Homework 3

Jarod Klion

February 23rd, 2022

Chapter 5:

1. **Consider a binary classification problem with the following set of attributes and attribute values:**
    a. **Are the rules mutually exclusive?**
        i. No
    b. **Is the rule set exhaustive?**
        i. No
    c. **Is ordering needed for this set of rules?**
        i. Yes
    d. **Do you need a default class for the rule set?**
        i. Yes
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:**

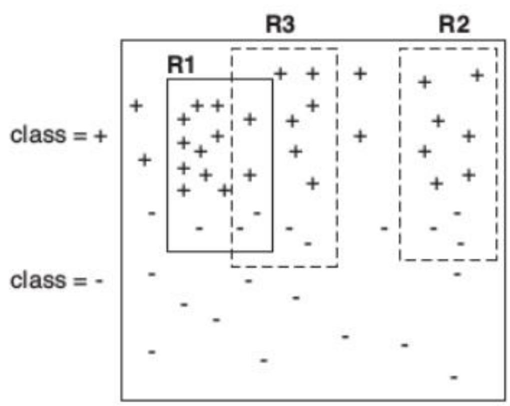

Figure 5.1. Elimination of training records by the sequential covering algorithm. $R1$, $R2$, and $R3$ represent regions covered by three different rules.

R1: 12 pos, 3 neg
R2: 7 pos, 3 neg
R3: 8 pos, 4 neg
Total: 29 positive examples, 21 negative examples

a. **The likelihood ratio statistic:**
    i. R1: $2\ x\ \left[12log_2(\frac{12}{15*(29/50)})\ +\ 3log_2(\frac{3}{15*(21/50)})\right]\ =\ 4.71$
    ii. R2: $2\ x\ \left[7log_2(\frac{7}{10*(29/50)})\ +\ 3log_2(\frac{3}{10*(21/50)})\right]\ =\ 0.89$
    iii. R3: $2\ x\ \left[8(\frac{8}{12*(29/50)})\ +\ 4log_2(\frac{4}{12*(21/50)})\right]\ =\ 0.54$
    iv. R1 is the best rule, and R3 is the worst rule by the likelihood ratio statistic

b. **The Laplace measure:**
   i. R1: $\frac{f_+ + 1}{n+k} = \frac{12+1}{15+2} = 76.47\%$
   ii. R2: $\frac{f_+ + 1}{n+k} = \frac{7+1}{10+2} = 66.67\%$
   iii. R3: $\frac{f_+ + 1}{n+k} = \frac{8+1}{12+2} = 64.29\%$
   iv. R1 is the best rule, and R3 is the worst rule by the Laplace measure
c. **The m-estimate measure (with k = 2 and p_+ = 0.58):**
   i. R1: $\frac{f_+ + kp_+}{n+k} = \frac{12 + 2*0.58}{15+2} = 77.41\%$
   ii. R2: $\frac{f_+ + kp_+}{n+k} = \frac{7 + 2*0.58}{10+2} = 68.00\%$
   iii. R3: $\frac{f_+ + kp_+}{n+k} = \frac{8 + 2*0.58}{12+2} = 65.43\%$
   iv. R1 is the best rule, and R3 is the worst rule by the m-estimate measure
d. **The rule accuracy after R1 has been discovered, where none of the examples covered by R1 are discarded.**
   i. R2: $\frac{f_+}{n} = \frac{7}{10} = 70\%$
   ii. R3: $\frac{f_+}{n} = \frac{8}{12} = 66.67\%$
   iii. R2 is chosen because it has higher accuracy than R3
e. **The rule accuracy after R1 has been discovered, where only the positive examples covered by R1 are discarded.**
   i. R2: 70%
   ii. R3: 60%
   iii. R2 is preferred because it has higher accuracy than R3
f. **The rule accuracy after R1 has been discovered, where both positive and negative examples covered by R1 are discarded.**
   i. R2: 70%
   ii. R3: 75%
   iii. R3 is preferred because it has higher accuracy than R2

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?**
      i. Given probabilities:
         P(S|UG) = 0.15
         P(S|G) = 0.23
         P(G) = 0.2
         P(UG) = 0.8
      ii. $P(G|S) = \frac{P(S|G)P(G)}{P(S|UG)P(UG) + P(S|G)P(G)} = \frac{0.23 \; x \; 0.2}{0.15 \; x \; 0.8 + 0.23 \; x \; 0.2} = 0.277$
   b. **Given the information in part (a), is a randomly chosen college student more likely to be a graduate or undergraduate student?**

i. P(UG) > P(G), so more likely to be an undergraduate.

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

i. $P(UG|S) = 1 - P(G|S)$, so more likely to be an undergraduate still.

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.**

i. Given probabilities:
P(D|UG) = 0.1
P(D|G) = 0.3
Needed probabilities:
P(D) = $P(UG)P(D|UG) + P(G)P(D|G) = 0.8\ x\ 0.1 + 0.2\ x\ 0.3\ =\ 0.14$
P(S) = $P(UG)P(S|UG) + P(G)P(S|G) = 0.8x0.15 + 0.2x0.23 = 0.166$
Conditional independence assumption:
P(DS|UG) = $P(D|UG)\ x\ P(S|UG)\ =\ 0.1\ x\ 0.15\ =\ 0.015$
P(DS|G) = $P(D|G)\ x\ P(S|G) = 0.3\ x\ 0.23 = 0.069$
P(UG|DS) = $\frac{P(DS|UG)\ x\ P(UG)}{P(DS)} = \frac{0.015 * 0.8}{P(DS)} = \frac{0.012}{P(DS)}$
P(G|DS) = $\frac{P(DS|G)\ x\ P(G)}{P(DS)} = \frac{0.069 * 0.2}{P(DS)} = \frac{0.0138}{P(DS)}$

ii. P(G|DS) > P(UG|DS) so 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|−).**

   i. $P(A = 0|+) = 0.4$ $P(A = 1|+) = 0.6$
   ii. $P(B = 0|+) = 0.8$ $P(B = 1|+) = 0.2$
   iii. $P(C = 0|+) = 0.2$ $P(C = 1|+) = 0.8$
   iv. $P(A = 0|-) = 0.6$ $P(A = 1|-) = 0.4$
   v. $P(B = 0|-) = 0.6$ $P(B = 1|-) = 0.4$
   vi. $P(C = 0|-) = 0.0$ $P(C = 1|-) = 1.0$

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.**

   i. $P(+|A = 0, B = 1, C = 0) = \frac{P(A=0,B=1,C=0|+)\ P(+)}{P(A=0,B=1,C=0)} = \frac{P(A=0|+)P(B=1|+)P(C=0|+)P(+)}{P(A=0,B=1,C=0)} = \frac{0.4\ x\ 0.2\ x\ 0.2\ x\ 0.5}{P(A=0,B=1,C=0)} = \frac{0.008}{P(A=0,B=1,C=0)}$
   ii. $P(-|A = 0, B = 1, C = 0) = \frac{P(A=0,B=1,C=0|-)\ P(-)}{P(A=0,B=1,C=0)} = \frac{P(A=0|-)P(B=1|-)P(C=0|-)P(-)}{P(A=0,B=1,C=0)} = \frac{0.6\ x\ 0.4\ x\ 0.0\ x\ 0.5}{P(A=0,B=1,C=0)} = \frac{0}{P(A=0,B=1,C=0)}$
   iii. The class label should be '+'

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

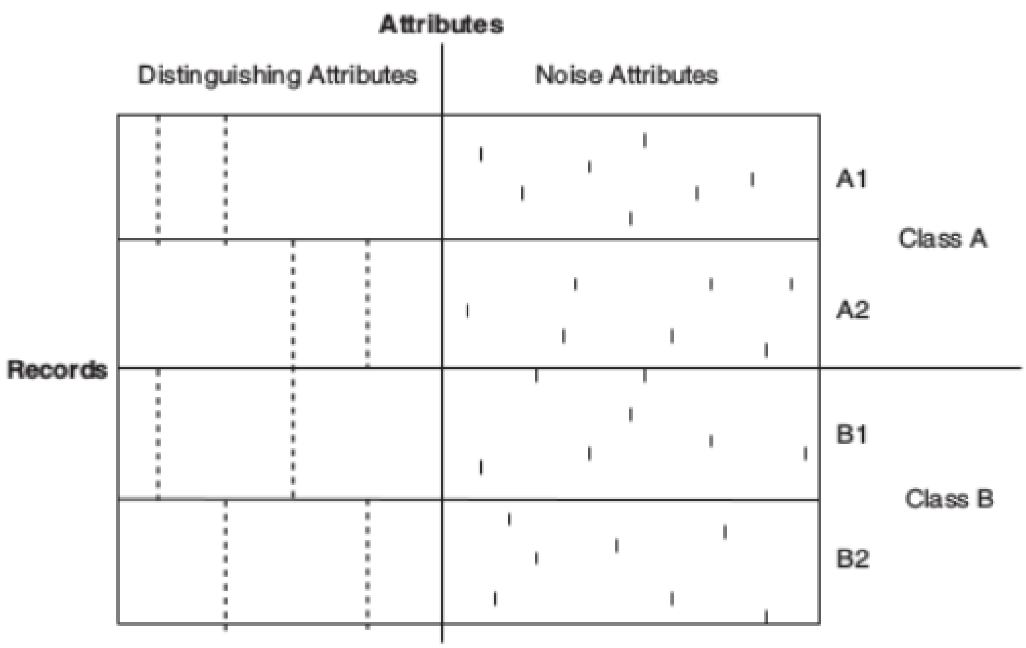

Figure 5.2. Data set for Exercise 9.

a. **Explain how naïve Bayes performs on the data set shown in Figure 5.2.**

   i. The conditional probabilities for each attribute are the same for both class A and class B, so naïve Bayes will perform poorly on this data set.

b. **If each class if further divided such that there are four classes (A1, A2, B1, and B2), will naïve Bayes perform better?**
    i. Yes, naïve Bayes will perform better

c. **How will a decision tree perform on this data set (for the two-class problem)? What if there are four classes?**
    i. A decision tree will perform poorly for the two-class problem as there will be no improvement in entropy after splitting; however, four classes will greatly improve the decision tree's performance.

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(Crocodile) = 2 x P(Alligator).**
    i. $2P(X = \hat{x}|Crocodile) \ = \ P(X = \hat{x}|Alligator) \rightarrow \hat{x} = 12.576$

b. **The prior probabilities are P(Alligator) = 2 x P(Crocodile).**
    i. $P(X = \hat{x}|Crocodile) \ = 2P(X = \hat{x}|Alligator) \rightarrow \hat{x} = 14.424$

c. **The prior probabilities are the same, but their standard deviations are different; i.e., σ(Crocodile) = 4 and σ(Alligator) = 2.**
    i. $\hat{x} \ = \ 7.625 \ or \ \hat{x} = 14.375$
    ii. For $x \leq 7.625$, animals would be classified as crocodiles.
    For $7.625 < x < 14.375$, animals would be classified as alligators.
    For $x \geq 14.375$, animals would be classified as crocodiles.

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).**
    i. 1-nearest neighbor: +
    ii. 3-nearest neighbor: -
    iii. 5-nearest neighbor: +
    iv. 9-nearest neighbor: -

b. **Repeat the previous analysis using the distance-weighted voting approach described in Section 5.2.1.**
    i. 1-nearest neighbor: +
    ii. 3-nearest neighbor: +
    iii. 5-nearest neighbor: +
    iv. 9-nearest neighbor: +

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.**

      i. AND: $y = sgn[x_1 + x_2 - 1.5]$

      ii. OR: $y = sgn[x_1 + x_2 - 0.5]$

   b. **Comment on the disadvantage of using linear functions as activation functions for the multilayer neural networks.**

      i. The disadvantage of using linear functions is that not all functions can be represented as a linear function, so the network will be less expressive.