STAT 542: Homework 7

Due: May 5 midnight on Canvas

Please make sure that your solutions are readable and the file size is reasonable. Typing the answers is highly encouraged.

Problem 1.

Suppose that a naive text generation model works in the following way: it looks at the string of words it has already produced, picks one (uniformly) at random, and then adds one more that word to the string. For example, suppose that at the initialization, the existing string of words is "popular language model model". In the next iteration the string becomes the following with probabilities 1/4, 1/4, 1/2:

popular language model model popular
popular language model model language
popular language model model model

and the iterations will continue.

- [1pts] Suppose that at the initialization the string is "popular language". What is the distribution of the number of occurrences of "popular" after n new random words are added by this generative model?

- [1pts] Suppose that at the initialization the string is "popular language model model". Let X_n be the number of occurrence of "popular" after n new words are added, and let $X := \lim_{n\to\infty} X_n$. Note that X is random because X_n is random. What is the distribution of X?

  Hint: The answers are distributions we encountered in class. Your answer should be specific about what are the parameters of the distribution and how you obtained them.

Problem 2.

An independent set of a graph is a set of vertices in a graph, no two of which are adjacent. (Intuitively this models a community in which people do not talk to each other.)

- [1pts] Consider the pentagon graph consisting of 5 vertices and 5 edges corresponding to the sides of a pentagon. How many independent sets of sizes 1,2,3,4,5 does this graph have?

- [1pts] Note that a subset of the vertices in the pentagon graph can be represented by a vector $x = (x_1, \ldots, x_5) \in \{0,1\}^5$, where $x_i = 1$ iff the i-th vertex is in this subset. Consider the following probability distribution of x:

$$P(x) = \frac{\lambda^{\|x\|_1}}{Z} 1\{x \text{ is an independent set}\} \tag{1}$$

where $\lambda > 0$ and Z is a normalization constant, and $\|x\|_1$ is the size of the subset represented by x. Show that

$$P(x) = \prod_{\{i,j\} \in E} \phi_{i,j}(x_i, x_j) \tag{2}$$

for some functions $(\phi_{i,j})_{\{i,j\} \in E}$, where E is the edge set of the pentagon graph. Explain how $\phi_{i,j}$ is defined using λ.

Remark: the independence set of the pentagon graph (and the product graph) played an important role in Shannon's celebrated result (nonadditivity) on zero-error capacity.

For the second question, it may be tempting to use Hammersley-Clifford, but a more intuitive construction directly yields the result. The pairwise factorization here actually holds for general graphs. While finding largest independent set is NP-hard, the factor form suggests the possibility of sampling a large independent set by taking λ large and using statistical sampling techniques.

Problem 3.

[3pts] Solve Ex. 14.22 in "Elements of Statistical Learning", copied below. (a) and (b) are worth 1pts and 2pts respectively.

> Ex. 14.22
>
> (a) Show that definition (14.108) implies that the sum of the *PageRanks* p_i is N, the number of web pages.
>
> (b) Write a program to compute the *PageRank* solutions by the power method using formulation (14.107). Apply it to the network of Figure 14.47.

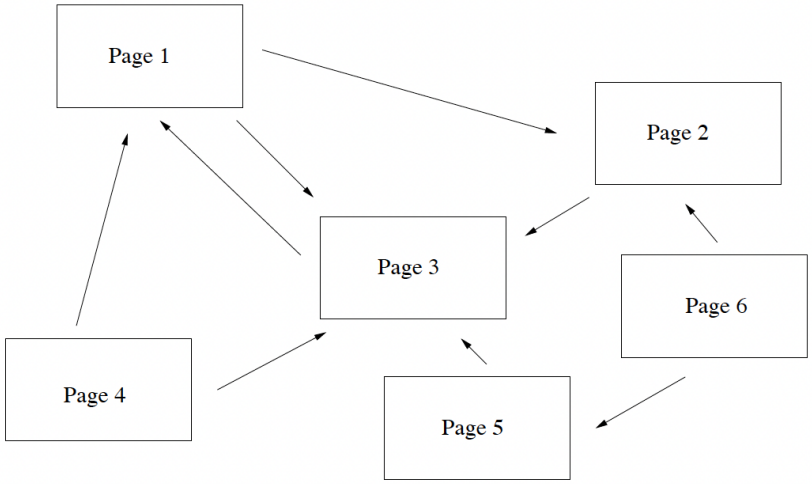

FIGURE 14.47. *Example of a small network.*