5

Classification: Alternative Techniques

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 $\longrightarrow$ Value = Low
Mileage = Low $\longrightarrow$ Value = High
Air Conditioner = Working, Engine = Good $\longrightarrow$ Value = High
Air Conditioner = Working, Engine = Bad $\longrightarrow$ Value = Low
Air Conditioner = Broken $\longrightarrow$ Value = Low

   (a) Are the rules mutually exclustive?
   Answer: No

   (b) Is the rule set exhaustive?
   Answer: Yes

   (c) Is ordering needed for this set of rules?
   Answer: Yes because a test instance may trigger more than one rule.

   (d) Do you need a default class for the rule set?
   Answer: No because every instance is guaranteed to trigger at least one rule.

2. The RIPPER algorithm (by Cohen [1]) is an extension of an earlier algorithm called IREP (by Fürnkranz and Widmer [3]). Both algorithms apply the **reduced-error pruning** method to determine whether a rule needs to be pruned. The reduced error pruning method uses a validation set to estimate the generalization error of a classifier. Consider the following pair of rules:

$$R_1: \quad A \longrightarrow C$$
$$R_2: \quad A \wedge B \longrightarrow C$$

R_2 is obtained by adding a new conjunct, B, to the left-hand side of R_1. For this question, you will be asked to determine whether R_2 is preferred over R_1 from the perspectives of rule-growing and rule-pruning. To determine whether a rule should be pruned, IREP computes the following measure:

$$v_{IREP} = \frac{p + (N - n)}{P + N},$$

where P is the total number of positive examples in the validation set, N is the total number of negative examples in the validation set, p is the number of positive examples in the validation set covered by the rule, and n is the number of negative examples in the validation set covered by the rule. v_{IREP} is actually similar to classification accuracy for the validation set. IREP favors rules that have higher values of v_{IREP}. On the other hand, RIPPER applies the following measure to determine whether a rule should be pruned:

$$v_{RIPPER} = \frac{p - n}{p + n}.$$

(a) Suppose R_1 is covered by 350 positive examples and 150 negative examples, while R_2 is covered by 300 positive examples and 50 negative examples. Compute the FOIL's information gain for the rule R_2 with respect to R_1.

Answer:

For this problem, $p_0 = 350$, $n_0 = 150$, $p_1 = 300$, and $n_1 = 50$. Therefore, the FOIL's information gain for R_2 with respect to R_1 is:

$$Gain = 300 \times \left[\log_2 \frac{300}{350} - \log_2 \frac{350}{500} \right] = 87.65$$

(b) Consider a validation set that contains 500 positive examples and 500 negative examples. For R_1, suppose the number of positive examples covered by the rule is 200, and the number of negative examples covered by the rule is 50. For R_2, suppose the number of positive examples covered by the rule is 100 and the number of negative examples is 5. Compute v_{IREP} for both rules. Which rule does IREP prefer?

Answer:

For this problem, $P = 500$, and $N = 500$.

For rule $R1$, $p = 200$ and $n = 50$. Therefore,

$$V_{IREP}(R1) = \frac{p + (N - n)}{P + N} = \frac{200 + (500 - 50)}{1000} = 0.65$$

For rule $R2$, $p = 100$ and $n = 5$.

$$V_{IREP}(R2) = \frac{p + (N - n)}{P + N} = \frac{100 + (500 - 5)}{1000} = 0.595$$

Thus, IREP prefers rule R1.

(c) Compute v_{RIPPER} for the previous problem. Which rule does RIPPER prefer?

Answer:

$$V_{RIPPER}(R1) = \frac{p - n}{p + n} = \frac{150}{250} = 0.6$$

$$V_{RIPPER}(R2) = \frac{p - n}{p + n} = \frac{95}{105} = 0.9$$

Thus, RIPPER prefers the rule R2.

3. C4.5rules is an implementation of an indirect method for generating rules from a decision tree. RIPPER is an implementation of a direct method for generating rules directly from data.

   (a) Discuss the strengths and weaknesses of both methods.

   Answer:

   The C4.5 rules algorithm generates classification rules from a global perspective. This is because the rules are derived from decision trees, which are induced with the objective of partitioning the feature space into homogeneous regions, without focusing on any classes. In contrast, RIPPER generates rules one-class-at-a-time. Thus, it is more biased towards the classes that are generated first.

   (b) Consider a data set that has a large difference in the class size (i.e., some classes are much bigger than others). Which method (between C4.5rules and RIPPER) is better in terms of finding high accuracy rules for the small classes?

   Answer:

   The class-ordering scheme used by C4.5rules has an easier interpretation than the scheme used by RIPPER.

4. Consider a training set that contains 100 positive examples and 400 negative examples. For each of the following candidate rules,

   R_1: $A \longrightarrow +$ (covers 4 positive and 1 negative examples),
   R_2: $B \longrightarrow +$ (covers 30 positive and 10 negative examples),
   R_3: $C \longrightarrow +$ (covers 100 positive and 90 negative examples),

   determine which is the best and worst candidate rule according to:

   (a) Rule accuracy.

   Answer:

   The accuracies of the rules are 80% (for R_1), 75% (for R_2), and 52.6% (for R_3), respectively. Therefore R_1 is the best candidate and R_3 is the worst candidate according to rule accuracy.

   (b) FOIL's information gain.

   Answer:

   Assume the initial rule is $\emptyset \longrightarrow +$. This rule covers $p_0 = 100$ positive examples and $n_0 = 400$ negative examples.

   The rule R_1 covers $p_1 = 4$ positive examples and $n_1 = 1$ negative example. Therefore, the FOIL's information gain for this rule is

   $$4 \times \left(\log_2 \frac{4}{5} - \log_2 \frac{100}{500} \right) = 8.$$

   The rule R_2 covers $p_1 = 30$ positive examples and $n_1 = 10$ negative example. Therefore, the FOIL's information gain for this rule is

   $$30 \times \left(\log_2 \frac{30}{40} - \log_2 \frac{100}{500} \right) = 57.2.$$

   The rule R_3 covers $p_1 = 100$ positive examples and $n_1 = 90$ negative example. Therefore, the FOIL's information gain for this rule is

   $$100 \times \left(\log_2 \frac{100}{190} - \log_2 \frac{100}{500} \right) = 139.6.$$

   Therefore, R_3 is the best candidate and R_1 is the worst candidate according to FOIL's information gain.

   (c) The likelihood ratio statistic.

   Answer:

   For R_1, the expected frequency for the positive class is $5 \times 100/500 = 1$ and the expected frequency for the negative class is $5 \times 400/500 = 4$. Therefore, the likelihood ratio for R_1 is

   $$2 \times \left[4 \times \log_2(4/1) + 1 \times \log_2(1/4) \right] = 12.$$

For R_2, the expected frequency for the positive class is $40 \times 100/500 = 8$ and the expected frequency for the negative class is $40 \times 400/500 = 32$. Therefore, the likelihood ratio for R_2 is

$$2 \times \left[30 \times \log_2(30/8) + 10 \times \log_2(10/32)\right] = 80.85$$

For R_3, the expected frequency for the positive class is $190 \times 100/500 = 38$ and the expected frequency for the negative class is $190 \times 400/500 = 152$. Therefore, the likelihood ratio for R_3 is

$$2 \times \left[100 \times \log_2(100/38) + 90 \times \log_2(90/152)\right] = 143.09$$

Therefore, R_3 is the best candidate and R_1 is the worst candidate according to the likelihood ratio statistic.

(d) The Laplace measure.

Answer:

The Laplace measure of the rules are 71.43% (for R_1), 73.81% (for R_2), and 52.6% (for R_3), respectively. Therefore R_2 is the best candidate and R_3 is the worst candidate according to the Laplace measure.

(e) The m-estimate measure (with $k = 2$ and $p_+ = 0.2$).

Answer:

The m-estimate measure of the rules are 62.86% (for R_1), 73.38% (for R_2), and 52.3% (for R_3), respectively. Therefore R_2 is the best candidate and R_3 is the worst candidate according to the m-estimate measure.

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.

Answer:

There are 29 positive examples and 21 negative examples in the data set. $R1$ covers 12 positive examples and 3 negative examples. The expected frequency for the positive class is $15 \times 29/50 = 8.7$ and the expected frequency for the negative class is $15 \times 21/50 = 6.3$. Therefore, the likelihood ratio for $R1$ is

$$2 \times \left[12 \times \log_2(12/8.7) + 3 \times \log_2(3/6.3)\right] = 4.71.$$

$R2$ covers 7 positive examples and 3 negative examples. The expected frequency for the positive class is $10 \times 29/50 = 5.8$ and the expected

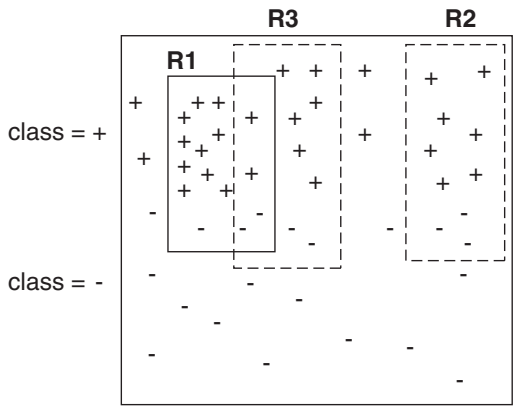


Figure 5.1. Elimination of training records by the sequential covering algorithm. $R1$, $R2$, and $R3$ represent regions covered by three different rules.

frequency for the negative class is $10 \times 21/50 = 4.2$. Therefore, the likelihood ratio for $R2$ is

$$2 \times \left[7 \times \log_2(7/5.8) + 3 \times \log_2(3/4.2) \right] = 0.89.$$

$R3$ covers 8 positive examples and 4 negative examples. The expected frequency for the positive class is $12 \times 29/50 = 6.96$ and the expected frequency for the negative class is $12 \times 21/50 = 5.04$. Therefore, the likelihood ratio for $R3$ is

$$2 \times \left[8 \times \log_2(8/6.96) + 4 \times \log_2(4/5.04) \right] = 0.5472.$$

$R1$ is the best rule and $R3$ is the worst rule according to the likelihood ratio statistic.

(b) The Laplace measure.

Answer:

The Laplace measure for the rules are 76.47% (for $R1$), 66.67% (for $R2$), and 64.29% (for $R3$), respectively. Therefore $R1$ is the best rule and $R3$ is the worst rule according to the Laplace measure.

(c) The m-estimate measure (with $k = 2$ and $p_+ = 0.58$).

Answer:

The m-estimate measure for the rules are 77.41% (for $R1$), 68.0% (for $R2$), and 65.43% (for $R3$), respectively. Therefore $R1$ is the best rule and $R3$ is the worst rule according to the m-estimate measure.

(d) The rule accuracy after $R1$ has been discovered, where none of the examples covered by $R1$ are discarded).

Answer:

If the examples for $R1$ are not discarded, then $R2$ will be chosen because it has a higher accuracy (70%) than $R3$ (66.7%).

(e) The rule accuracy after $R1$ has been discovered, where only the positive examples covered by $R1$ are discarded).

Answer:

If the positive examples covered by $R1$ are discarded, the new accuracies for $R2$ and $R3$ are 70% and 60%, respectively. Therefore $R2$ is preferred over $R3$.

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

Answer:

If the positive and negative examples covered by $R1$ are discarded, the new accuracies for $R2$ and $R3$ are 70% and 75%, respectively. In this case, $R3$ is preferred over $R2$.

6. (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?

Answer:

Given $P(S|UG) = 0.15$, $P(S|G) = 0.23$, $P(G) = 0.2$, $P(UG) = 0.8$. We want to compute $P(G|S)$.

According to Bayesian Theorem,

$$P(G|S) = \frac{0.23 \times 0.2}{0.15 \times 0.8 + 0.23 \times 0.2} = 0.277. \tag{5.1}$$

(b) Given the information in part (a), is a randomly chosen college student more likely to be a graduate or undergraduate student?

Answer:

An undergraduate student, because $P(UG) > P(G)$.

(c) Repeat part (b) assuming that the student is a smoker.

Answer:

An undergraduate student because $P(UG|S) > P(G|S)$.

(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.

Answer:

First, we need to estimate all the probabilities.

$P(D|UG) = 0.1$, $P(D|G) = 0.3$.
$P(D) = P(UG).P(D|UG) + P(G).P(D|G) = 0.8 * 0.1 + 0.2 * 0.3 = 0.14$.
$P(S) = P(S|UG)P(UG) + P(S|G)P(G) = 0.15 * 0.8 + 0.23 * 0.2 = 0.166$.
$P(DS|G) = P(D|G) \times P(S|G) = 0.3 \times 0.23 = 0.069$ (using conditional independent assumption)
$P(DS|UG) = P(D|UG) \times P(S|UG) = 0.1 \times 0.15 = 0.015$.
We need to compute $P(G|DS)$ and $P(UG|DS)$.

$$P(G|DS) = \frac{0.069 \times 0.2}{P(DS)} = \frac{0.0138}{P(DS)}$$

$$P(UG|DS) = \frac{0.015 \times 0.8}{P(DS)} = \frac{0.012}{P(DS)}$$

Since $P(G|DS) > P(UG|DS)$, he/she is more likely to be a graduate student.

7. Consider the data set shown in Table 5.1

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	+

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

Answer:

$P(A = 1|-) = 2/5 = 0.4$, $P(B = 1|-) = 2/5 = 0.4$,
$P(C = 1|-) = 1$, $P(A = 0|-) = 3/5 = 0.6$,
$P(B = 0|-) = 3/5 = 0.6, P(C = 0|-) = 0$; $P(A = 1|+) = 3/5 = 0.6$,
$P(B = 1|+) = 1/5 = 0.2$, $P(C = 1|+) = 2/5 = 0.4$,
$P(A = 0|+) = 2/5 = 0.4$, $P(B = 0|+) = 4/5 = 0.8$,
$P(C = 0|+) = 3/5 = 0.6$.

(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 naïve Bayes approach.

Answer:

Let $P(A = 0, B = 1, C = 0) = K$.

$$\begin{aligned} & P(+|A = 0, B = 1, C = 0) \\ = & \frac{P(A = 0, B = 1, C = 0|+) \times P(+)}{P(A = 0, B = 1, C = 0)} \\ = & \frac{P(A = 0|+)P(B = 1|+)P(C = 0|+) \times P(+)}{K} \\ = & 0.4 \times 0.2 \times 0.6 \times 0.5/K \\ = & 0.024/K. \end{aligned}$$

$$\begin{aligned} & P(-|A = 0, B = 1, C = 0) \\ = & \frac{P(A = 0, B = 1, C = 0|-) \times P(-)}{P(A = 0, B = 1, C = 0)} \\ = & \frac{P(A = 0|-) \times P(B = 1|-) \times P(C = 0|-) \times P(-)}{K} \\ = & 0/K \end{aligned}$$

The class label should be '+'.

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

Answer:

$P(A = 0|+) = (2 + 2)/(5 + 4) = 4/9$,
$P(A = 0|-) = (3 + 2)/(5 + 4) = 5/9$,
$P(B = 1|+) = (1 + 2)/(5 + 4) = 3/9$,
$P(B = 1|-) = (2 + 2)/(5 + 4) = 4/9$,
$P(C = 0|+) = (3 + 2)/(5 + 4) = 5/9$,
$P(C = 0|-) = (0 + 2)/(5 + 4) = 2/9$.

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

Answer:

Let $P(A = 0, B = 1, C = 0) = K$

$$
\begin{aligned}
& P(+|A=0,B=1,C=0) \\
= & \frac{P(A=0,B=1,C=0|+) \times P(+)}{P(A=0,B=1,C=0)} \\
= & \frac{P(A=0|+)P(B=1|+)P(C=0|+) \times P(+)}{K} \\
= & \frac{(4/9) \times (3/9) \times (5/9) \times 0.5}{K} \\
= & 0.0412/K
\end{aligned}
$$

$$
\begin{aligned}
& P(-|A=0,B=1,C=0) \\
= & \frac{P(A=0,B=1,C=0|-) \times P(-)}{P(A=0,B=1,C=0)} \\
= & \frac{P(A=0|-) \times P(B=1|-) \times P(C=0|-) \times P(-)}{K} \\
= & \frac{(5/9) \times (4/9) \times (2/9) \times 0.5}{K} \\
= & 0.0274/K
\end{aligned}
$$

The class label should be '+'.

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

Answer:

When one of the conditional probability is zero, the estimate for conditional probabilities using the m-estimate probability approach is better, since we don't want the entire expression becomes zero.

8. Consider the data set shown in Table 5.2.

(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.

Answer:

$P(A = 1|+) = 0.6$, $P(B = 1|+) = 0.4$, $P(C = 1|+) = 0.8$, $P(A = 1|-) = 0.4$, $P(B = 1|-) = 0.4$, and $P(C = 1|-) = 0.2$

(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 naïve Bayes approach.

Answer:

Let $R : (A = 1, B = 1, C = 1)$ be the test record. To determine its class, we need to compute $P(+|R)$ and $P(-|R)$. Using Bayes theorem,

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	+

$P(+|R) = P(R|+)P(+)/P(R)$ and $P(-|R) = P(R|-)P(-)/P(R)$. Since $P(+) = P(-) = 0.5$ and $P(R)$ is constant, R can be classified by comparing $P(+|R)$ and $P(-|R)$.

For this question,

$$\begin{aligned} P(R|+) &= P(A = 1|+) \times P(B = 1|+) \times P(C = 1|+) = 0.192 \\ P(R|-) &= P(A = 1|-) \times P(B = 1|-) \times P(C = 1|-) = 0.032 \end{aligned}$$

Since $P(R|+)$ is larger, the record is assigned to (+) class.

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

Answer:

$P(A = 1) = 0.5$, $P(B = 1) = 0.4$ and $P(A = 1, B = 1) = P(A) \times P(B) = 0.2$. Therefore, A and B are independent.

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

Answer:

$P(A = 1) = 0.5$, $P(B = 0) = 0.6$, and $P(A = 1, B = 0) = P(A = 1) \times P(B = 0) = 0.3$. A and B are still independent.

(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?

Answer:

Compare $P(A = 1, B = 1|+) = 0.2$ against $P(A = 1|+) = 0.6$ and $P(B = 1|Class = +) = 0.4$. Since the product between $P(A = 1|+)$ and $P(A = 1|-)$ are not the same as $P(A = 1, B = 1|+)$, A and B are not conditionally independent given the class.

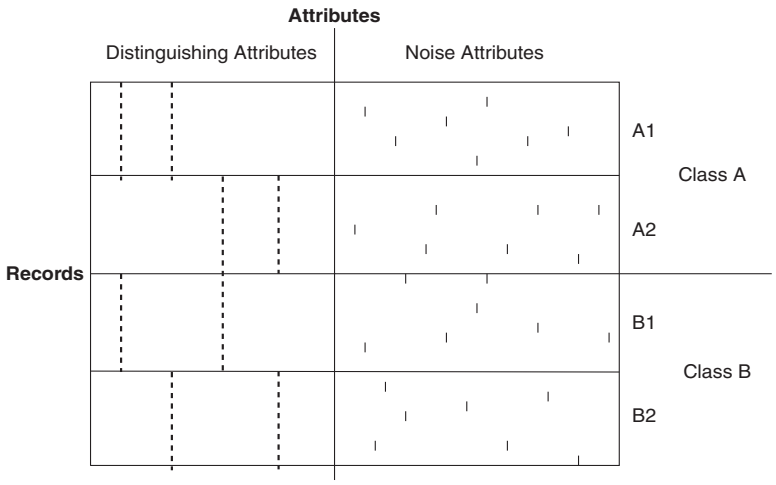


Figure 5.2. Data set for Exercise 9.

9. (a) Explain how naïve Bayes performs on the data set shown in Figure 5.2.

   Answer:

   NB will not do well on this data set because the conditional probabilities for each distinguishing attribute given the class are the same for both class A and class B.

   (b) If each class is further divided such that there are four classes ($A1$, $A2$, $B1$, and $B2$), will naïve Bayes perform better?

   Answer:

   The performance of NB will improve on the subclasses because the product of conditional probabilities among the distinguishing attributes will be different for each subclass.

   (c) How will a decision tree perform on this data set (for the two-class problem)? What if there are four classes?

   Answer:

   For the two-class problem, decision tree will not perform well because the entropy will not improve after splitting the data using the distinguishing attributes. If there are four classes, then decision tree will improve considerably.

10. Repeat the analysis shown in Example 5.3 for finding the location of a decision boundary using the following information:

    (a) The prior probabilities are $P(\texttt{Crocodile}) = 2 \times P(\texttt{Alligator})$.
    Answer: $\hat{x} = 13.0379$.

(b) The prior probabilities are $P(\texttt{Alligator}) = 2 \times P(\texttt{Crocodile})$.
Answer: $\hat{x} = 13.9621$.

(c) The prior probabilities are the same, but their standard deviations are different; i.e., $\sigma(\texttt{Crocodile}) = 4$ and $\sigma(\texttt{Alligator}) = 2$.
Answer: $\hat{x} = 22.1668$.

11. Figure 5.3 illustrates the Bayesian belief network for the data set shown in Table 5.3. (Assume that all the attributes are binary).

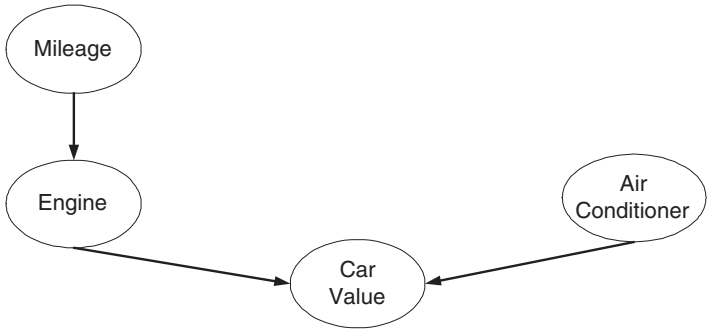


Figure 5.3. Bayesian belief network.

Table 5.3. Data set for Exercise 11.

Mileage	Engine	Air Conditioner	Number of Records with Car Value=Hi	Number of Records with Car Value=Lo
Hi	Good	Working	3	4
Hi	Good	Broken	1	2
Hi	Bad	Working	1	5
Hi	Bad	Broken	0	4
Lo	Good	Working	9	0
Lo	Good	Broken	5	1
Lo	Bad	Working	1	2
Lo	Bad	Broken	0	2

(a) Draw the probability table for each node in the network.
P(Mileage=Hi) = 0.5
P(Air Cond=Working) = 0.625
P(Engine=Good|Mileage=Hi) = 0.5
P(Engine=Good|Mileage=Lo) = 0.75

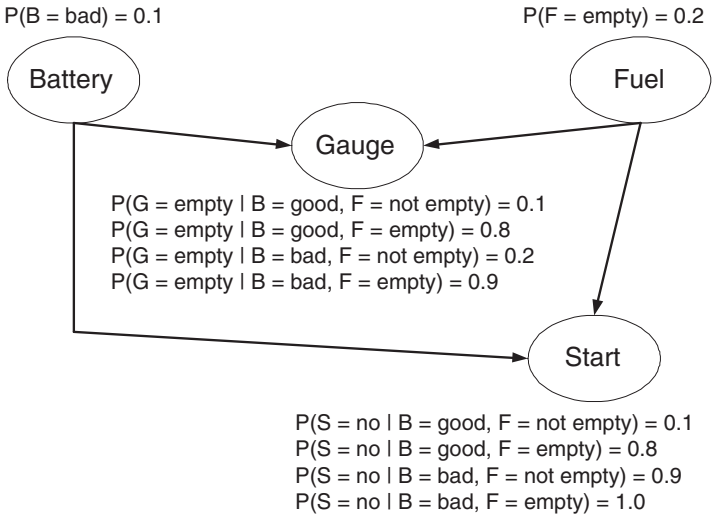


Figure 5.4. Bayesian belief network for Exercise 12.

P(Value=High|Engine=Good, Air Cond=Working) = 0.750
P(Value=High|Engine=Good, Air Cond=Broken) = 0.667
P(Value=High|Engine=Bad, Air Cond=Working) = 0.222
P(Value=High|Engine=Bad, Air Cond=Broken) = 0

(b) Use the Bayesian network to compute P(Engine = Bad, Air Conditioner = Broken).

$$\begin{aligned} & P(Engine = Bad, Air\ Cond = Broken) \\ = & \sum_{\alpha\beta} P(Engine = Bad, Air\ Cond = Broken, Mileage = \alpha, Value = \beta) \\ = & \sum_{\alpha\beta} P(Value = \beta | Engine = Bad, Air\ Cond = Broken) \\ & \times P(Engine = Bad | Mileage = \alpha) P(Mileage = \alpha) P(Air\ Cond = Broken) \\ = & \ 0.1453. \end{aligned}$$

12. Given the Bayesian network shown in Figure 5.4, compute the following probabilities:

(a) P(B = good, F = empty, G = empty, S = yes).

Answer:

$$\begin{aligned} & P(B = good, F = empty, G = empty, S = yes) \\ = \; & P(B = good) \times P(F = empty) \times P(G = empty | B = good, F = empty) \\ & \times P(S = yes | B = good, F = empty) \\ = \; & 0.9 \times 0.2 \times 0.8 \times 0.2 = 0.0288. \end{aligned}$$

(b) P(B = bad, F = empty, G = not empty, S = no).

Answer:

$$\begin{aligned} & P(B = bad, F = empty, G = not\ empty, S = no) \\ = \; & P(B = bad) \times P(F = empty) \times P(G = not\ empty | B = bad, F = empty) \\ & \times P(S = no | B = bad, F = empty) \\ = \; & 0.1 \times 0.2 \times 0.1 \times 1.0 = 0.002. \end{aligned}$$

(c) Given that the battery is bad, compute the probability that the car will start.

Answer:

$$\begin{aligned} & P(S = yes | B = bad) \\ = \; & \sum_{\alpha} P(S = yes | B = bad, F = \alpha) P(B = bad) P(F = \alpha) \\ = \; & 0.1 \times 0.1 \times 0.8 \\ = \; & 0.008 \end{aligned}$$

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	−	−	+	+	+	−	−	+	−	−

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

Answer:

1-nearest neighbor: +,

3-nearest neighbor: −,

5-nearest neighbor: +,

9-nearest neighbor: −.

(b) Repeat the previous analysis using the distance-weighted voting approach described in Section 5.2.1.

Answer:

1-nearest neighbor: +,

3-nearest neighbor: +,

5-nearest neighbor: +,

9-nearest neighbor: +.

14. The nearest-neighbor algorithm described in Section 5.2 can be extended to handle nominal attributes. A variant of the algorithm called PEBLS (Parallel Examplar-Based Learning System) by Cost and Salzberg [2] measures the distance between two values of a nominal attribute using the modified value difference metric (MVDM). Given a pair of nominal attribute values, V_1 and V_2, the distance between them is defined as follows:

$$d(V_1, V_2) = \sum_{i=1}^{k} \left| \frac{n_{i1}}{n_1} - \frac{n_{i2}}{n_2} \right|, \tag{5.2}$$

where n_{ij} is the number of examples from class i with attribute value V_j and n_j is the number of examples with attribute value V_j.

Consider the training set for the loan classification problem shown in Figure 5.9. Use the MVDM measure to compute the distance between every pair of attribute values for the `Home Owner` and `Marital Status` attributes.

Answer:

The training set shown in Figure 5.9 can be summarized for the `Home Owner` and `Marital Status` attributes as follows.

	Marital Status		
Class	Single	Married	Divorced
Yes	2	0	1
No	2	4	1

	Home Owner	
Class	Yes	No
Yes	0	3
No	3	4

d(Single, Married) = 1

d(Single, Divorced) = 0

d(Married, Divorced) = 1

d(Refund=Yes, Refund=No) = 6/7

15. For each of the Boolean functions given below, state whether the problem is linearly separable.

    (a) A AND B AND C
    Answer: Yes

    (b) NOT A AND B
    Answer: Yes

    (c) (A OR B) AND (A OR C)
    Answer: Yes

    (d) (A XOR B) AND (A OR B)
    Answer: No

16. (a) Demonstrate how the perceptron model can be used to represent the AND and OR functions between a pair of Boolean variables.
    Answer:
    Let x_1 and x_2 be a pair of Boolean variables and y be the output. For AND function, a possible perceptron model is:

$$y = sgn\Big[x_1 + x_2 - 1.5\Big].$$

    For OR function, a possible perceptron model is:

$$y = sgn\Big[x_1 + x_2 - 0.5\Big].$$

    (b) Comment on the disadvantage of using linear functions as activation functions for multilayer neural networks.
    Answer:
    Multilayer neural networks is useful for modeling nonlinear relationships between the input and output attributes. However, if linear functions are used as activation functions (instead of sigmoid or hyperbolic tangent function), the output is still a linear combination of its input attributes. Such a network is just as expressive as a perceptron.

17. You are asked to evaluate the performance of two classification models, M_1 and M_2. The test set you have chosen contains 26 binary attributes, labeled as A through Z.

    Table 5.5 shows the posterior probabilities obtained by applying the models to the test set. (Only the posterior probabilities for the positive class are shown). As this is a two-class problem, $P(-) = 1 - P(+)$ and $P(-|A, \ldots, Z) = 1 - P(+|A, \ldots, Z)$. Assume that we are mostly interested in detecting instances from the positive class.

Table 5.5. Posterior probabilities for Exercise 17.

Instance	True Class	$P(+\|A,\ldots,Z,M_1)$	$P(+\|A,\ldots,Z,M_2)$
1	+	0.73	0.61
2	+	0.69	0.03
3	−	0.44	0.68
4	−	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	−	0.08	0.38
8	−	0.15	0.05
9	+	0.45	0.01
10	−	0.35	0.04

(a) Plot the ROC curve for both M_1 and M_2. (You should plot them on the same graph.) Which model do you think is better? Explain your reasons.

Answer:

The ROC curve for $M1$ and $M2$ are shown in the Figure 5.5.

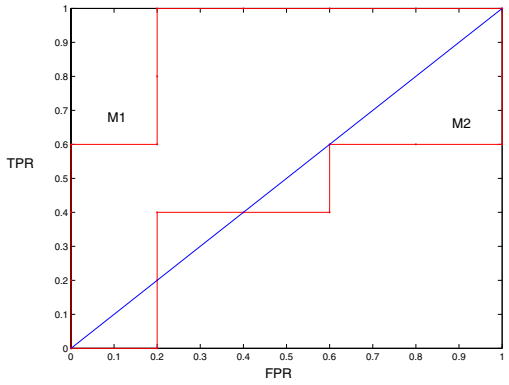


Figure 5.5. ROC curve.

$M1$ is better, since its area under the ROC curve is larger than the area under ROC curve for $M2$.

(b) For model M_1, suppose you choose the cutoff threshold to be $t = 0.5$. In other words, any test instances whose posterior probability is greater than t will be classified as a positive example. Compute the precision, recall, and F-measure for the model at this threshold value.

When $t = 0.5$, the confusion matrix for $M1$ is shown below.

		+	-
Actual	+	3	2
	-	1	4

Precision $= 3/4 = 75\%$.

Recall $= 3/5 = 60\%$.

F-measure $= (2 \times .75 \times .6)/(.75 + .6) = 0.667$.

(c) Repeat the analysis for part (c) using the same cutoff threshold on model M_2. Compare the F-measure results for both models. Which model is better? Are the results consistent with what you expect from the ROC curve?

Answer:

When $t = 0.5$, the confusion matrix for $M2$ is shown below.

		+	-
Actual	+	1	4
	-	1	4

Precision $= 1/2 = 50\%$.

Recall $= 1/5 = 20\%$.

F-measure $= (2 \times .5 \times .2)/(.5 + .2) = 0.2857$.

Based on F-measure, $M1$ is still better than $M2$. This result is consistent with the ROC plot.

(d) Repeat part (c) for model M_1 using the threshold $t = 0.1$. Which threshold do you prefer, $t = 0.5$ or $t = 0.1$? Are the results consistent with what you expect from the ROC curve?

Answer:

When $t = 0.1$, the confusion matrix for $M1$ is shown below.

		+	-
Actual	+	5	0
	-	4	1

Precision $= 5/9 = 55.6\%$.

Recall $= 5/5 = 100\%$.

F-measure $= (2 \times .556 \times 1)/(.556 + 1) = 0.715$.

According to F-measure, $t = 0.1$ is better than $t = 0.5$.

When $t = 0.1$, $FPR = 0.8$ and $TPR = 1$. On the other hand, when $t = 0.5$, $FPR = 0.2$ and $TRP = 0.6$. Since (0.2, 0.6) is closer to the point (0, 1), we favor $t = 0.5$. This result is inconsistent with the results using F-measure. We can also show this by computing the area under the ROC curve

For $t = 0.5$, area $= 0.6 \times (1 - 0.2) = 0.6 \times 0.8 = 0.48$.
For $t = 0.1$, area $= 1 \times (1 - 0.8) = 1 \times 0.2 = 0.2$.
Since the area for $t = 0.5$ is larger than the area for $t = 0.1$, we prefer $t = 0.5$.

18. Following is a data set that contains two attributes, X and Y, and two class labels, "+" and "−". Each attribute can take three different values: $0, 1$, or 2.

X	Y	Number of Instances	
		+	−
0	0	0	100
1	0	0	0
2	0	0	100
0	1	10	100
1	1	10	0
2	1	10	100
0	2	0	100
1	2	0	0
2	2	0	100

The concept for the "+" class is $Y = 1$ and the concept for the "−" class is $X = 0 \vee X = 2$.

(a) Build a decision tree on the data set. Does the tree capture the "+" and "−" concepts?

Answer:

There are 30 positive and 600 negative examples in the data. Therefore, at the root node, the error rate is

$$E_{orig} = 1 - \max(30/630, 600/630) = 30/630.$$

If we split on X, the gain in error rate is:

	$X=0$	$X=1$	$X=2$
+	10	10	10
−	300	0	300

$$E_{X=0} = 10/310$$
$$E_{X=1} = 0$$
$$E_{X=2} = 10/310$$

$$\Delta_X = E_{orig} - \frac{310}{630}\frac{10}{310} - \frac{10}{630}0 - \frac{310}{630}\frac{10}{310} = 10/630.$$

If we split on Y, the gain in error rate is:

	$Y=0$	$Y=1$	$Y=2$
+	0	30	0
−	200	200	200

$$\begin{aligned} E_{Y=0} &= 0 \\ E_{Y=1} &= 30/230 \\ E_{Y=2} &= 0 \end{aligned}$$

$$\Delta_Y = E_{orig} - \frac{230}{630}\frac{30}{230} = 0.$$

Therefore, X is chosen to be the first splitting attribute. Since the $X = 1$ child node is pure, it does not require further splitting. We may use attribute Y to split the impure nodes, $X = 0$ and $X = 2$, as follows:

- The $Y = 0$ and $Y = 2$ nodes contain $100-$ instances.
- The $Y = 1$ node contains $100-$ and $10+$ instances.

In all three cases for Y, the child nodes are labeled as $-$. The resulting concept is

$$\text{class} = \begin{cases} +, & \text{X = 1;} \\ -, & \text{otherwise.} \end{cases}$$

(b) What are the accuracy, precision, recall, and F_1-measure of the decision tree? (Note that precision, recall, and F_1-measure are defined with respect to the "+" class.)

Answer: The confusion matrix on the training data:

		Predicted +	Predicted −
Actual	+	10	20
	−	0	600

$$\begin{aligned} \text{accuracy} &: \frac{610}{630} = 0.9683 \\ \text{precision} &: \frac{10}{10} = 1.0 \\ \text{recall} &: \frac{10}{30} = 0.3333 \\ \text{F}-\text{measure} &: \frac{2 * 0.3333 * 1.0}{1.0 + 0.3333} = 0.5 \end{aligned}$$

(c) Build a new decision tree with the following cost function:

$$C(i,j) = \begin{cases} 0, & \text{if } i = j; \\ 1, & \text{if } i = +,\ j = -; \\ \frac{\text{Number of } - \text{ instances}}{\text{Number of } + \text{ instances}}, & \text{if } i = -,\ j = +. \end{cases}$$

(Hint: only the leaves of the old decision tree need to be changed.) Does the decision tree capture the "+" concept?

Answer:

The cost matrix can be summarized as follows:

		Predicted +	Predicted −
Actual	+	0	600/30=20
	−	1	0

The decision tree in part (a) has 7 leaf nodes, $X = 1$, $X = 0 \wedge Y = 0$, $X = 0 \wedge Y = 1$, $X = 0 \wedge Y = 2$, $X = 2 \wedge Y = 0$, $X = 2 \wedge Y = 1$, and $X = 2 \wedge Y = 2$. Only $X = 0 \wedge Y = 1$ and $X = 2 \wedge Y = 1$ are impure nodes. The cost of misclassifying these impure nodes as positive class is:

$$10 * 0 + 1 * 100 = 100$$

while the cost of misclassifying them as negative class is:

$$10 * 20 + 0 * 100 = 200.$$

These nodes are therefore labeled as +.

The resulting concept is

$$\text{class} = \begin{cases} +, & X = 1 \vee (X = 0 \wedge Y = 1) \vee (X = 2 \wedge Y = 2); \\ -, & \text{otherwise.} \end{cases}$$

(d) What are the accuracy, precision, recall, and F_1-measure of the new decision tree?

Answer:

The confusion matrix of the new tree

		Predicted +	Predicted −
Actual	+	30	0
	−	200	400

$$\text{accuracy} : \frac{430}{630} = 0.6825$$

$$\text{precision} : \frac{30}{230} = 0.1304$$

$$\text{recall} : \frac{30}{30} = 1.0$$

$$\text{F} - \text{measure} : \frac{2 * 0.1304 * 1.0}{1.0 + 0.1304} = 0.2307$$

19. (a) Consider the cost matrix for a two-class problem. Let $C(+,+) = C(-,-) = p$, $C(+,-) = C(-,+) = q$, and $q > p$. Show that minimizing the cost function is equivalent to maximizing the classifier's accuracy.

Answer:

Confusion Matrix	+	−
+	a	b
−	c	d

Cost Matrix	+	-
+	p	q
−	q	p

The total cost is $F = p(a+d) + q(b+c)$.
Since $acc = \frac{a+d}{N}$, where $N = a+b+c+d$, we may write

$$F = N\big[acc(p-q) + q\big].$$

Because $p - q$ is negative, minimizing the total cost is equivalent to maximizing accuracy.

(b) Show that a cost matrix is scale-invariant. For example, if the cost matrix is rescaled from $C(i,j) \longrightarrow \beta C(i,j)$, where β is the scaling factor, the decision threshold (Equation 5.82) will remain unchanged.

Answer:

The cost matrix is:

Cost Matrix	+	−
+	$c(+,+)$	$c(+,-)$
−	$c(-,+)$	$c(-,-)$

A node t is classified as positive if:

$$\begin{aligned}
& c(+,-)p(+|t) + c(-,-)p(-|t) > c(-,+)p(-|t) + c(+,+)p(+|t) \\
\Longrightarrow\quad & c(+,-)p(+|t) + c(-,-)[1-p(+|t)] > c(-,+)[1-p(+|t)] + c(+,+)p(+|t) \\
\Longrightarrow\quad & p(+|t) > \frac{c(-,+) - c(-,-)}{[c(-,+) - c(-,-)] + [c(+,-) - c(+,+)]}
\end{aligned}$$

The transformed cost matrix is:

Cost Matrix	+	−
+	$\beta c(+,+)$	$\beta c(+,-)$
−	$\beta c(-,+)$	$\beta c(-,-)$

Therefore, the decision rule is:

$$\begin{aligned}
p(+|t) &> \frac{\beta c(-,+) - \beta c(-,-)}{[\beta c(-,+) - \beta c(-,-)] + [\beta c(+,-) - \beta c(+,+)]} \\
&= \frac{c(-,+) - c(-,-)}{[c(-,+) - c(-,-)] + [c(+,-) - c(+,+)]}
\end{aligned}$$

which is the same as the original decision rule.

(c) Show that a cost matrix is translation-invariant. In other words, adding a constant factor to all entries in the cost matrix will not affect the decision threshold (Equation 5.82).

Answer:

The transformed cost matrix is:

Cost Matrix	+	−
+	$c(+,+)+\beta$	$c(+,-)+\beta$
−	$c(-,+)+\beta$	$c(-,-)+\beta$

Therefore, the decision rule is:

$$\begin{aligned} p(+|t) &> \frac{\beta + c(-,+) - \beta - c(-,-)}{[\beta + c(-,+) - \beta - c(-,-)] + [\beta + c(+,-) - \beta - c(+,+)]} \\ &= \frac{c(-,+) - c(-,-)}{[c(-,+) - c(-,-)] + [c(+,-) - c(+,+)]} \end{aligned}$$

which is the same as the original decision rule.

20. Consider the task of building a classifier from random data, where the attribute values are generated randomly irrespective of the class labels. Assume the data set contains records from two classes, "+" and "−." Half of the data set is used for training while the remaining half is used for testing.

    (a) Suppose there are an equal number of positive and negative records in the data and the decision tree classifier predicts every test record to be positive. What is the expected error rate of the classifier on the test data?

    Answer: 50%.

    (b) Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 0.8 and negative class with probability 0.2.

    Answer: 50%.

    (c) Suppose two-thirds of the data belong to the positive class and the remaining one-third belong to the negative class. What is the expected error of a classifier that predicts every test record to be positive?

    Answer: 33%.

    (d) Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 2/3 and negative class with probability 1/3.

    Answer: 44.4%.

21. Derive the dual Lagrangian for the linear SVM with nonseparable data where the objective function is

$$f(\mathbf{w}) = \frac{\|\mathbf{w}\|^2}{2} + C\big(\sum_{i=1}^{N} \xi_i\big)^2.$$

Answer:

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j} - C\bigg(\sum_i \xi_i\bigg)^2.$$

Notice that the dual Lagrangian depends on the slack variables ξ_i's.

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.

Answer:

The decision boundary is $f(x_1, x_2) = x_1x_2$.

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

Answer:

(a) Both decision tree and NB will do well on this data set because the distinguishing attributes have better discriminating power than noise attributes in terms of entropy gain and conditional probability. k-NN will not do as well due to relatively large number of noise attributes.

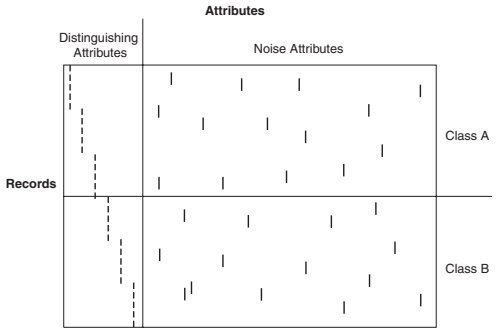
(b) NB will not work at all with this data set due to attribute dependency. Other schemes will do better than NB.

(c) NB will do very well in this data set, because each discriminating attribute has higher conditional probability in one class over the other and the overall classification is done by multiplying these individual conditional probabilities. Decision tree will not do as well, due to the relatively large number of distinguishing attributes. It will have an overfitting problem. k-NN will do reasonably well.

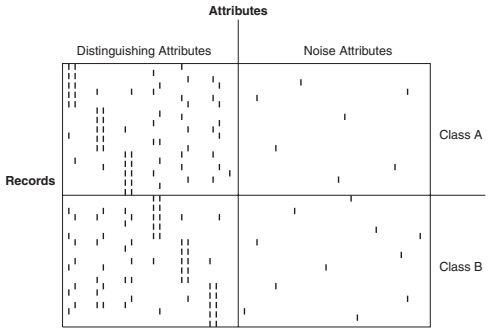
(d) k-NN will do well on this data set. Decision trees will also work, but will result in a fairly large decision tree. The first few splits will be quite random, because it may not find a good initial split at the beginning. NB will not perform quite as well due to the attribute dependency.

(e) k-NN will do well on this data set. Decision trees will also work, but will result in a large decision tree. If decision tree uses an oblique split instead of just vertical and horizontal splits, then the resulting decision tree will be more compact and highly accurate. NB will not perform quite as well due to attribute dependency.

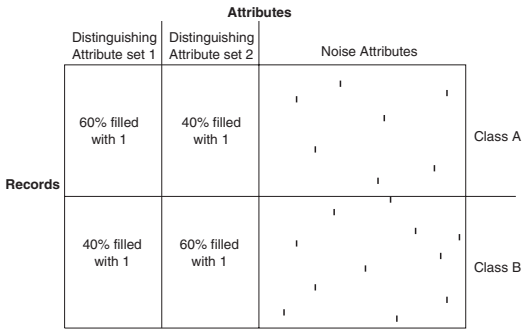
(f) kNN works the best. NB does not work well for this data set due to attribute dependency. Decision tree will have a large tree in order to capture the circular decision boundaries.



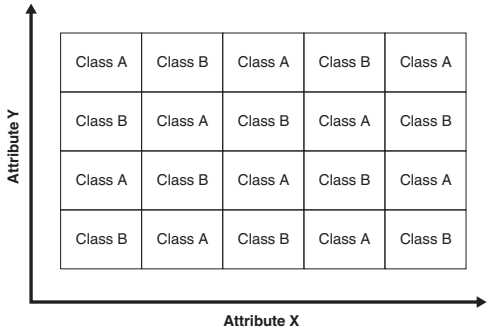
(a) Synthetic data set 1.



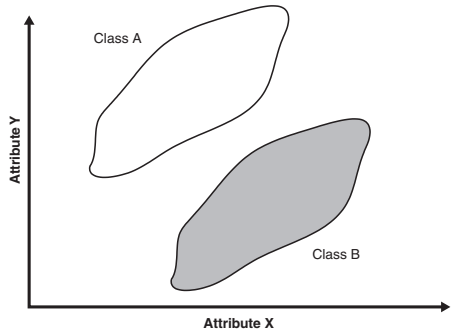
(b) Synthetic data set 2.



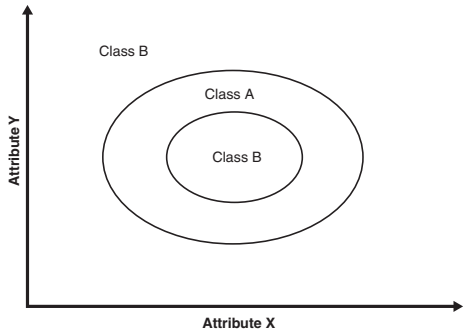
(c) Synthetic data set 3.



(d) Synthetic data set 4



(e) Synthetic data set 5.



(f) Synthetic data set 6.

Figure 5.6. Data set for Exercise 23.