This code uses a **linear combination** with a weight vector as policy to determine which action to take.

The program consists of 2 parts:

1. Exploration: Finding a good parameter set with random search
2. Exploitation: Using the found parameter set on a few episodes to check the parameter set

**Exploration:**

In every episode, a random weight vector is created and used until the cartpole fails or reaches a total reward of 200. (200 is the maximum possible step number, after that the done flag is set and the task is considered solved) It also saves the weight parameter set which performed best but finally returns the first parameter set which reaches a total reward of 200.

Usually, the random search finds a suited parameter set within the first 20-30 episodes.

**Exploitation:**

The found parameter set is then used again for a linear combination on 5 episodes just to see if the greedily found parameter set performs well on other starting points.

Sometimes the found parameter set will do great for all runs, sometimes it won’t solve the task in a single run.

This program was kept as simple as possible to show the basic idea of the exploration & exploitation policy.