Bayesian Statistics HW-1

Weijia Zhao

January 30, 2022

Caution: I keep 6 decimal digits for infinite decimals.

Exercise 1

Carpal Tunnel Syndrome Tests
Notice that the question expression is a bit vague. Here I believe that for Tinel, Phalen, NCV tests, the sensitivity is (0.97,0.92,0.93) respectively and the specificity is (0.92,0.88,0.87) respectively.

By definition, sensitivity is the probability that a test is positive conditional on the presence of the syndrome, specificity is the probability that a test is negative conditional on the absence of the syndrome, positive predictive value (PVV) is the probability that the case has syndrome conditional on a positive testing result

$$\text{sensitivity} = \frac{TP}{TP + FN} = \frac{TP}{nD}$$

$$\text{specificity} = \frac{TN}{FP + TN} = \frac{TN}{nC}$$

$$\text{PPV} = \frac{TP}{TP + FP} = \frac{TP}{nP}$$

(a) In serial manner, we only declare positive if all the three tests give positive results. For a case with syndrome (i.e. True), the probability that all tests are positive is given by

$$\mathbb{P}_{\text{Serial}}(P|D) = \mathbb{P}_{\text{Tinel}}(P|D) \cdot \mathbb{P}_{\text{Phalen}}(P|D) \cdot \mathbb{P}_{\text{NCV}}(P|D) = 0.97 \cdot 0.92 \cdot 0.93 = 0.829932$$

For a case without syndrome (i.e. False), the probability that at least one test is negative is given by

$$\begin{aligned}\mathbb{P}_{\text{Serial}}(N|C) &= 1 - \mathbb{P}_{\text{Serial}}(P|C) \\ &1 - \mathbb{P}_{\text{Tinel}}(P|C) \cdot \mathbb{P}_{\text{Phalen}}(P|C) \cdot \mathbb{P}_{\text{NCV}}(P|C) \\ &1 - (1 - \mathbb{P}_{\text{Tinel}}(N|C)) \cdot (1 - \mathbb{P}_{\text{Phalen}}(N|C)) \cdot (1 - \mathbb{P}_{\text{NCV}}(N|C)) \\ &= 1 - (1 - 0.91) \cdot (1 - 0.88) \cdot (1 - 0.87) = 0.998596\end{aligned}$$

Thus sensitivity and specificity are 0.829993 and 0.998596 respectively

(b) In parallel manner, we declare positive if at least one of the three tests gives positive result. For a case with syndrome (i.e. True), the probability that at least one test is positive is given by

$$\begin{aligned}\mathbb{P}_{\text{Parallel}}(P|D) &= 1 - \mathbb{P}_{\text{Parallel}}(N|D) \\ &= 1 - \mathbb{P}_{\text{Tinel}}(N|D) \cdot \mathbb{P}_{\text{Phalen}}(N|D) \cdot \mathbb{P}_{\text{NCV}}(N|D) \\ &= 1 - (1 - \mathbb{P}_{\text{Tinel}}(P|D)) \cdot (1 - \mathbb{P}_{\text{Phalen}}(P|D)) \cdot (1 - \mathbb{P}_{\text{NCV}}(P|D)) \\ &= 1 - (1 - 0.97) \cdot (1 - 0.92) \cdot (1 - 0.93) = 0.999832\end{aligned}$$

For a case without syndrome (i.e. False), the probability that all tests are negative is given by

$$\mathbb{P}_{\text{Parallel}}(N|C) = \mathbb{P}_{\text{Tinel}}(N|C) \cdot \mathbb{P}_{\text{Phalen}}(N|C) \cdot \mathbb{P}_{\text{NCV}}(N|C) = 0.91 \cdot 0.88 \cdot 0.87 = 0.696696$$

Thus the sensitivity and specificity are 0.999832 and 0.696696 respectively

(c) We know $\mathbb{P}(D) = \frac{50}{1000} = 0.05$ and $\mathbb{P}(C) = 1 - \frac{50}{1000} = 0.95$

For serial test

$$\begin{aligned}PPV &= \frac{\mathbb{P}(D) \cdot \mathbb{P}_{\text{Serial}}(P|D)}{\mathbb{P}(D) \cdot \mathbb{P}_{\text{Serial}}(P|D) + \mathbb{P}(C) \cdot \mathbb{P}_{\text{Serial}}(P|C)} \\ &= \frac{0.05 \cdot 0.829932}{0.05 \cdot 0.829932 + 0.95 \cdot (1 - 0.998596)} = 0.968859\end{aligned}$$

For parallel test

$$\begin{aligned}PPV &= \frac{\mathbb{P}(D) \cdot \mathbb{P}_{\text{Parallel}}(P|D)}{\mathbb{P}(D) \cdot \mathbb{P}_{\text{Parallel}}(P|D) + \mathbb{P}(C) \cdot \mathbb{P}_{\text{Parallel}}(P|C)} \\ &= \frac{0.05 \cdot 0.999832}{0.05 \cdot 0.999832 + 0.95 \cdot (1 - 0.696696)} = 0.147847\end{aligned}$$

Thus the PPV for the test in (a) and (b) are 0.968859 and 0.147847 respectively.

Exercise 2

A Simple Naive Bayes Classifier: 6420 Students going to Beach
We first grab some summary statistics from the table:
Raw probabilities ("Midterm" for "Satisfied with Midterm Results", "Finances" for "Personal finances good", "Forecast" for "Weather forecast good", "Gender" for "Female"):

$$\mathbb{P}(\text{Beach=1}) = \frac{40}{100} = 0.4$$
$$\mathbb{P}(\text{Midterm=1}) = \frac{62}{100} = 0.62$$
$$\mathbb{P}(\text{Finances=1}) = \frac{54}{100} = 0.54$$
$$\mathbb{P}(\text{Friends=1}) = \frac{38}{100} = 0.38$$
$$\mathbb{P}(\text{Forecast=1}) = \frac{76}{100} = 0.76$$
$$\mathbb{P}(\text{Gender=1}) = \frac{32}{100} = 0.32$$

Conditional Probabilities

$$\mathbb{P}(\text{Midterm=1}|\text{Beach=1}) = \frac{35}{40}, \mathbb{P}(\text{Midterm=1}|\text{Beach=0}) = \frac{27}{60}$$
$$\mathbb{P}(\text{Finances=1}|\text{Beach=1}) = \frac{29}{40}, \mathbb{P}(\text{Finances=1}|\text{Beach=0}) = \frac{25}{60}$$
$$\mathbb{P}(\text{Friends=1}|\text{Beach=1}) = \frac{31}{40}, \mathbb{P}(\text{Friends=1}|\text{Beach=0}) = \frac{7}{60}$$
$$\mathbb{P}(\text{Forecast=1}|\text{Beach=1}) = \frac{33}{40}, \mathbb{P}(\text{Forecast=1}|\text{Beach=0}) = \frac{43}{60}$$
$$\mathbb{P}(\text{Gender=1}|\text{Beach=1}) = \frac{9}{40}, \mathbb{P}(\text{Gender=1}|\text{Beach=0}) = \frac{23}{60}$$

(a) For Michael,

$$\mathbb{P}(\text{Beach}) = \frac{(1-\frac{35}{40})\cdot(1-\frac{29}{40})\cdot(\frac{31}{40})\cdot(\frac{33}{40})\cdot(1-\frac{9}{40})\cdot 0.4}{(1-\frac{35}{40})\cdot(1-\frac{29}{40}))\cdot(\frac{31}{40})\cdot(\frac{33}{40})\cdot(1-\frac{9}{40})\cdot 0.4 + (1-\frac{27}{60})\cdot(1-\frac{25}{60})\cdot(\frac{7}{60})\cdot(\frac{43}{60})\cdot(1-\frac{23}{60})\cdot 0.6}$$
$$= 0.407042$$

$$\mathbb{P}(\text{NoBeach}) = 1 - \mathbb{P}(\text{Beach}) = 0.592958$$

So naive Bayes assigns the probability of 0.592958 of class "not going to beach" to Michael.

Note that I stack all the calculations in one single equation to avoid rounding errors. But similar to the example given, pbpropto = $(1-\frac{35}{40})\cdot(1-\frac{29}{40})\cdot(\frac{31}{40})\cdot(\frac{33}{40})\cdot(1-\frac{9}{40})\cdot 0.4$, pnbpropto = $(1-\frac{27}{60})\cdot(1-\frac{25}{60})\cdot(\frac{7}{60})\cdot(\frac{43}{60})\cdot(1-\frac{23}{60})\cdot 0.6$, and pbeach = $\frac{\text{pbpropto}}{\text{pbpropto+pnbpropto}}$

(b) For Melissa

$$\mathbb{P}(\text{Beach}) = \frac{(\frac{35}{40})\cdot(\frac{29}{40})\cdot(1-\frac{31}{40})\cdot(\frac{33}{40})\cdot(\frac{9}{40})\cdot 0.4}{(\frac{35}{40})\cdot(\frac{29}{40})\cdot(1-\frac{31}{40})\cdot(\frac{33}{40})\cdot(\frac{9}{40})\cdot 0.4 + (\frac{27}{60})\cdot(\frac{25}{60})\cdot(1-\frac{7}{60})\cdot(\frac{43}{60})\cdot(\frac{23}{60})\cdot 0.6}$$
$$= 0.279642$$

$$\mathbb{P}(\text{NoBeach}) = 1 - \mathbb{P}(\text{Beach}) = 0.720358$$

So naive Bayes assigns the probability of 0.592958 of class “not going to beach" to Melissa. Again, we can instead break up the long equation into pieces similar to (a).

Exercise 3

Multiple Choice Exam

(a) Note that when the student knows the correct answer, the probability that (s)he answers it correctly is 1. The probability that one question is answered correctly is

$$P1 = \mathbb{P}(\text{Known}) + \mathbb{P}(\text{Unknown}) \cdot \mathbb{P}(\text{Correct}|\text{Unknown}) = 0.8 + 0.2 \cdot 0.25 = 0.85$$

Since two questions are independent, the probability that both questions will be answered correctly is $\mathbb{P}(\text{Correct}) = P1 \cdot P1 = 0.7225$

(b) The probability that the student really knew the correct answer conditional on answered correctly is given by

$$\mathbb{P}(\text{Known}|\text{Correct}) = \frac{\mathbb{P}(\text{Known \& Correct})}{\mathbb{P}(\text{Correct})} = \frac{\mathbb{P}(\text{Known})}{\mathbb{P}(\text{Correct})} = \frac{0.8 \cdot 0.8}{0.7225} = 0.885813$$

(c) When there are n independent questions, the probability that the student answers all of them correctly is given by 0.85^n, the probability that the student really knew the answer to all questions is given by $\frac{0.8^n}{0.85^n} = (\frac{16}{17})^n$. Both quantities go to 0 if $n \longrightarrow \infty$