CSC 5800 : Intelligent Systems

Homework 5

Due Date: December 9, 2018 **Total: 100 Points**

Problem 1. Clustering

(15 Points; 10 + 5)

Consider the following is a set of one-dimensional points: {6, 12, 18, 24, 30, 42, 48}

(a) Show both the clusters and the total squared error for each set of initial centroids using K-means algorithm. (i) {18,45} and (ii) {15,40}. Also, compute the Cohesion and Separation values.

(b) What are the two clusters produced by single-linkage clustering?

Problem 2. Hierarchical Clustering

(10 Points; 5 + 5)

Use the similarity matrix in the following Table to perform single and complete linkage hierarchical clustering. Show your results by drawing a dendrogram that will clearly show the order in which the points are merged. Also, give the updated similarity matrix after each merge.

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

Problem 3. Association Analysis

(20 Points; 8 + 12)

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

(a) Draw a contingency table for each of the following rules using the transactions shown in Table above:

$$\text{Rules: } \{b\} \longrightarrow \{c\}, \{a\} \longrightarrow \{d\}, \{b\} \longrightarrow \{d\}, \{e\} \longrightarrow \{c\}, \{c\} \longrightarrow \{a\}.$$

(b) Use the contingency tables in part (a) to compute and rank the rules in decreasing order according to support and confidence

Problem 4. WEKA – K-means Clustering

(15 Points)

Load iris.arff file into Weka. Click on the *Cluster* tab and choose "SimpleKMeans" algorithm for clustering and set "numClusters" to 3. Select "Classes to cluster evaluation" and click on the "Ignore attributes" and select "class". Start the clustering.

(a) How many instances were clustered incorrectly?
(b) How many instances are in cluster2? How many of these instances were incorrectly clustered and which cluster they should belong to?
(c) Right-click on the result list and click on "visualize cluster assignments". Set the x-axis to instance_numbr and y-axis to sepallength. Change the color to class. Which type of iris flower has all instances clustered correctly?

Problem 5. WEKA – Association Rule-Learner

(15 Points)

This exercise aims to familiarize you with an association rule algorithm using Weka. We want to discover association rules in a given dataset by invoking the Apriori algorithm. First thing is to make sure that all the attributes in the dataset are of type nominal, in other words there should not be any numerical attribute(s) in the dataset and if there are any, they should be discretized before you can perform the association rule algorithm.

(a) Load zoo.arff into Weka. Find the numerical attribute(s) in this dataset and choose the right filter under *Filter* to discretize the numerical attribute(s). Report the attribute(s) and the filter you used.
(b) Now, go to the *Associate* tab and choose Apriori algorithm and start the algorithm. List 4 interesting rules.
(c) Which rule(s) are going to be always true according to the algorithm and the confidence?

Problem 6. Match the Following: (Find the closest possible match.)

(25 Points)

	COLUMN A	Ans.		COLUMN B
1.	Complete Linkage		a.	Impurity measure
2.	Covariance		b.	Expectation Maximization
3.	Perceptron		c.	Attribute Transformation
4.	Centroids		d.	Boosting
5.	Nearest Neighbors		e.	Hierarchical Clustering
6.	Overfitting		f.	Multidimensional Scaling
7.	Attribute Independence		g.	Model Comparison
8.	Document Similarity		h.	Rule Growing
9.	z-score normalization		i.	Patient ID
10.	Percentile Plot		j.	Backpropagation
11.	Confidence		k.	Apriori Principle
12.	Nominal Attribute		l.	Lazy Learners
13.	Sample re-weighting		m.	Partitional Clustering
14.	Curse of Dimensionality		n.	Naïve Bayes
15.	Kappa Statistic		o.	Height of a Person
16.	Lagrange Multipliers		p.	Support Vector Machines
17.	Parametric Model		q.	Sigmoid Unit
18.	Entropy		r.	Asymmetric Attributes
19.	Ordinal Attribute		s.	Bootstrap
20.	Artificial Neural Networks		t.	Correlation
21.	Sampling		u.	Density-based Clustering
22.	Frequent Itemset		v.	Cumulative Distribution
23.	Quantitative Attribute		w.	Occam's Razor
24.	Core Points		x.	Class Grades
25.	Sequential Covering		y.	Rule Evaluation

NOTE: Problems 7, 8 and 9 are extra-credit.

Problem 7. WEKA - Statistics and Visualization

(20 Points)

The goal of this exercise is to familiarize you with statistics and visualizations in Weka.

(a) Load labor.arff file into Weka. Click on the "preprocess" tab and use the histogram and statistics in this panel to answer the following questions:

i. Give the number of instances, features and class labels in this dataset.
ii. What is the attribute type of "vacation" and how many distinct vales does it have?
iii. What is the range of attribute "working-hours" and what is its standard deviation?
iv. Considering "working-hours" attribute, how many instances are classified as *good* for the range of (31.333 to 33.5) and how many are classified as *bad*? (where good and bad are the class labels)

(b) The *Visualize* panel helps you visualize a dataset- not the result of a classification or clustering model, but the dataset itself. It displays a matrix of two-dimensional scatter plots of every pair of attributes. Using the same dataset as in part (a) click on the visualize tab and answer the following questions:

i. Suppose we have a classification problem and our feature space consists of only two attributes: "wage-increase-second-year" and "wage-increase-first-year". Furthermore, suppose we are using *KNN* as our classifier and $k=5$. What class labels will be assigned to data points (4.5, 4.5) and (3, 0)?
ii. In this panel use *Jitter* on different scatter plots and describe briefly what it does and when it is useful?
iii. Now load iris.arff file into Weka and go to the *Visualize* panel. Examine different scatter plots for this dataset. Which of the numeric attributes are highly correlated and which ones are not correlated?

Problem 8. WEKA – Attribute Selection

(15 Points)

This exercise aims to familiarize you with the feature selection algorithm(s) implemented in Weka. Load labor.arff file into Weka.

(a) Using the preprocess tab, select 6 attributes arbitrarily (including the class label). Now, use this dataset and a classifier of your choice to be trained on this dataset. Report the attributes you selected and classification error using the 10 fold cross-validation as evaluation method.

(b) Re-load the dataset to get all the 17 attributes again. We will now find the most informative features using the attribute selection algorithm in Weka as follows:

i. Click on "select attributes" tab and choose "infoGainAttributeEval" under "Attribute Evaluator". Note that "Ranker" search method will be chosen

whenever you choose this attribute evaluator. Now click on start button. The output shows the attributes ranked by information gain. We now use this metric (information gain) to choose the 5 best attributes for our model building purpose.

ii. Now repeat part (a) but this time use the 5 best features from the previous step and make sure you use the same exact classifier you used in part (a). Report the attributes selected and the error rate. Did the classifier improve?

Problem 9. WEKA – Classification and Clustering

(25 Points)

The goal of this exercise is to introduce the concept of semi-supervised classification and using it to improve the classification performance. Semi-supervised classification is a classification task where the training set contains a limited number of labeled examples but a large number of unlabeled examples. Instead of building a classification model using the labeled examples alone, the goal of semi-supervised classification is to use the additional unlabeled examples to improve classification performance. For this exercise, you will use Weka to perform semi-supervised classification and compare its performance against using labeled examples only.

(a) Download the three data sets(trainfile.arff, testfile.arff, mergrefile.arff) from class website. The trainfile.arff file contains only labeled examples. The mergefile.arff file contains both labeled and unlabeled examples. You will use the mergefile.arff file for semi-supervised classification. The testfile.arff file contains data points for evaluating the performance of your induced model. The class labels are denoted as 1, 2 and 3.

(b) In this exercise, you will investigate the performance of the classifier trained using labeled samples only. You will use the Naïve Bayes classifier for model building. After loading the files, select Naïve Bayes classifier and train the classifier using trainfile.arff and re-evaluate it on testfile.arff. Write down the confusion matrix generated by the classifier. What is the error rate of the classifier?

(c) In this exercise, you will combine labeled and unlabeled examples for building a model. The basic idea is as follows: first, you will use k-means algorithm to cluster together the labeled and unlabeled examples. For each cluster, you will determine its class label by choosing the majority class labeled examples assigned to the cluster. You will then label the unlabeled examples using the majority class of its corresponding cluster. Finally, you will construct a naïve Bayes classifier on the new data set consisting of old labeled examples and newly labeled examples.

i. Load the meregfile.arff file into Weka. Note that the examples are labeled as 1, 2, 3 and 4(where 4 corresponds to unlabeled examples)

ii. Click on cluster tab. Select Simple k-means to be your clustering algorithm. Set numclusters = 6 and make sure "classes to cluster evaluation" is checked under the cluster mode.

iii. Click on the start button to perform the clustering and right click on the "Result list".

iv. You need to save the cluster assignment for each record. To do this, click on "visualize classifier assignment". Save the resulting cluster assignment in a file called clusterResult.arff. This file looks exactly the same as mergefile.arff file except the first column contains the instance number and the last column contains the cluster Id. From the confusion matrix produced by the k-means algorithm, determine which class label (1,2 or 3) should you assign to each cluster Id. Use the majority class of the labeled examples to assign the class label for each cluster. (remember: 4 is not a valid class label.)

v. Open the clusterResult.arff file with a text editor. Replace the cluster labels(i.e cluster0, cluster1, etc) in the last column with their corresponding class names found in the previous step(i.e based on the majority class of the labeled examples for the cluster.) save your changes.

vi. Load the modified clusterResult.arff file into Weka. Under the "preprocess" tab, remove the "instance number" and "class" fields. Save the data set as newTraining.arff file.

vii. Open the newTraining.arff file and test.arff files using Word-pad or you favorite text editor. Modify the declaration at the beginning of newTraining.arff file so that it has the same relation name and fields as your test file. In other words, you need to change the name of the relation to : @relation cmc and modify @attribute Cluster{cluster0, cluster1,……}to @attribute class{1, 2, 3} Save your changes.

viii. The previous step will re-label the unlabeled examples with the class label of its corresponding cluster. In this step, you will perform classification on this larger data set. Load the newTraining.arff file into Weka and run naïve Bayes classifier corresponding to the steps given in part (a). Write down its confusion matrix and error rate of the new classifier.

ix. Compare the error rates of the two naïve Bayes models you had obtained. For this data set, does semi-supervised classification improve your classification performance.