Problem Set 06

1. Commercials (6p)

Let's manually do a 5-fold cross-validation to get an error estimation of predicting the gender for the dataset below. You need multiple arff files. Let WEKA build the decision trees for each fold with the training data using the J48 algorithm, then compute the error of the test data. Now use WEKA for the whole 5-fold cross-validation and compare the results (accuracy and confusion matrix).

Sample	Watches advertisement in/on				Age	Sex
	Magazine	Paper	Internet	TV		
S01	Yes	No	No	No	45	Male
S02	Yes	Yes	Yes	Yes	40	Female
S03	No	No	No	No	42	Male
S04	Yes	Yes	Yes	Yes	30	Male
S05	Yes	Yes	Yes	No	38	Female
S06	No	No	No	No	55	Female
S07	Yes	Yes	Yes	Yes	35	Male
S08	No	No	No	No	27	Male
S09	Yes	No	No	No	43	Male
S10	Yes	Yes	Yes	No	41	Female

2. Books (4p)

A dataset about 76 booklovers shows some information (gender, age, number of books, likes Dan Brown, …, bought your bestseller). You want to send publicity to other customers who might be interested in your book. Based on the known readers you build decision trees using the k-fold cross-validation. Which of these trees do you use to determine which person to send advertisement to?

What's the size of the first, sixth, seventh, and tenth test set if you are using 10-fold cross-validation?

Deadline:

November 14, 2016 at 8:00 AM