Homework 2

Classification for Matrix Data

1. Decision Tree

Construct a decision tree for the following training data, where "Edible" is the class we are going to predict. Information gain is used to select the attributes. Please write down the major steps in the construction process (you need to show the information gain for each candidate attribute when a new node is created in the tree).

Color	Size	Shape	Edible?
Yellow	Small	Round	+
Yellow	Small	Round	-
Green	Small	Irregular	+
Green	Large	Irregular	-
Yellow	Large	Round	+
Yellow	Small	Round	+
Yellow	Small	Round	+
Yellow	Small	Round	+
Green	Small	Round	-
Yellow	Large	Round	-
Yellow	Large	Round	+
Yellow	Large	Round	-
Yellow	Large	Round	-
Yellow	Large	Round	-
Yellow	Small	Irregular	+
Yellow	Large	Irregular	+

2. Support Vector Machine

#	x1	x2	class
1	2.46	2.59	1
2	3.05	2.87	1
3	1.12	1.64	1
4	0.01	1.44	1
5	2.20	3.04	1
6	0.41	2.04	1
7	0.53	0.77	1
8	1.89	2.64	1
9	-0.39	0.96	1
10	-0.96	0.08	1
11	2.65	-1.33	-1

12	1.57	-1.70	-1
13	3.05	0.01	-1
14	2.66	-1.15	-1
15	4.51	-0.52	-1
16	3.06	-0.82	-1
17	3.16	-0.56	-1
18	2.05	-0.62	-1
19	0.71	-2.47	-1
20	1.63	-0.91	-1

Given 20 data points and their class labels in the above, suppose by solving the dual form of the quadratic programming of svm, we can derive the $\alpha's$ for each data point as follows:
$\alpha_7 = 0.4952$
$\alpha_{18} = 0.0459$
$\alpha_{20} = 0.4493$
Others = 0

(1) Please point out the support vectors in the training points.
(2) Calculate the normal vector of the hyperplane: w
(3) Calculate the bias b, according to $b = \sum_{k:\alpha_k \neq 0}(y_k - w'x_k)/N_k$, where $x_k = (x_{k1}, x_{k2})'$ indicate the support vectors and N_k is the total number of support vectors.
(4) Write down the learned decision boundary function $f(x) = w'x + b$ (the hyperplane) by substituting w and b with learned values in the formula.
(5) Suppose there is a new data point $x = (-1,2)$, please use the decision boundary to predict its class label.

Clustering for Matrix Data

3. List four limitations of K-means, and name one algorithm for each limitation that can overcome that limitation.

4. Clustering Evaluation.

ID	Conference Name	Ground Truth Label	Algorithm output Label
1	IJCAI	3	2
2	AAAI	3	2
3	ICDE	1	3
4	VLDB	1	3
5	SIGMOD	1	3
6	SIGIR	4	4
7	ICML	3	2
8	CVPR	3	2
9	CIKM	4	3
10	KDD	2	1

11	WWW	4	4
12	PAKDD	2	1
13	PODS	1	3
14	ICDM	2	1
15	ECML	3	2
16	PKDD	2	1
17	EDBT	1	2
18	SDM	2	1
19	ECIR	4	4
20	WSDM	4	1

Suppose we want to cluster 20 above conferences into four areas, with ground truth label and algorithm output label shown in third and fourth column. Please evaluate the quality of the clustering algorithm according to purity, precision, recall, F-measure, and normalized mutual information, respectively.

Frequent Pattern Mining for Set Data

5. Textbook Chapter 6: 6.6