CS769 Advanced NLP Homework 2

Assigned 1/27/2009
Due 2/10/2009 in class

1 MLE

Download the particular version of *Alice's Adventures in Wonderland* from `http://pages.cs.wisc.edu/~jerryzhu/cs769/dataset/alice.txt`. This is the document we'll be working on.

1. **Question 1**. [10] How many *distinct, ASCII printable* characters are there in the file? For example, in the string "Aaaab" there are 3 such characters. See `http://en.wikipedia.org/wiki/ASCII#ASCII_printable_characters` for a definition of ASCII printable characters.

2. **Question 2**. [10] Assuming the file is generated from a multinomial distribution θ over characters. Find the MLE θ, sort it as:

   most-frequent-character, $\theta_{\text{most-freq-char}}$
   second-most-freq-char, $\theta_{\text{second-most-freq-char}}$
   ...
   Hand in this sorted list.

3. **Question 3**. [10] Let D be the whole *Alice's Adventures in Wonderland* data. Compute the likelihood $P(D|\theta)$ of your MLE θ using Matlab. Discuss any issues you might have encountered.

4. **Question 4**. [20] Repeat Questions 2 multiple times: Estimate the MLEs $\theta_{10}, \theta_{100}, \theta_{1000}, \theta_{10000}, \ldots$ on the first 10, 100, 1000, 10000, ... printable characters of the document. The last θ should be using the whole document. You do not need to hand in these θs. Instead, compute the Euclidean distance between each adjacent θ pairs. Discuss your observations.

2 Multinomial and Dirichlet Distributions

- **Question 1.** [10] Let us assume that a "monkey keyboard" has only three keys: `A B` and white space. A monkey hits them with probability p, p and $1 - 2p$, respectively. Derive the rank and frequency function relation on monkey words.

- **Question 2.** [10] In the first five letters the monkey produced, there are two A's, one B, and two white spaces. Plot the likelihood function of p on these five letters using matlab.

- **Question 3.** [10] For this question, assume the same five letters as in Question 2, but allow the character probabilities to be arbitrary (instead of constraining them to be $p, p, 1-2p$). Assuming a Dirichlet prior with parameters $\alpha_A = 0.2, \alpha_B = 0.3, \alpha_{whitespace} = 0.5$. What is the *mean and mode* of the posterior distribution, respectively?

3 Best Strategy

- **Question 1.** [10] Suppose the world generates an observation $x \sim \text{multinomial}(\theta)$. You know θ and want to guess what x is. What is the probability that your guess will be correct, using the following strategies respectively? Prove your answer.

  Strategy 1: Always guess $x^* = \arg\max_x \theta_x$, the outcome with the highest probability.

  Strategy 2: You mimic the World by generating an outcome $y \sim \theta$, and guess y.

- **Question 2.** [10] Repeat the above question. But this time, you believe the parameter is θ while the world is using some other parameter ϕ.