CS 481, Homework 4

Out: April 25, 2017, Due: May 01, 2017 (10:30 am, in class), Total: 90

Note:

- This homework will carry 9 points towards your final score
- Please answer all the questions below. Please type or write legibly.
- Homeworks are individual work, please do not collaborate with others inside or outside of the class.
- Please email the instructor or use the office hours for any questions.

Questions

1 (30 points). In this assignment we will use scikit-learn SVM tool for classification. You can learn more about this tool from `http://scikit-learn.org/stable/modules/svm.html`. We will use "Congressional Voting Records Data Set" available from UCI Machine Learning Repository. The dataset can be downloaded from `http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records`. This is a classification dataset with two classes, namely democratic, and republican. It has 16 features, all of them binary (y value represents yes, n value represents no, and ? value represents withdrawal from voting).

a. Convert this dataset into numeric by converting y to 1, n to -1 and ? to 0.

b. Break the dataset into 4 folds with approximately similar ratio of the classes (republican/democratic) in each fold. Use 1 fold for parameter tuning only. For the remaining three folds, report average 3-fold classification accuracy (along with standard deviation) for Linear SVM with soft-margin classifier.

c. Now use SVM with Gaussian Kernel for the same task.

2 (35 points). In this assignment we will use itemeset mining for building features for classification. We will use the "Congressional Voting Records Data Set" for this task also. Our task is to convert the dataset into an itemset dataset and then to extract classification association rule. You can make this conversion by using the following simple python script.

```
f = open(’house-votes-84.data’,’r’)
lines = f.readlines()
for line in lines:
        strpline = line.rstrip()
        arr = strpline.split(’,’)
        newline = [];
        for i in range(len(arr)):
                if arr[i] == ’y’:
                        newline.append(i)
        if arr[0] == ’republican’:
                newline.append(100)
        else:
                newline.append(200)
        print(*newline, sep=’,’)
```

Output of the above program is an itemset dataset. In this dataset we have listed each record as itemset of the ’y’ valued features. Each feature is represented with an integer between 1 and 16. We have also make the class label as one of the item (for republican we have used 100, for democratic we have used 200). Now download an implementation of Eclat algorithm from (`http://www.borgelt.net/eclat.html`). You can use this algorithm for finding frequent itemset from this dataset. Now, answer the following questions:

a. Run the itemset mining algorithm with 20% support. How many frequent itemsets are there?

b. Write top 10 itemsets (in terms of highest support value).

c. How many frequent itemsets have 100 as part of itemsets?

d. How many frequent itemsets have 200 as part of itemsets?

e. Write top 10 association rules (in terms of highest confidence value) where the rule’s head is 100

f. How many rules with head 100 are there for which the confidence value is more than 75%? List them. For this you need to write a small script that finds confidence of a rule from the support value of its body and the support value of its body plus head.

g. Write top 10 association rules (in terms of highest confidence value) where the rule’s head is 200

h. How many rules with head 200 are there for which the confidence value is more than 75%? List them.

i. Use the rules (which has more than 75% confidence) as binary feature and construct a new dataset, in which each rule is a feature and any transaction that has the body of the rule will have a feature value of 1 and if it does not have the body of the rule will have a feature value 0. Use soft-margin SVM with linear kernel to report 3-fold classification accuracy (after using 1 fold for parameter tuning). Report both average and standard deviation.

3 (10 points). Report the best 10 features in terms of odd-ratio for predicting republican and best 10 features in terms of odd-ratio for predicting democrat.

4 (15 points). For the following dataset, use the FP-growth method for finding frequent patterns with a minimum support value of 3. Show the FP-tree and all the refined FP-trees.

1	ABDE
2	ABCE
3	BCE
4	ABCD
5	ABCDE
6	ABDE

1 Deliverables

No code submission is necessary. However, you need to submit the printout of your scripts in paper. For question 1, submit your python script for calling SVM from scikit-learn with your best parameter values. Also submit the output of your execution. For the remaining questions, simply submit your answers of the above questions. Note that submission deadline is **in-class**.