Data Mining

Problem Set 4

Chapter 5 (pg 315)

1. Consider a binary classification problem with the following set of attributes and attribute values:

- Air Conditioner = {Working, Broken}
- Engine = {Good, Bad}
- Mileage = {High, Medium, Low}
- Rust = {Yes, No}

Suppose a rule-based classifier produces the following rule set:

Mileage = High → Value = Low
Mileage = Low → Value = High
Air Conditioner = Working, Engine = Good → Value = High
Air Conditioner = Working, Engine = Bad → Value = Low
Air Conditioner = Broken → Value = Low

(a) Are the rules mutually exclusive?
(b) Is the rule set exhaustive?
(c) Is ordering needed for this set of rules?
(d) Do you need a default class for the rule set?

5. **Figure 5.1** illustrates the coverage of the classification rules R1, R2, and R3. Determine which is the best and worst rule according to:

(a) The likelihood ratio statistic.
(b) The Laplace measure.
(c) The m-estimate measure (with $k = 2$ and $p_+ = 0.58$).
(d) The rule accuracy after R1 has been discovered, where none of the examples covered by R1 are discarded).
(e) The rule accuracy after R1 has been discovered, where only the positive examples covered by R1 are discarded).

(f) The rule accuracy after R1 has been discovered, where both positive and negative examples covered by R1 are discarded.

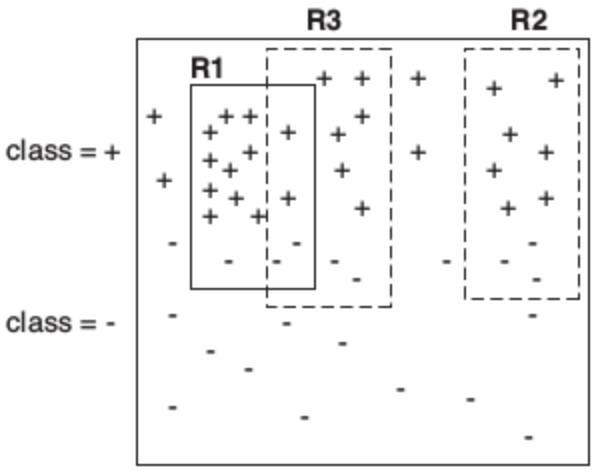


Figure 5.1. Elimination of training records by the sequential covering algorithm. $R1$, $R2$, and $R3$ represent regions covered by three different rules.

6. Answer the following probability questions about student smokers.

   (a) Suppose the fraction of undergraduate students who smoke is 15% and the fraction of graduate students who smoke is 23%. If one-fifth of the college students are graduate students and the rest are undergraduates, what is the probability that a student who smokes is a graduate student?
   (b) Given the information in part (a), is a randomly chosen college student more likely to be a graduate or undergraduate student?
   (c) Repeat part (b) assuming that the student is a smoker.
   (d) Suppose 30% of the graduate students live in a dorm but only 10% of the undergraduate students live in a dorm. If a student smokes and lives in the dorm, is he or she more likely to be a graduate or undergraduate student? You can assume independence between students who live in a dorm and those who smoke.

7. Consider the data set shown in **Table 5.1**

   (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?

Table 5.1 Data set for Exercise 7

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	+

9. Consider the plot shown in **Figure 5.2**

   (a) Explain how naive Bayes performs on the data set shown in **Figure 5.2.**
   (b) If each class is further divided such that there are four classes (A1, A2, B1, and B2), will naive Bayes perform better?
   (c) How will a decision tree perform on this data set (for the two-class problem)? What if there are four classes?

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

Table 5.4 Data set for Exercise 13

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

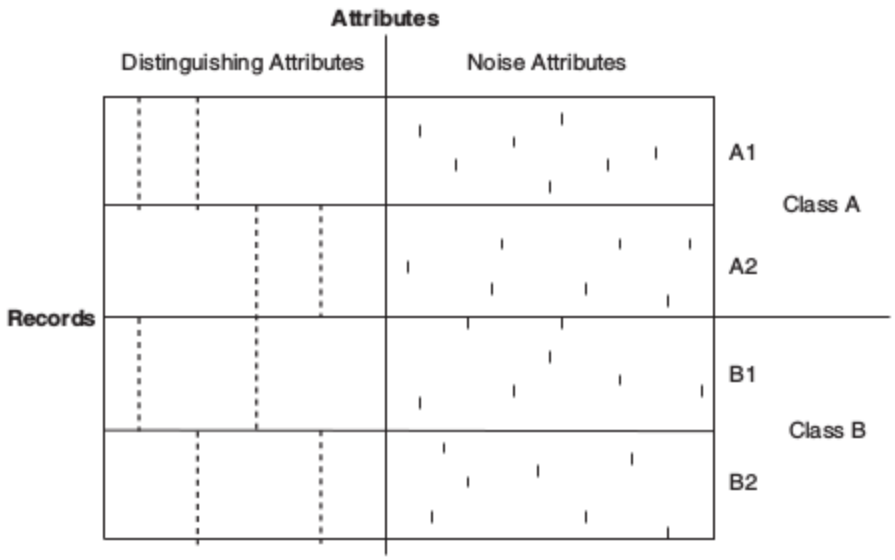


Figure 5.2. Data set for Exercise 9.

(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 described in Section 5.2.1.

16. Answer the following questions about neural networks.

(a) Demonstrate how the perceptron model can be used to represent the AND and OR functions between a pair of Boolean variables.
(b) Comment on the disadvantage of using linear functions as activation functions for multilayer neural networks.

22. Consider the XOR problem where there are four training points:

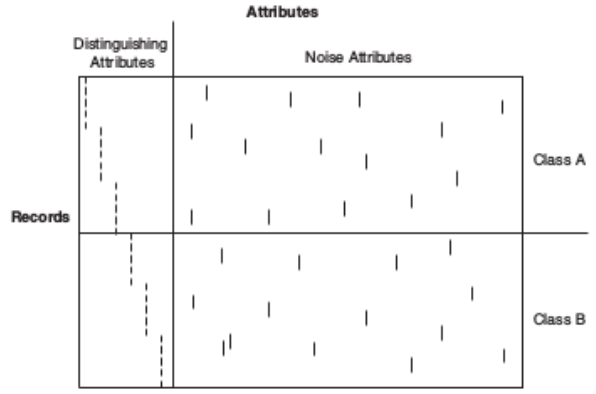
$$(1,1,-),\ (1,0,+),\ (0,1,+),\ (0,0,-)$$

Transform the data into the following feature space:

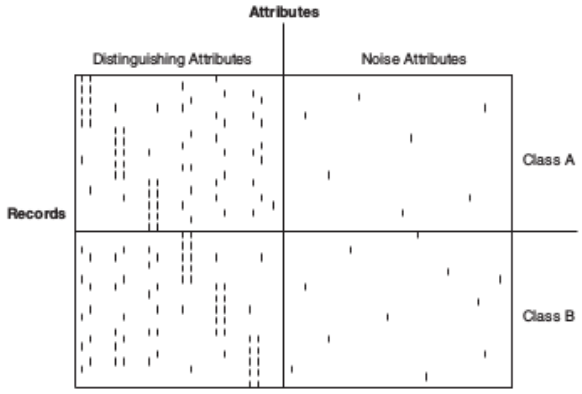
$$\Phi = (1,\ \sqrt{2}x_1,\ \sqrt{2}x_2,\ \sqrt{2}x_1x_2,\ x_1^2,\ x_2^2).$$

Find the maximum margin linear decision boundary in the transformed space.

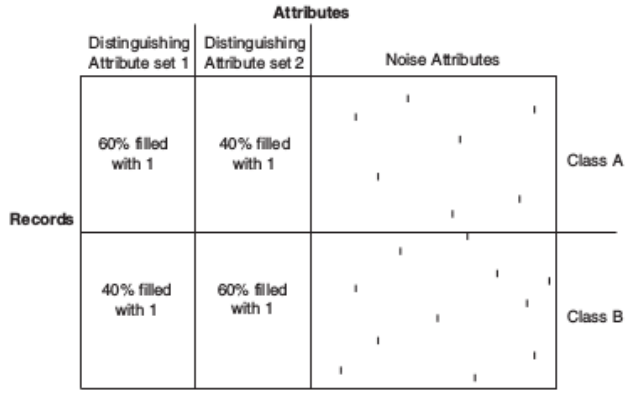
23. Given the data sets shown in **Figures 5.6**, explain how the decision tree, naive Bayes, and k-nearest neighbor classifiers would perform on these data sets.



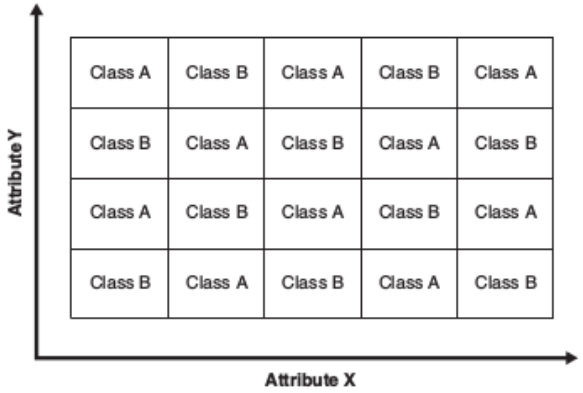
(a) Synthetic data set 1.



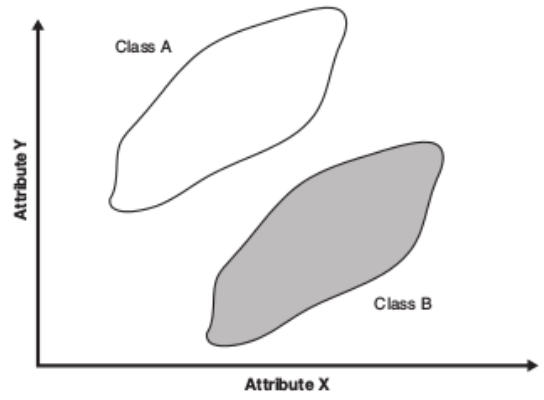
(b) Synthetic data set 2.



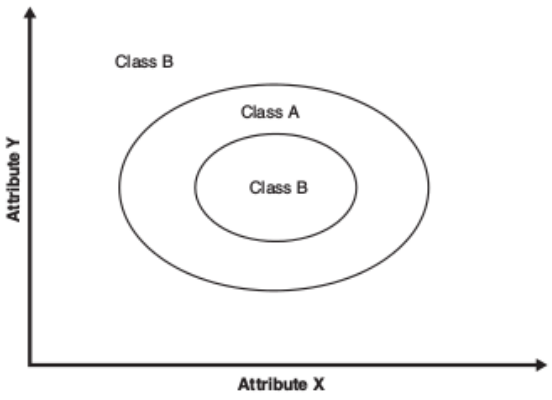
(c) Synthetic data set 3.



(d) Synthetic data set 4



(e) Synthetic data set 5.



(f) Synthetic data set 6.

Figure 5.6. Data set for Exercise 23.