CMSC 828G:Spring 2002

Sketch of Solutions to Homework 2

- Problem 1

Part	Time complexity	Space Complexity
1	O(rn)	O(r)
2*	O(dn)	O(d)
3	O(cn)	O(c)
4	O(rn)	O(r)
5	O(rn)	O(r)
6	O(r^2n)	O(r^2)
7	O(cn)	O(c)
8	O(c^2n)	O(c^2)

  * There is a simple solution to this part via binning by going over all the data in one pass. Some of you suggested *sorting* which is less efficient O($dn \log n$) time complexity and O($n + d$) space (points were not deducted for this solution).

- Problem 2

  - 1. Store the index of the cells which contain 1's in a list.

  - 2. Walk along the two lists and count mismatches, O($n_i + n_j$).

  - 3. The sample mean can be interpreted as the probability that the word will appear in a randomly selected document. The sample variance can be interpreted as the variability of the word occurrences in the different documents.

  - 4. The difference in the distribution of the word occurrences.

  - 5. There are a number of possible answers. Some of you suggested the representative document which is the closest to the mean or the furthest document which can be an outlier.

  - 6. This is not fair for the short document. One possible fix is normalization.

- Problem 3

  - 1.

$$P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\frac{\partial P(\theta)}{\partial \theta} = 0 = (\alpha - 1)\theta^{\alpha-2}(1-\theta)^{\beta-1} - (\beta-1)\theta^{\alpha-1}(1-\theta)^{\beta-2}$$

Which yields

$$(\alpha - 1)(1-\theta) - (\beta - 1)\theta = 0$$

Which implies that

$$\theta_{max} = \frac{\alpha - 1}{\alpha + \beta - 2}$$

You *should* check the second derivative to make sure that this value actually maximizes the density.

– 2. $\theta_{ML} = \frac{r}{n}$ and the prior mean $= \frac{\alpha}{\alpha+\beta}$
Assume

$$\frac{\alpha}{\alpha+\beta} < \frac{r}{n}$$

This implies that

$$\alpha n + rn < (\alpha + \beta)r + rn$$

Which implies

$$\frac{\alpha + r}{\alpha + \beta + n} < \frac{r}{n}$$

Repeat for the case of $>$ and $=$.