UECS3213 / UECS3453 DATA MINING

SESSION: January 2019

TUTORIAL 3
Chapter 3 - Classification

k-NN

1. Consider the one-dimensional data set shown in Table 5.4.

Table 5.4. Data set

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	−	−	+	+	+	−	−	+	−	−

a) Classify the data point x = 5.0 according to its 1-, 3-, 5-, and 9-nearest neighbors (using majority vote).

b) Repeat the previous analysis using the *distance-weighted voting* approach.

Decision Tree

2. Consider the training examples shown in Table 4.1 for a binary classification problem.

Table 4.1. Data set

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

a) Compute the Gini index for the overall collection of training examples.

b) Compute the Gini index for the Gender attribute.

c) Compute the Gini index for the Car Type attribute using multiway split.

d) Compute the Gini index for the Shirt Size attribute using multiway split.

e) Which attribute is **better**, Gender, Car Type, or Shirt Size?

3. Consider the training examples shown in Table 4.2 for a binary classification problem.

Table 4.2. Data set

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

a) What is the *entropy* of this collection of training examples with respect to the positive class?

b) What are the *information gains* of a_1 and a_2 relative to these training examples?

c) For a_3, which is a continuous attribute, compute the *information gain* for every possible split.

d) What is the best split (among a_1, a_2, and a_3) according to the information gain?

e) What is the best split (between a_1 and a_2) according to the classification error rate?

f) What is the best split (between a_1 and a_2) according to the Gini index?

Naive Bayes

4. Consider the data set shown in Table 5.1

Table 5.1. Data set

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

a) Estimate the *conditional probabilities* for P(A|+), P(B|+), P(C|+), P(A|−), P(B|−), and P(C|−).

b) Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample (A = 0, B = 1, C = 0) using the Naive Bayes approach.

c) Estimate the *conditional probabilities* using the m-estimate approach, with p = 1/2 and m = 4.

d) Repeat part (b) using the conditional probabilities given in part (c).

e) Compare the two methods for estimating probabilities. Which method is **better** and why?

5. Consider the data set shown in Table 5.2.

Table 5.2. Data set for Exercise 8.

Instance	A	B	C	Class
1	0	0	1	−
2	1	0	1	+
3	0	1	0	−
4	1	0	0	−
5	1	0	1	+
6	0	0	1	+
7	1	1	0	−
8	0	0	0	−
9	0	1	0	+
10	1	1	1	+

a) Estimate the conditional probabilities for P(A = 1|+), P(B = 1|+), P(C = 1|+), P(A = 1|–), P(B = 1|–), and P(C = 1|–) using the same approach as in the previous problem.

b) Use the conditional probabilities in part (a) to predict the class label for a test sample (A = 1,B = 1, C = 1) using the naive Bayes approach.

c) Compare P(A = 1), P(B = 1), and P(A = 1,B = 1). State the relationships between A and B.

d) Repeat the analysis in part (c) using P(A = 1), P(B = 0), and P(A = 1,B = 0).

e) Compare P(A = 1,B = 1|Class = +) against P(A = 1|Class = +) and P(B = 1|Class = +). Are the variables *conditionally independent* given the class?

Support Vector Machine

6. Suppose you are using a Linear SVM classifier with 2 class classification problem. Now you have been given the following data in which some points are circled red that are representing support vectors.
   a) If you remove the following any one red points from the data. Does the decision boundary will change? Yes/No
   b) If you remove the non-red circled points from the data, the decision boundary will change? True/False

7. What do we mean by generalization error in terms of the SVM?

   a) How far the hyperplane is from the support vectors
   b) How accurately the SVM can predict outcomes for unseen data
   c) The threshold amount of error in an SVM

8. What do we mean by a hard margin?

    a) The SVM allows very low error in classification
    b) The SVM allows high amount of error in classification
    c) None of the above

9. The SVM's are less effective when:

    a) The data is linearly separable
    b) The data is clean and ready to use
    c) The data is noisy and contains overlapping points

10. Which of the following are real world applications of the SVM?

    a) Text and Hypertext Categorization
    b) Image Classification
    c) Clustering of News Articles
    d) All of the above

The End