U: U72884742

Assignment: 4

**Question1] 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 G1 = (V1, E1) and G2 = (V2, E2), where |V1| = |V2| = n. We say that the two graphs are disjoint meaning that V1 ∩ V2 = ∅, i.e. they have no nodes in common. For example, V1 and V2 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 G1 and G2 together. This is formally defined as G1 ∪ G2 = (V1 ∪ V2, E1 ∪ E2). Answer the following questions:

1. What is the expected density of G1 U G2?
   1. Density is the ratio of the number of edges |E| that graph G has over the maximum it can have
   2. =
   3. A graph G = (V,E) is -dense if |E|
2. What is the expected local clustering coefficient of a node v ∈ V1 ∪ V2?
3. 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 ∈ V1 and v ∈ V2, we flip a coin and with probability p 0 we add a new edge {v, w}. What is the expected number of edges in G1 ∪ G2 after this operation?
   1. Adding new edge connecting members of the two groups {v,w}
   2. To do: expected number of edges in G1 U G2 after adding an edge
      1. |E1| + |E2| + number of edges

**Question2] 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>.

