CS492-Data Science
Assignment 02

Due Date: 18th March 2019 (In Class) **Total Marks: 100**

- What is the difference between labelled and unlabelled data? (Write in your own words) [**5 Points**]
- The following information is held in an employee dataset
  Name, Date of Birth, Annual Income, Tax Paid, Gender
  What is the type of each variable i.e. Categorical or Continuous? [**5 Points**]
- Use Naive Bayes Classifier to Classify Test Samples from Table **??**. Use Table **??** for Training. Show all steps clearly [**10 Points**].
- Use Table **??** to create Decision Tree (both C4.5 and CART). Show all steps clearly. Upto 2 levels are enough. You can use simple computer programs for continuous data [**20 Points**]
- Table **??** shows data for training. Use *k* nearest classifier to calculate the classification accuracy, precision, recall, F-measure. Show confusion matrix clearly. Use k=3 and city block distance. Show all steps clearly. Use 2 separate results using 3 Fold CV and Leave out CV. [**20 Points**]
- Biggest problem with kNN classifier is storing all data points in memory. What is the time complexity of traditional kNN classifier in terms of Big Theta? Review few research papers that has improved the time complexities of kNN classifier. [**30 Points**]

Refund	Status	Tax Income	Cheat
Yes	Single	125	No
Yes	Single	80	No
Yes	Single	20	Yes
Yes	Married	90	Yes
Yes	Divorced	45	No
Yes	Divorced	90	Yes
Yes	Divorced	80	No
No	Single	75	No
No	Single	65	Yes
No	Married	120	Yes
No	Married	80	No
No	Divorced	100	Yes
No	Divorced	75	Yes

Table 1: Training Data.

Refund	Status	Tax Income	Cheat
Yes	Single	115	?
Yes	Single	5	?
Yes	Married	100	?
No	Divorced	73	?

Table 2: Test Data.

Att1	Att2	Attr3	Class
2	6	5	1
6	3	7	1
8	4	2	0
9	3	3	1
10	5	9	0
1	1	1	0

Table 3: Data for k-NN Classifier.