HMM Homework Problems

Mark Stamp

February 11, 2022

1. For this problem, you will train HMMs on English text.

   a) Train multiple models on English text, experimenting with different random initializations. For all of your models, use $N = 2$ hidden states, $M = 27$ observation symbols, and $T = 50{,}000$ observations. When generating the observation sequence from English text, omit all symbols other than letters and word-space, and convert all uppercase letters to lowercase. Thus, the $M = 27$ observation symbols will consist of lowercase `a` through `z` and word-space. Give your best converged model $\lambda = (A, B, \pi)$ and explain what the model tells us about the training data.

   b) Repeat part a), but use $N = 4$ hidden states.

   c) Repeat part a), but use $N = 27$ hidden states.

2. In this problem, your will experiment with random restarts when training HMMs on English text. Repeat problem 1 a), but minimize T. That is, experiment to determine the smallest value of T that you can use to obtain a model that converges. Give your best model and the value of T that was used to train the model.

3. In this problem, you will use HMMs to break a simple substitution cipher.

   a) Randomly select a simple substitution key for $M = 27$ symbols (lowercase `a` through `z` and word-space), and encrypt an English plaintext message of length $T = 50{,}000$.

   b) Repeat problem 1 a) using your simple substitution ciphertext from part a) of this problem in place of English plaintext. Show that you can determine the ciphertext symbols that correspond to the plaintext consonants and vowels from the the converged model.

   c) Repeat problem 1 c) using your simple substitution ciphertext from part a) of this problem in place of English plaintext. Show that you can determine the key from the the converged model.

4. In this problem, you will experiment with HMMs to break a Vigenère cipher.

   a) Select a 5-letter keyword and encrypt English plaintext of length $T = 50{,}000$. In this case, omit all symbols other than letters, and convert all uppercase letters to lowercase. Since we do not include word-space, there are $M = 26$ symbols.

b) Train an HMM on your Vigenère ciphertext from part a) of this problem with $N = 5$ hidden states. Show that the converged model can be used to determine the key.

c) Explain how you could use HMMs to break a Vigenère ciphertext message in a case where you do not know the keyword length.

5. In this problem, you will experiment generating fake English text using a trained HMM.

a) Use the best HMM that you generated in problem 1 a) to generate at least 1,000 characters of fake English text. That is, generate uniform random numbers and use the probabilities in the A matrix to determine state transitions, and the probabilities in the B matrix to determine which observation symbol to output at each step. Comment on the quality of this fake English text.

b) Repeat part a) of this problem, but using the model you generated in problem 1 c). Is there any noticeable improvement in the quality of the fake English text generated with $N = 27$ hidden states, as compared to that generated with $N = 2$ hidden states?