DSP HW1

Discrete Hidden Markov Model Implementation

工海三 柯哲邦 B05505053

- **Environment:**

  Macbook pro 2016 late

  CPU - 2 GHz Intel Core i5

  RAM – 8 GB 1867 MHz LPDDR3

  g++ compiler – Apple LLVM version 10.0.0 (clang-1000.11.45.2)

- **Compile:**

  $ make : to compile *test_hmm.cpp* & *train_hmm.cpp*

  $ make clean : to remove the executable files test & train

- **Execute:**

  $ ./train #iter *model_init.txt seq_model_0?.txt model_0?.txt*

  (#iter: the iteration number to train, '?' = 1-5)

  $ ./test *modellist.txt testing_data?.txt result?.txt*

  ('?' = 1-2)

- **Result:**

  因為不同的 iteration 次數產生的 model 不同，也會因此影響到 test 的 accuracy。所以本次實作觀察 accuracy 如何隨#iteration 而改變。(另寫了一個檔案去算 accuracy)

  數據顯示，當#iteration = 10 時，accuracy 會是最低點，之後就會回升，然後在#iteration = 880 之後都趨於穩定。也在#iteration 880 達到最高峰

#iteration	1	10	50	100	300	500	700	850	880	1000
accuracy	0.766	0.540	0.822	0.810	0.848	0.856	0.8656	0.8692	0.8696	0.8696

Best_Model:

#iteration = 880

Initial model = model_init.txt

Accuracy of result1 to testing_data1.txt = 0.8696

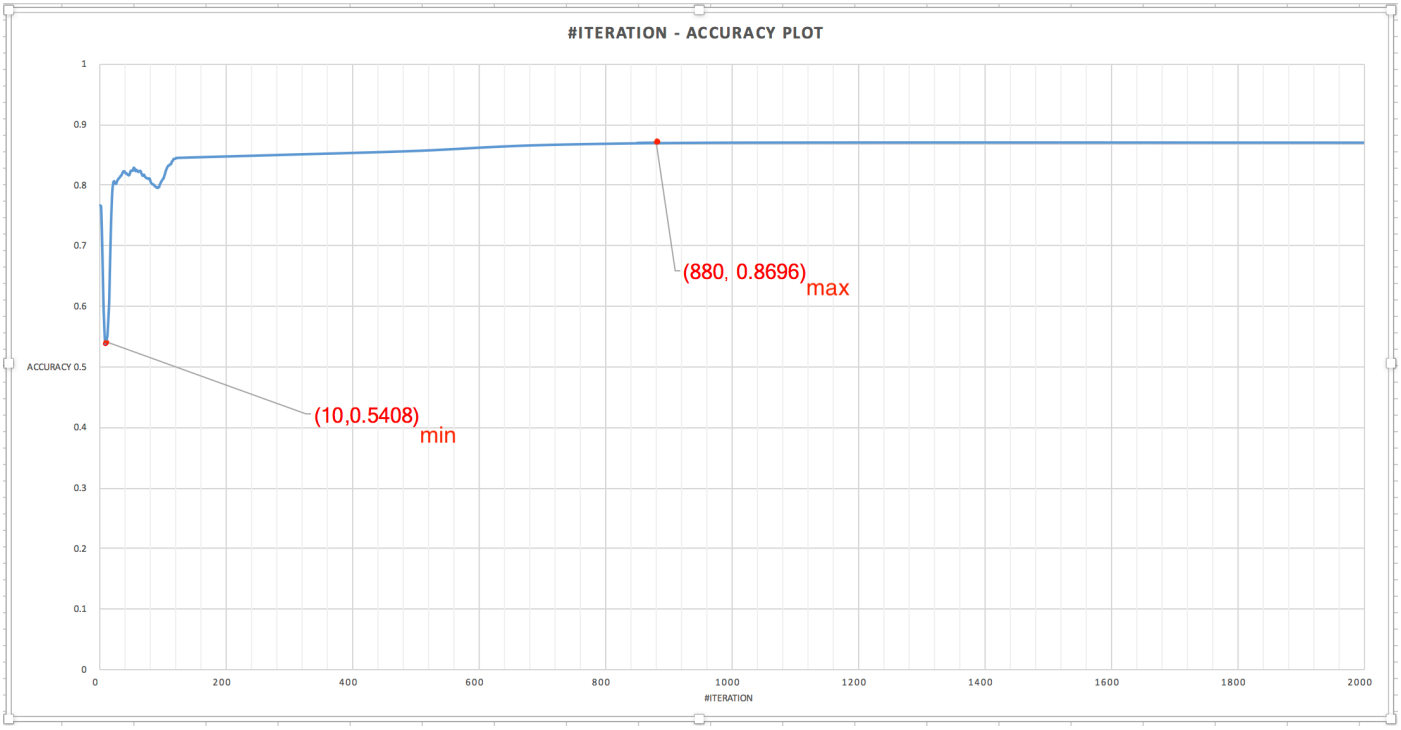
DSP HW1

Worst_Model:

#iteration = 10

Initial model = model_init.txt

Accuracy of result1 to testing_data1.txt = 0.5400



- **Additional experiment:**

考慮到 iteration 的次數會影響到 model 的大小，就會想到那如果我不取整個 seq_model，而是只取片段的 seq_model 來 train，比較完整和片段的結果。

Batch size: 1000 (only take 1000 sequences of the seq_model_0?.txt) and #iteration = 50:

Accuracy = 0.8153

- **Conclusion:**

#iteration 的變化相較 batch 的 size 更能影響機率。可以看到，#iteration 的不同，影響的機率最大可以達 0.3，而 batch size 取的不同影響非常些微。

- **Problems to discuss:**

1. The best way to calculate P(O|lambda).
2. Auto adjust the HMM parameters to avoid overfitting.
3. Compare the performance of the EM version and convex optimization version of Baum-Welch.
4. Is using small batch update still coverage?