Math 656 HW5

Jeff Gould

9/29/2020

Chapter 4, Problem 7

```
table4.9 <- data.frame(Instance = c(1:10),
                       A = c(rep(0,5), rep(1,5)),
                       B = c(0,0,1,1,0,0,0,0,1,0),
                       C = c(0,rep(1,9)),
                       Class = c("\\+", "--", "--", "--", "\\+", "\\+", "--", "--", "\\+", "\\+"))
table4.9$Class = as.character(table4.9$Class)

kable(table4.9, escape = F, align = "r") %>%
  kable_classic(full_width = F)
```

Instance	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 probabilites for P(A|+), P(B|+), P(C|+), P(A|-), P(B|-), P(C|-)

```
table4.9 %>%
  group_by(Class) %>%
  summarise(P_A = mean(A == 1),
            P_B = mean(B == 1),
            P_C = mean(C == 1)) %>%
  kable(escape = F, align = "r") %>%
  kable_classic(full_width = F)
```

Class	P_A	P_B	P_C
–	0.4	0.4	1.0
+	0.6	0.2	0.8

Notation for the assignemt will be in this form: $P(A = 1)$ will be written as $P(A)$, $P(A = 0)$ will be written as $P(A')$

$P(A|+) = 0.6, P(A|-) = 0.4, P(A'|+) = 0.4, P(A'|-) = 0.6$

$P(B|+) = 0.2, P(B|-) = 0.4, P(B'|+) = 0.8, PB'|-) = 0.6$

$P(C|+) = 0.8, P(C|-) = 1, P(C'|+) = 0.2, P(C'|-) = 0$

b) Use the estimate of conditional probabilites given in the previous question to predict the class label for a test sample $(A = 0, B = 1, C = 0)$

$P(+|A', B, C') \approx P(A'|+)P(B|+)P(C'|+)P(+) = (0.4)(0.2)(0.2)(0.5) = 0.008$

$P(-|A', B, C') \approx P(A'|-)P(B|-)P(C'|-)P(-) = (0.6)(0.4)(0.0)(0.5) = 0$

$P(+|A', B, C') \approx \frac{0.008}{0.008+0} = 1$

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

```
p = 1/2
m = 4


table4.9 %>%
  group_by(Class) %>%
  summarise(P_A = (sum(A == 1) + m*p) / (n() + m),
            P_B = (sum(B == 1) + m*p) / (n() + m),
            P_C = (sum(C == 1) + m*p) / (n() + m)) %>%
  kable(escape = F, align = "r", digits = 3) %>%
  kable_classic(full_width = F)
```

Class	P_A	P_B	P_C
–	0.444	0.444	0.778
+	0.556	0.333	0.667

Using the m-estimate approach, our conditional probabilties are:

$P(A|+) = \frac{5}{9}, P(A|-) = \frac{4}{9}$

$P(B|+) = \frac{1}{3}, P(B|-) = \frac{4}{9}$

$P(C|+) = \frac{2}{3}, P(C|-) = \frac{7}{9}$

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

$P(+|A', B, C') \approx P(A'|+)P(B|+)P(C'|+)P(+) = (4/9)(1/3)(1/3)(1/2) = \frac{2}{81}$

$P(-|A', B, C') \approx P(A'|-)P(B|-)P(C'|-)P(-) = (5/9)(4/9)(2/9)(1/2) = \frac{20}{729}$

$P(+|A', B, C') \approx \frac{2/81}{2/81+20/729} \approx 0.4737$

So if we were given A', B, and C', and using a threshold of 0.5, then we would predict an instance of "-", which differs from step (b)

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

In step (b) we classified the instance of having class "+" with probability 1. In step (c) we classified it has having class "-" with probability 0.526. This is due to $P(C'|-) = 0$, so any Naive Bayes classifier on this example without some sort of smoothing, where one of the conditions is C', will always classify as "+" with probability 1. Because of this, a smoothing method, like the m-estimate method used here, is generally a better method.