University of Ioannina,
Department of Computer Science & Engineering

Data Mining – CSE 012/CS 059
Fall 2015

# Assignment 4

The deadline for Assignment 4 is Februady 12, by the end of the day. Submit everything electronically, either through turn-in or email. For late submissions the late policy on the page of the course will be applied. Details for the turn-in, and how to write reports are on the Assignments web page of the course.

## Question 1

Let $U = \{x_1, \dots, x_N\}$ be a universe of elements. We define a weighted subset $S$ of $U$ as an $N$-dimensional vector $W_S$, where $W_S[x]$ is the weight of the element $x$ in S, if $x \in S$, and zero if $x \notin S$. Let $C = \{S_1, \dots, S_k\}$ denote a collection of weighted subsets of $U$. We define the weight vector of the collection $C$ as an $N$-dimensional vector $W_C$, where $W_C[x] = \max_{S \in C} W_S[x]$.

Given a collection $C$ we define the function $f(C) = \sum_{x \in U} W_C[x]$, to be the total weight of $C$ for all elements. (Note that this can also be defined for a set, which is a special case of the collection). Prove that the function $f$ is submodular, that is, for any two collections $A, B$, such that $A \subseteq B$, and set $S \notin B$, $f(A \cup \{S\}) - f(A) \geq f(B \cup \{S\}) - f(B)$.

## Question 2

Let $G = (V, E)$ be an undirected graph that represents a social network. A triplet of nodes $\{u, v, w\}$, defines an **open triangle**, if $(u, v) \in E$ and $(v, w) \in E$ but $(u, w) \notin E$. Assume that the edges of the graph are labeled as Strong or Weak. We say that the labeling satisfies the **Strong Triadic Closure property** if for every open triangle $\{u, v, w\}$, at least one of the edges $(u, v)$, $(v, w)$ is labeled as Weak. The intuition of this definition is that it cannot be the case that $v$ has strong relationships with both $u$ and $w$, but $u$ and $w$ do not know each other.

We are given as input a graph $G = (V, E)$ without any labels on the edges. We want to find a labeling of the edges such that it satisfies the Strong Triadic Closure property, **and** the number of Weak labels is minimized. In other words, we want to find the minimum subset of edges $S \subseteq E$, such that, if labeled Weak, all open triangles will satisfy the Strong Triadic Closure requirement.

Show how this problem can be mapped to a coverage problem. Based on the mapping, derive an approximation bound for the problem.

## Question 3

In this question you will use the data that you created for Assignment 2 on Recommendation Systems. The goal is to use the social network between the Yelp users in order to predict their ratings for new businesses.

Start with the dataset from Assignment 2, consisting of businesses in Las Vegas that have at least 10 reviews from users with at least 10 reviews. Using these users as nodes of the graph, construct the graph consisting of friendship edges between them, which you will obtain from the file yelp_academic_dataset_user.json. From this graph keep the largest connected component. This will define the graph $G$ that you will work with, and the nodes in the component the set of users that we are interested in (the set of businesses remains the same).

Remove randomly 10% of the ratings of the users, and try to predict the rating for the user-business pair $(u, b)$ using a random walk with absorbing nodes as follows. Given a pair $(u, b)$ , in the graph $G$, make every node $v$ that has rated the business $b$ to be absorbing, and assign to that node the value of the rating $R(v, b)$. Using the value propagation method we described in class compute a rating $P(v', b)$ for every non-absorbing node $v'$ in the graph. The predicted rating for node $u$ will be the value $P(u, b)$.

Compare the Mean Sum of Square Errors (MSSE) that you obtain for this solution with the MSSE you get for the user-based and item-based collaborative filtering approaches that you implemented in Assignment 2 (methods 2,3 from Question 2).

Reminder: The MMSE is defined as:

$$SSE = \frac{1}{n}\sum_{i=1}^{n}(r_i - p_i)^2$$

Where $r_i$ is the actual rating and $p_i$ is the predicted rating.

**Bonus**: Propose, implement, and test a different method for predicting the ratings using random walks on the social graph.

## Question 4

In Kaggle there is an active competition from AirBnB for predicting the first booking of a new user (here is the link to the competition). Using the account you created for the previous assignment submit a solution to the competition. The goal is not to win the competition (although, again, your position in the leaderboard will count towards a bonus grade), but rather to work on a real problem, that does not have a known solution.

Create a report which should contain the following:

a. A description of how you decided to model the problem (as a classification problem, as a clustering problem, frequent itemsets etc).

b. A description of the solution you implemented. What features you used, what techniques you applied, and a short explanation of the rationale of your choices.
c. Your results on the Kaggle test dataset.
d. A commentary on the above: What seems to work and why? What insight did you gain into the data and the problem?

Hand in your code, and the report. Make sure to note your user-name in Kaggle, and your standing at the time that you submitted the report.