Business Data Mining (IDS 572)

Homework 5 (Question 1)-Solution

Question 1

(a) P(senior)= 5/11 and P(junior) = 6/11.
(b) P(department|status)

Class	Sales	Systems	Marketing	Secretary
Senior	1/5	2/5	1/5	1/5
Junior	2/6 = 1/3	2/6 = 1/3	1/6	1/6

P(age|status)

Class	21…25	26…30	31…35	36…40	41…45	46…50
Senior	0/5 = 0	0/5 = 0	1/5	2/5	1/5	1/5
Junior	1/6	3/6=1/2	2/6 = 1/3	0/6 = 0	0/6 = 0	0/6 = 0

P(salary|status)

Class	26K-30K	31K-35K	36K-40K	41K-45K	46K-50K	66K-70K
Senior	0/5 = 0	0/5 = 0	1/5	0/5 = 0	2/5	2/5
Junior	2/6 = 1/3	1/6	0/6 = 0	1/6	1/6	1/6

(c) For the test instance A = {marketing, 31…35, 46K-50K} we have,
P(senior|A) = P(senior)P(marketing|senior)P(31…35|senior)P(46K-50K|senior)
$= 5/11 \times 1/5 \times 1/5 \times 2/5 = 0.007$
P(junior|A) = P(junior)P(marketing|junior)P(31…35|junior)P(46K-50K|junior)
$= 6/11 \times 1/6 \times 1/3 \times 1/6 = 0.005$

Therefore the label for this instance is "senior" since P(senior|A) > P(junior|A)

For the instance B = {sale, 31…35,66K-70K},

P(senior|B) = P(senior)P(sale|senior)P(31…35|senior)P(66K-70K|senior)
$= 5/11 \times 1/5 \times 1/5 \times 2/5 = 0.007$
P(junior|B) = P(junior)P(sale|junior)P(31…35|junior)P(66K-70K|junior)
$= 6/11 \times 1/3 \times 1/3 \times 1/6 = 0.01$

Therefore the label for this instance is “junior” since P(senior|B) < P(junior|B).

(d) Now suppose A = {marketing, 31…35, 46K-50K, 46K-50K} and B = {sale, 31…35, 66K-70K, 66K-70K}. Therefore we have,
P(senior|A) = P(senior)P(marketing|senior)P(31…35|senior)P(46K-50K|senior) P(46K-50K|senior) = 5/11 × 1/5 × 1/5 × 2/5 × 2/5 = 0.003
P(junior|A) = P(junior)P(marketing|junior)P(31…35|junior)P(46K-50K|junior) P(46K-50K|junior) = 6/11 × 1/6 × 1/3 × 1/6 × 1/6 = 0.0008

Therefore the label for the instance A is “senior” since P(senior|A) > P(junior|A)

P(senior|B) = P(senior)P(sale|senior)P(31…35|senior)P(66K-70K|senior) P(66K-70K|senior) = 5/11 × 1/5 × 1/5 × 2/5 × 2/5 = 0.003

P(junior|B) = P(junior)P(sale|junior)P(31…35|junior)P(66K-70K|junior) P(66K-70K|junior) = 6/11 × 1/3 × 1/3 × 1/6 × 1/6 = 0.02

Therefore the label for the instance B is “junior” since P(senior|B) < P(junior|B).

(e) The assumption of independent inputs is clearly violated in part (d). But Naïve Bayes model does not consider the dependency of the input variables. In addition, the occurrence of A and B is zero. In general the Naïve Bayes model could compute large probabilities even for the cases that have very low occurrence.