Fundamentals of Speech Signal Processing

HW1 : Hidden Markov Model

I. Environment Setting and How to Execute

- 4.2.5-1-ARCH
- 進入 hw1_b02901085 資料夾中，輸入‘make‘ ，

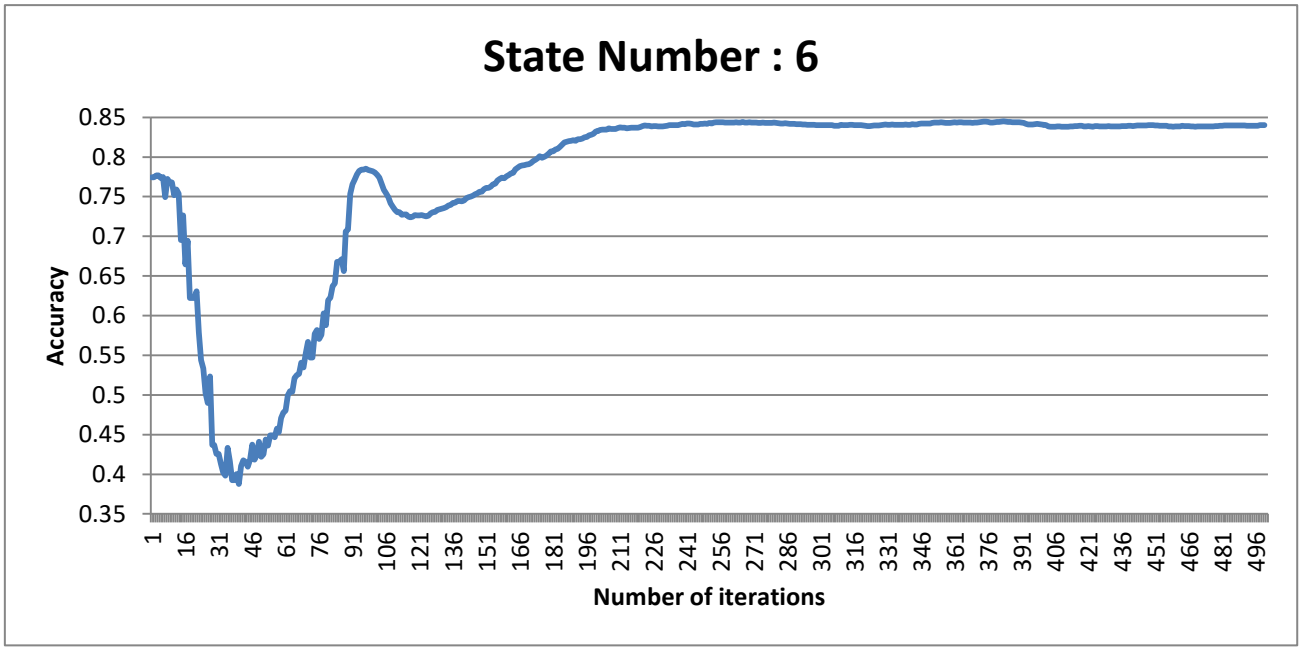
./train <iterations> <preTrainedModel> <training data> <saveModelName>

./test <modelList> <testingdata> <result> [accLog]

執行以上程式之前，請確定 testing_data1.txt ，
testing_data2.txt ， testing_answer.txt ， 所有的 sequence
model ， modellist.txt ， model_0*.txt 都在此資料夾中。

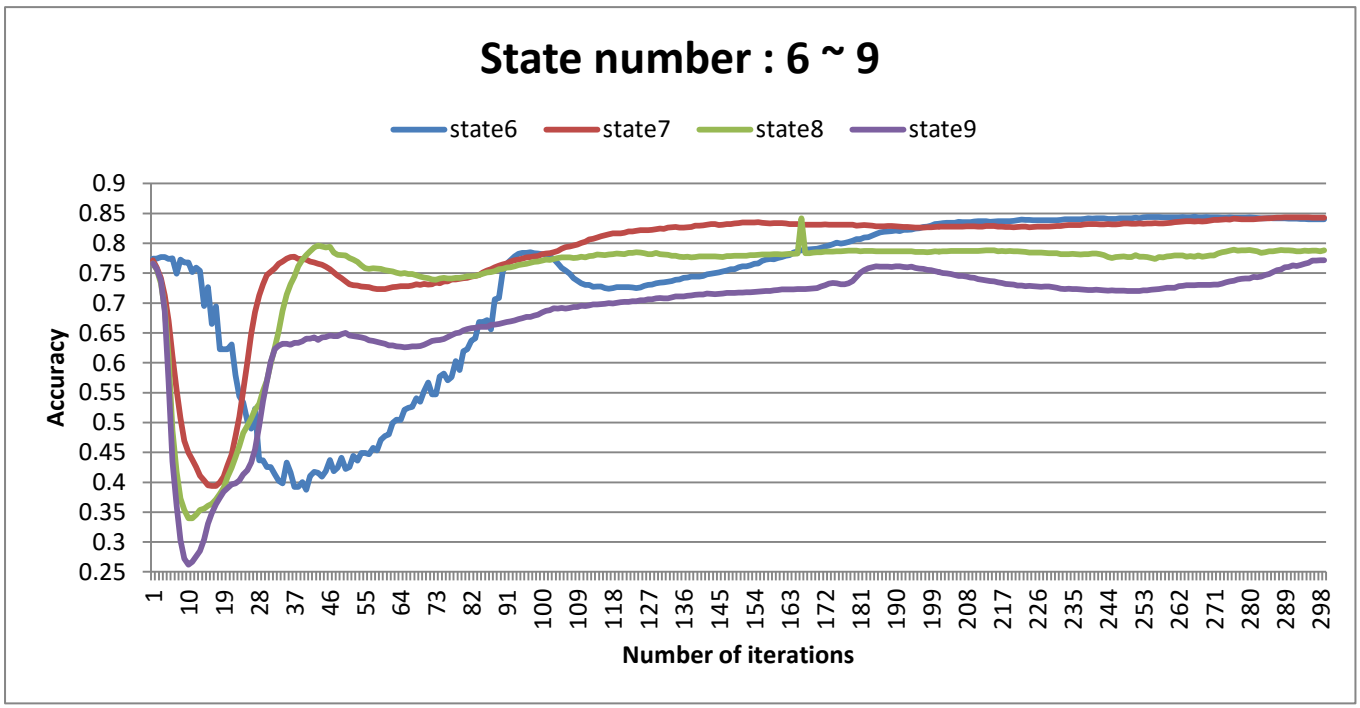
II. Result and Analysis

- 改動 iteration 的次數，如下圖



從圖中可以發現到，**iteration 次數越多，其準確率是慢慢提升**的，原先預期看到的 overfitting 現象(當 iteration 次數超過一個 threshold 後，其在 validation data 上的表現會下降)並沒有發生，結果是會**逐漸收斂**。

➢ 改動 state 數目，如下圖



從圖中可以觀察到，盲目地增加 state 數目是沒有幫助的，原先以為模型較為複雜，需要更多的 iteration 次數才會有好的效果，然而，縱使將 iteration 次數提高到 300 次，會發現其準確率依然較差，適合這個 task 的 state number 為 6 和 7，這說明了**以過度複雜的模型，去 model 較為簡單的 task，會造成 overfit**。關於這個推論，比方說今天有一序列 ABAB，我們分別以兩個 state 去描述它，則其中一個 state 經過 training 後，會代表 A，另一個則代表 B，但若今天我們以三個 state 去描述它，則 HMM model 的 state 轉移便多了好幾種可能，而且這些轉移的過程相似度並不高，使得 viterbi algorithm 不容易選出最佳的狀態轉移，在比較 model 之間對此 sequence 的機率時，會產生偏差。雖說使用太複雜的 model，準確率會降低，但太簡單也不行，會無法合適地描述這個 HMM，所以如何決定 state 數目也是一門大學問，查了一下資料後發現，關於最優 state 數目的選擇，有以下理論 *Bayesian Information Criterion*, *Akaike Information Criterion*, *Minimum Message Length Criterion*，與模型中可自由調整的參數(各種機率)，likelihood function，training data 的數量有關。