University of South Florida
CIS 6930 Special Topics (Social Media Mining)

Homework Assignment 4

Due: Fri Nov. 1, 11:59 p.m. on Canvas.

Instructions (do's and don'ts)

Late submissions will be assessed a penalty. Please *do* review the policy about late work on the syllabus.

Do submit exactly three (3) files on Canvas:

1. a text document (accepted formats: `.pdf`, `.doc`, `.docx`, and `.rtf`),
2. a Jupyter Notebook saved as a Python script (*File -> Download as -> Python (.py)*),
3. a Jupyter Notebook saved as a PDF file (*File -> Download as -> PDF via LaTeX (.pdf)*).

Submit each file separately; clicking on *Add Another File* in the upload form will let you add extra files. *Do not* combine the files in a compressed archive, as they cannot be parsed by Turnitin. For the same reason, please *do not* include scanned hand-writing in the first file.

Speaking about plagiarism, please *do* review the course policy on the syllabus, as well as the general USF policy on academic integrity (available online at: www.usf.edu/provost/faculty/core-syllabus-policy-statements.aspx).

All submissions will be graded anonymously, to ensure fairness. Please *do* include your UID in the submission. *Do not* include any other indentifying information (name, etc.). The grading rubric is on Canvas; it is displayed on the assignment page. Please *do* review it as well.

If you have any question, please *do* ask the instructor or the TA.

1 Theory: The segregated random network

Let us consider two disjoint realizations of the $G(n, p)$ random graph model, that is, let us consider two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, where $|V_1| = |V_2| = n$. We say that the two graphs are *disjoint* meaning that $V_1 \cap V_2 = \varnothing$, i.e. they have no nodes in common. For example, V_1 and V_2 could represent two groups of users of opposite political affiliation (like liberals and conservatives in the U.S., or tories and labor in the U.K.), or fans of two different sport teams in the same city.

Let us now consider the graph obtained by combining G_1 and G_2 together. This is formally defined as $G_1 \cup G_2 = (V_1 \cup V_2, E_1 \cup E_2)$. Answer the following questions:

(i) What is the expected density of $G_1 \cup G_2$?

(ii) What is the expected local clustering coefficient of a node $v \in V_1 \cup V_2$?

(iii) Let us assume now that we add edges connecting members of the two groups in the following way: for any two nodes v and w such that $v \in V_1$ and $v \in V_2$, we flip a coin and with probability p' we add a new edge $\{v, w\}$. What is the expected number of edges in $G_1 \cup G_2$ after this operation?

2 Coding: Classification with the decision tree model

Write a self-contained and fully functional Jupyter Notebook that:

(a) Loads the necessary libraries to run (e.g., scikit-learn).

(b) Loads the dataset "digits" (see below for description) from the scikit-learn example datasets package.

(c) Instantiates a `sklearn.tree.DecisionTreeClassifier`.

(d) Splits the dataset in training and testing sets using `sklearn.model_selection.KFold` cross validation (use $K = 5$).

(e) Trains and tests the performance of the classifier in identifying the digits.

(f) Computes and outputs the accuracy for each of the K folds of cross validation.

(g) Computes and outputs the average accuracy across the K folds.

A description of the dataset is available at: `http://goo.gl/MZNP9N`.