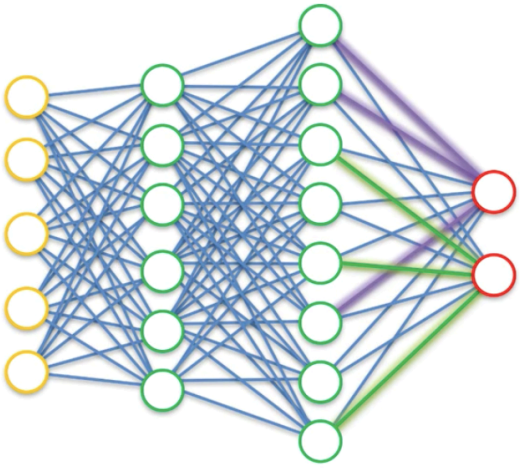
Artificial Intelligence A-Z
© SuperDataScience
udemy



The reason for having different action selection: Exploration and Exploitation.
Sometimes the environment forces the agent to explore different actions.
Exploration and exploitation is really important since sometimes the agent get stuck in the local maximum. For example, the agent might find based on its initial exploration that a certain action is the best action even though it's not and there is a bias toward this action and the agent thinks this is a very good action and it keeps getting good reward and it keeps doing that but there might be a better action than this. Here is when action selection going to help us.
$\varepsilon - greedy$: This will select the one with the highest Q value except for the epsilon percent of the time. For instance, if the epsilon is equal to 0.1 or 10% then 10% of the time, action is going to be selected at random and 90% it is going to be selected based on the best action based on the highest Q value.
That's why it's called greedy because we greedily choose good action except for a little epsilon. The lower the epsilon, the greedier algorithm is.
$\varepsilon - soft\ (1 - \varepsilon)$: This is exactly opposite of the $\varepsilon - greedy$ that it will select at random except epsilon percent of the time.

Action Selection Policies

Flattening

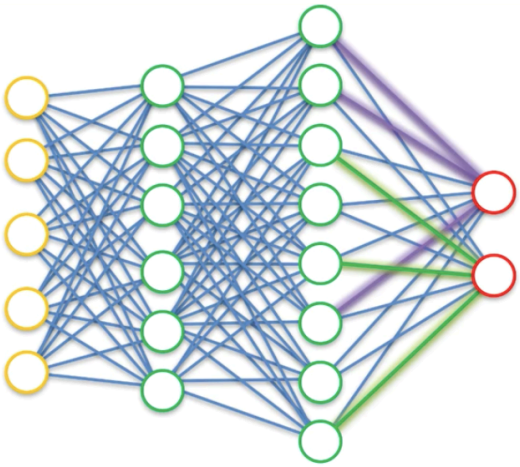


Dog 0.95

Cat 0.05

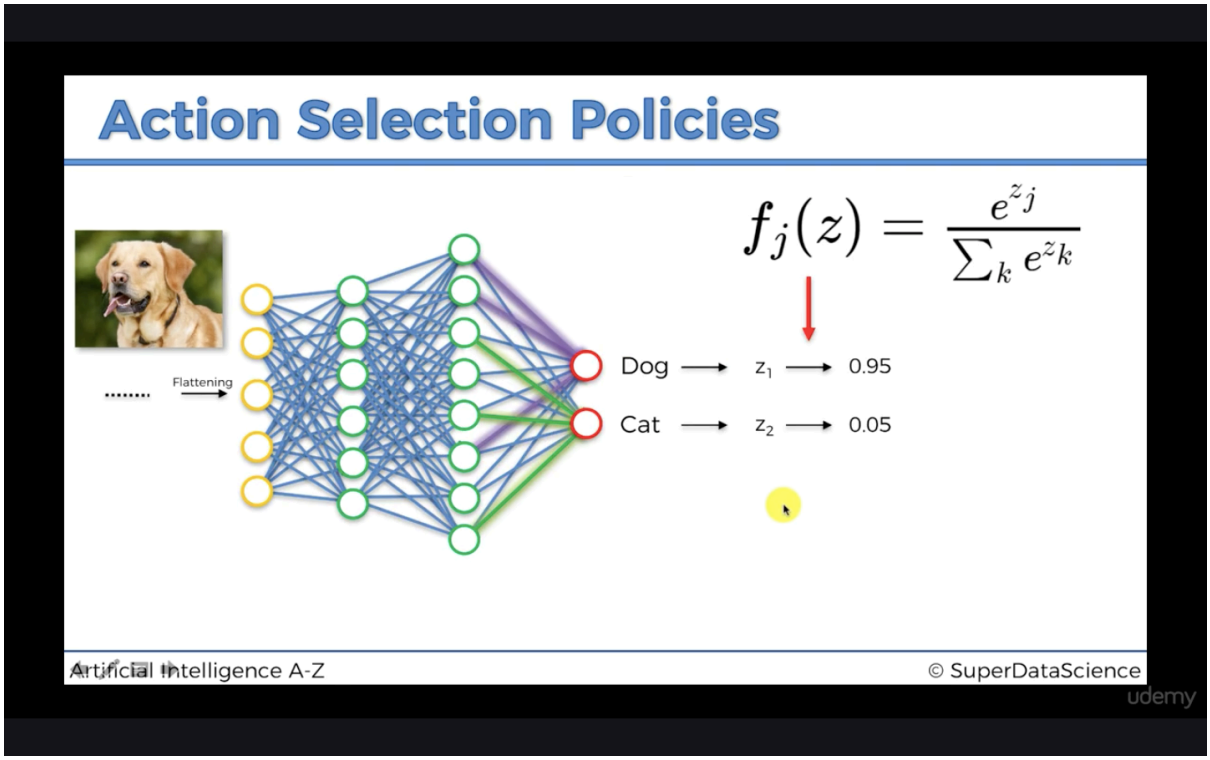
Action Selection Policies

Flattening

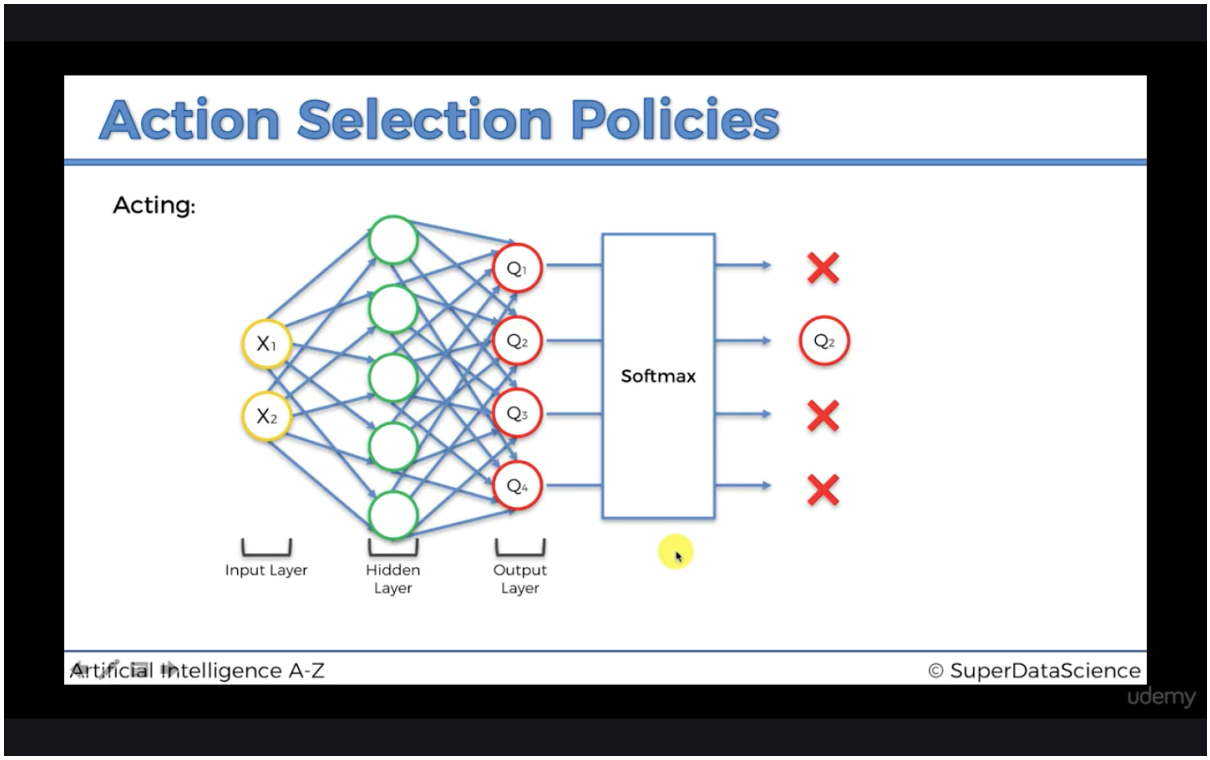


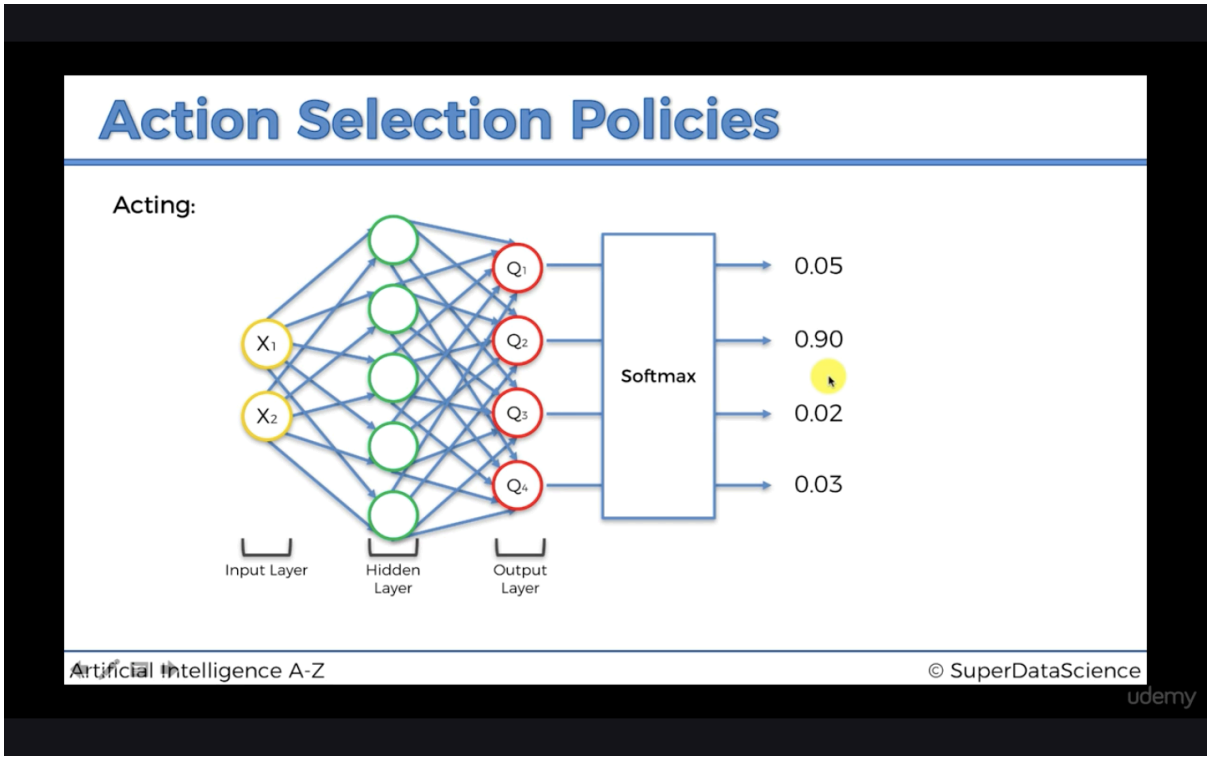
Dog 0.95

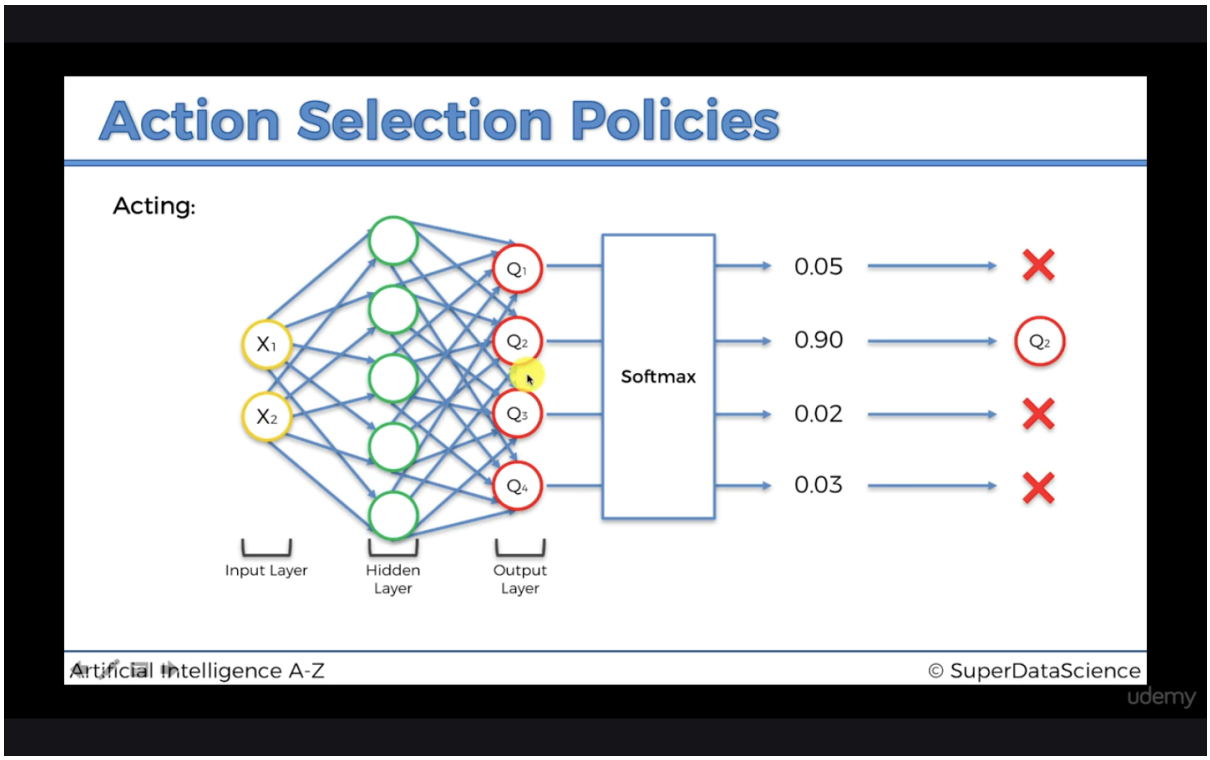
Cat 0.05



This is the softmax function that we are applying. After applying we can add our number in the total of number 1.

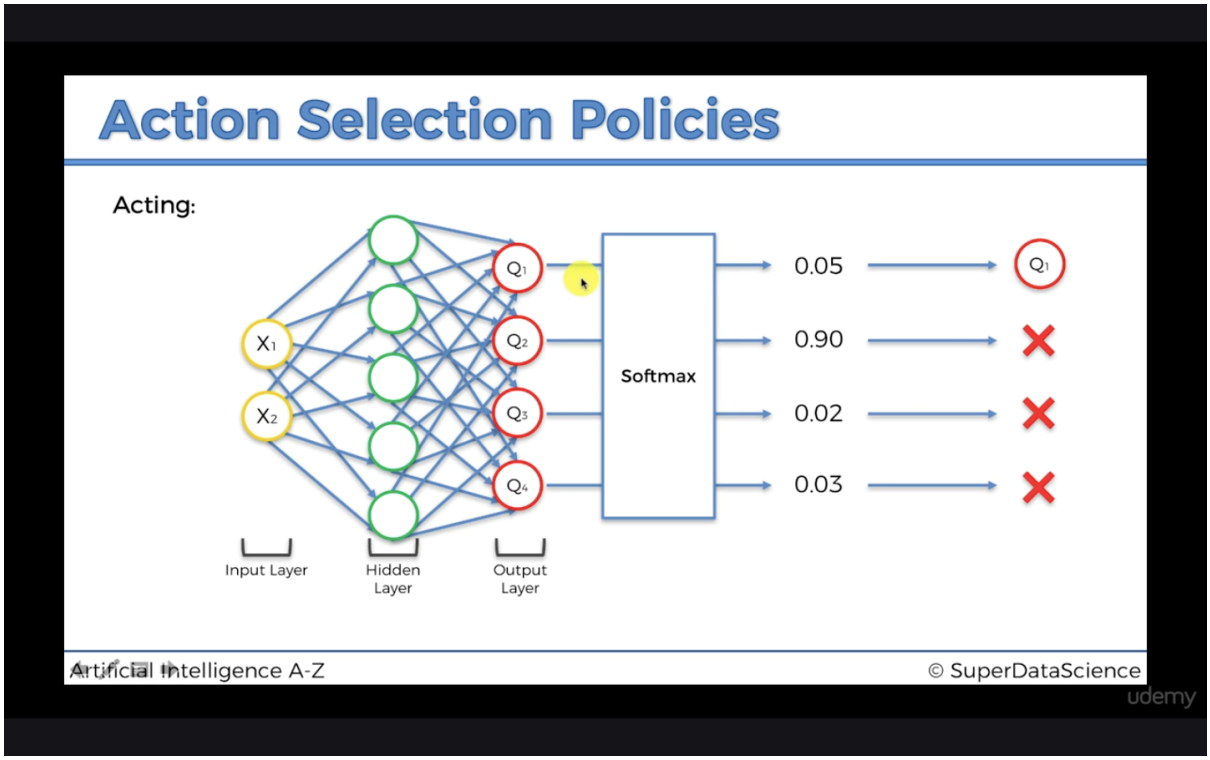




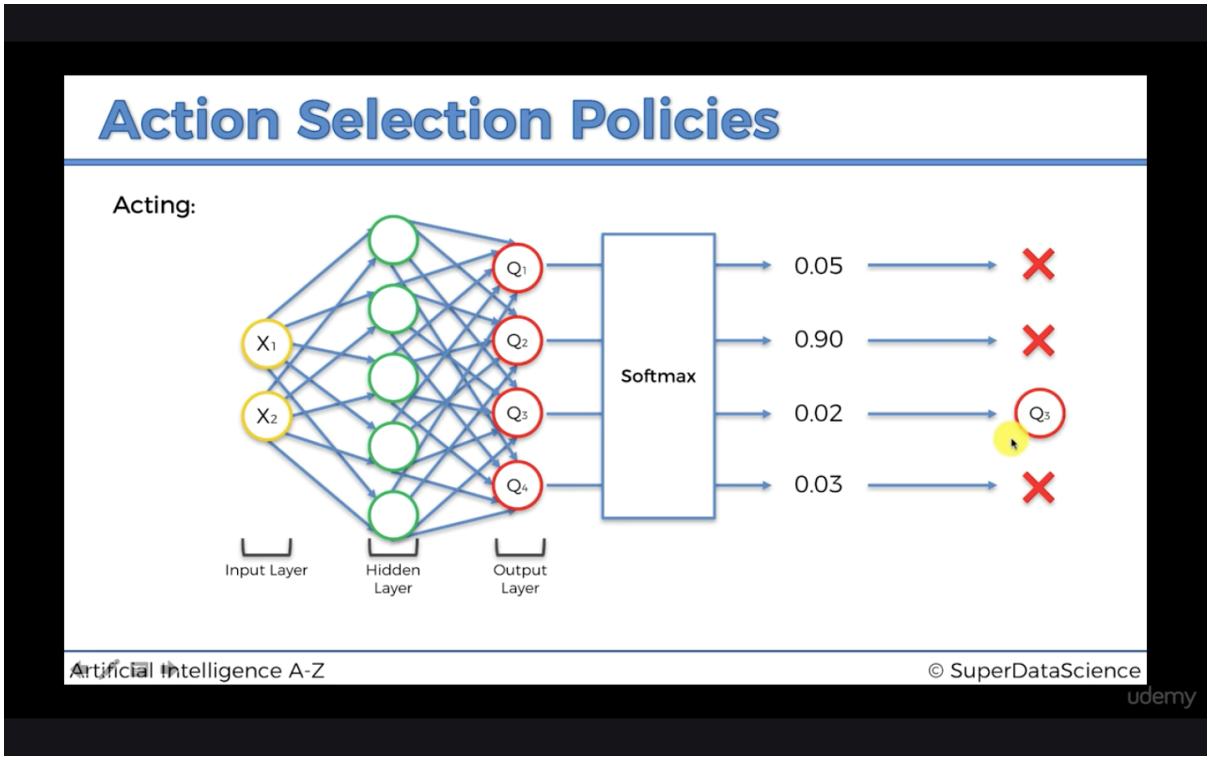


In here the best action is going to have the highest probability because it has the highest Q value. In here we are going to take Q2 90% of the times but still 5% of times we are going to take Q1 and so on.

After each update, these values are going to be change.



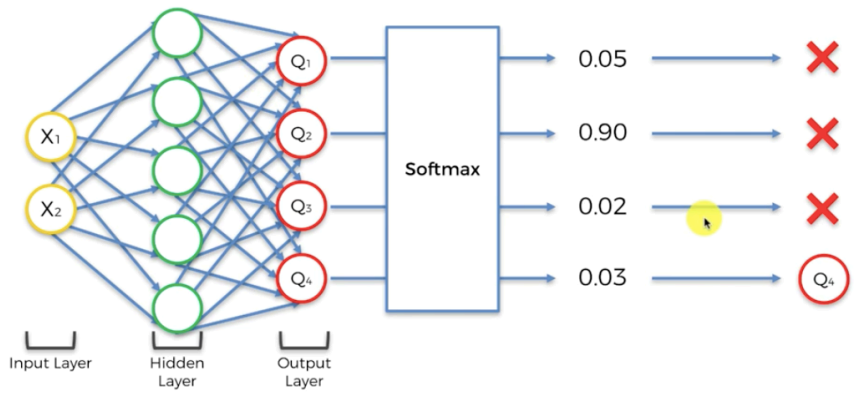
5% of time we choose Q1.



2% of time we choose Q3.

Action Selection Policies

Acting:



3% of time we choose Q4.

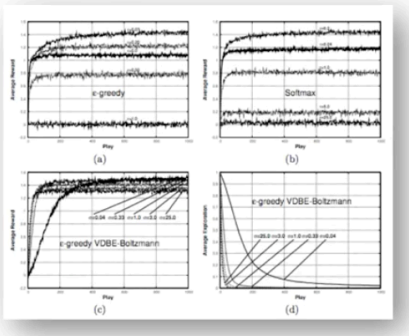
Additional Reading

Additional Reading:

Adaptive ε-greedy Exploration in Reinforcement Learning Based on Value Differences

Michel Tokic (2010)

Link:



http://tokic.com/www/tokicm/publikationen/papers/AdaptiveEpsilonGreedyExploration.pdf