**RL HW1**

110705013 沈昱宏

* Plot – 400k

一張含有 文字, 螢幕擷取畫面, 繪圖, 行 的圖片

自動產生的描述

I set alpha = 0.1 for the first 200k, alpha = 0.01 for the subsequent 100k, and alpha = 0.001 for the last 100k.

* Implementation & usage of n-tuple network

There are 16 blocks on the board, and each block consist of 16 bossible values. We can’t directly compute using the state of whole board due to memory limit, so we have to use n-tuple network to solve the problem. I used 4\*6-tuple networks with iso = 8 (the same as the sample code) (iso is for making more blocks considered with the same feature by rotating or mirroring the feature.

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

The patterns are then stored into class “learning” using a vector of pointer of “feature” class (the content is a derived class “pattern”). In the array of vector <int> named isomorphic under the class “pattern”, all the pattern (including its isomorphic) are stored using its index and will be used to (1) search for index, (2) update the weights stored in its base class “feature” (3) estimate the best choice.

The following codes are the implementation of the 3 functions I mentioned. Other explanation can be found in the comment.

一張含有 文字, 螢幕擷取畫面, 陳列, 軟體 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 軟體, 字型 的圖片

自動產生的描述

* TD(0)

In TD(0), you only have to consider two consecutive steps. The thing you want to learn is the value of V(s). By definition, we want the value of the next state V(s’’) equal to the value of the current state V(s) plus the reward r, so we update the V(s) with the TDerror (the difference that we want to update) and alpha.

一張含有 文字, 圖表, 螢幕擷取畫面, 行 的圖片

自動產生的描述

* Implementation – action selection

一張含有 文字, 螢幕擷取畫面, 陳列, 軟體 的圖片

自動產生的描述

一張含有 文字, 字型, 白色, 印刷術 的圖片

自動產生的描述

In the code, I iterate through all the possible moves and select the one that has the biggest value (the value is update by calculating + reward. The psuedo code is given by TA for estimating actions. However, there isn’t a function that can find , so I add pop up a tile directly when estimating the value of the future state. When the agent face the same situation multiple times, the mean of the pop up tile will gradually become , which is the desired value.

* TD backup

一張含有 文字, 螢幕擷取畫面, 軟體, 字型 的圖片

自動產生的描述一張含有 文字, 字型, 筆跡, 圖表 的圖片

自動產生的描述

I update the path from the ending state to the first step. I use a variable “afterstate\_value” to save the value of the before\_state on the last updated step, which will further be used to compute TD target and TD error.