STAT 598Z: Course Project

Due: 24th April 2012

1 Sparse Matrices

In many applications of statistics and machine learning, the data is sparse, that is, only a small number of features are non-zero. It is therefore important to represent and manipulate sparse vectors. In this project we will study one simple representation. Each vector is represented as a list of tuples, and each tuple contains two elements namely an index and a value. Only non-zero elements are represented. For instance, the sparse vector $[10, 0, 20]$ is represented as follows: $[(0, 10), (2, 20)]$, where 0 indicates the index of the first element and 10 indicates its value. This idea can be extended to matrices in a straightforward manner by representing a matrix as a list of sparse vectors. For instance,

$$\begin{bmatrix} 10 & 0 & 20 \\ 0 & 30 & 0 \\ 0 & 40 & 50 \end{bmatrix}$$

can be represented as $[[(0, 10), (2, 20)], [(1, 30)], [(1, 40), (2, 50)]]$.

2 Project Description

1. This project will contribute 20 points towards your final score.
2. Download the `sparse_skeleton.py` file and follow the instructions to complete the implementation of various functions.
3. You will need to download the a9a dataset from `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary/a9a` to complete the project.
4. Hand in your Project (including print outs of your source code) at the beginning of the class on 24th April 2012. Additionally your source code should be emailed to `stat598z@gmail.com` **before** the project is submitted in the class. No late submissions will be accepted!