EPOKA UNIVERSITY

CEN376 Data Mining

HOMEWORK ASSIGNMENT

PROBLEM 1

In this problem you are required to apply various classification techniques on a benchmark dataset: Diabetes Risk Prediction given as a CSV file in the attachment (source: Kaggle). This dataset is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). The data includes features such as age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, visual blurring, itching, irritability, delayed healing etc. This dataset can be used to build machine learning models to predict diabetes in patients based on their medical history and demographic information.

In order to assess the performance of the classification techniques, you are going to apply a simplified version of the k-fold cross validation method. The simplified k-fold cross validation method proceeds in this way:

- Shuffle the dataset
- Divide the dataset into k equal partitions
- For each of the partitions:
  - Apply the classification model using the union of (k-1) other partitions as training set
  - Test it on the current partition
  - Analyze the performance (precision, recall, accuracy) for this case
- Generate the overall performance report, taking the averages of the found results.

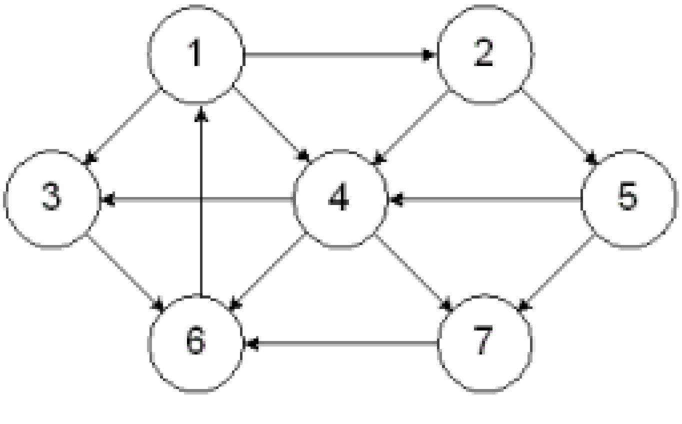
Apply a 5-fold cross validation methodology, as described above, for all the classification techniques that we have studied in our course:

i. Decision trees
ii. KNN
iii. Support Vector Machines
iv. Logistic Regression
v. Naïve Bayes

For each classification technique that is applied, the overall performance report should be printed. Briefly compare the results.

PROBLEM 2

Apply the PageRank algorithm on the graph given in the figure. Print the nodes in a sorted order according to relevance values (evaluated by PageRank)



PROBLEM 3

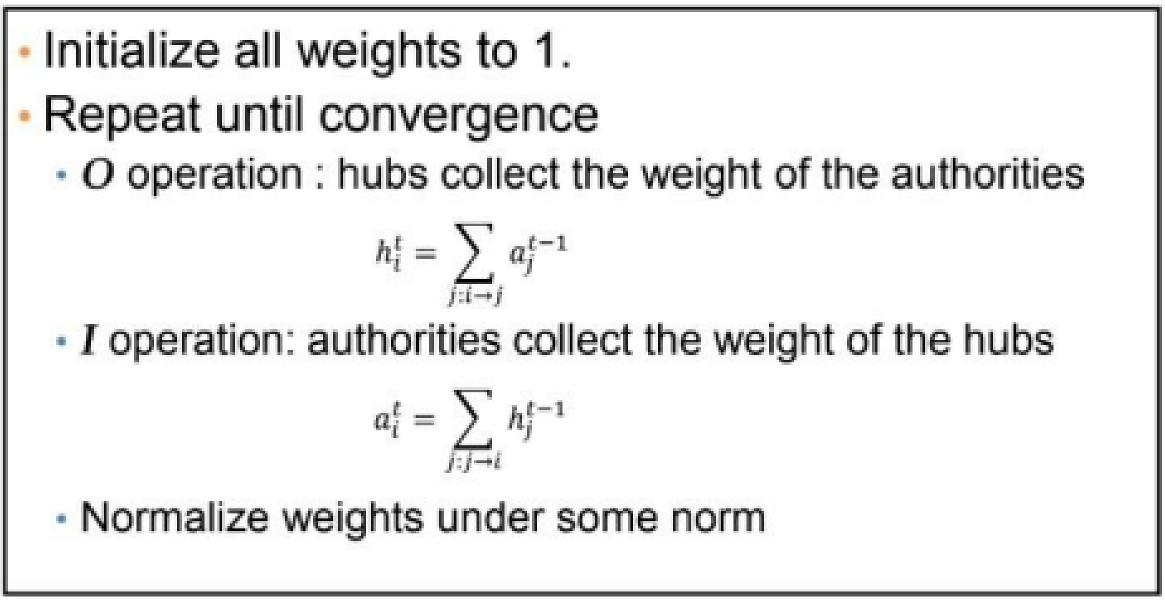
The given file kosarak.dat provides (anonymized) click-stream data of a hungarian on-line news portal (http://fimi.uantwerpen.be/data/).

a. Apply the Apriori algorithm to find the frequent itemsets of books with minSupport = 0.15. b. Find the largest value of minSupport (with a scale of precision 0.001) for which there would exist at least one frequent itemset of size 4.

PROBLEM 4

In this problem you are required to compare two versions of the HITS algorithm discussed in our lectures.

Figure 1 - HITS Version 1



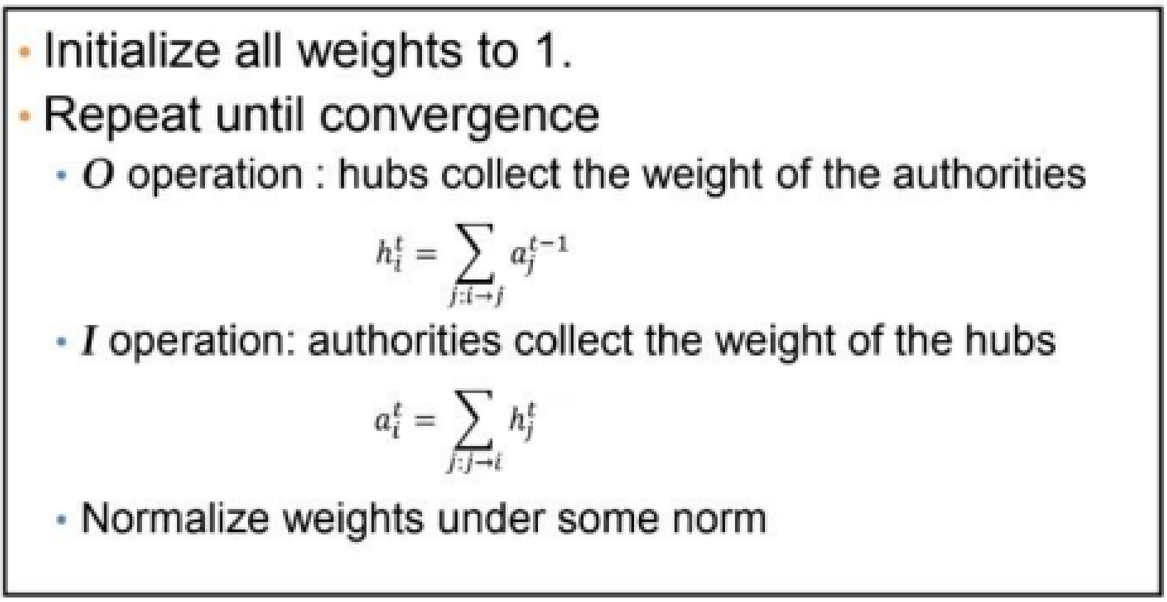
- Initialize all weights to 1.
- Repeat until convergence
  - *O* operation : hubs collect the weight of the authorities

$$h_i^t = \sum_{j:i \to j} a_j^{t-1}$$

  - *I* operation: authorities collect the weight of the hubs

$$a_i^t = \sum_{j:j \to i} h_j^{t-1}$$

  - Normalize weights under some norm



- Initialize all weights to 1.
- Repeat until convergence
  - *O* operation : hubs collect the weight of the authorities

$$h_i^t = \sum_{j:i \to j} a_j^{t-1}$$

  - *I* operation: authorities collect the weight of the hubs

$$a_i^t = \sum_{j:j \to i} h_j^{t}$$

  - Normalize weights under some norm

You will pick 3 graphs (of your choice) and a scale of tolerance of your choice and the comparison will be done in two aspects:

- The hub and authority values that the algorithms converge to
- The number of steps that the convergence is achieved

Note: in this problem you are not allowed to use any libraries.