Zac Monroe

B365

HW3

1. Problem 1
   (a) T and S are not independent. P(T) = 0.20. P(S | T) = 0.80. P(S | !T) = P(S | not T) = 0.30. So since P(S | T) ≠ P(S | !T), the probability of S is dependent on T, so S and T are not independent events.
   (b) P(T | S) = P(S, T) / P(S) = (P(S | T) * P(T)) / (P(S | T) * P(T) + P(S | !T) * P(!T))
   = (0.80 * 0.20) / (0.80 * 0.20 + 0.30 * 0.80) = 0.40
   (c) See prob1.r.
   (d) From my simulation, my estimated 95% confidence interval for P(T | S) is 0.40505 +- 0.01557, or [0.3895, 0.4206]. This is consistent with my answer of 0.40 from part (b), because it is contained within this interval.
2. Problem 2 : F = person favors the ballot measure, !F = person does not favor the ballot measure
   (a) P(A | !F) = P(A) * P(!F | A) / (P(A) * P(!F | A) + P(B) * P(!F | B) + P(C) * P(!F | C))
   = (0.20 * 0.60) / (0.20 * 0.60 + 0.30 * 0.40 + 0.50 * 0.20) = ~0.353
   (b) P(B | !F) = P(B) * P(!F | B) / (P(A) * P(!F | A) + P(B) * P(!F | B) + P(C) * P(!F | C))
   = (0.30 * 0.40) / (0.20 * 0.60 + 0.30 * 0.40 + 0.50 * 0.20) = ~0.353
   (c) P(C | !F) = P(C) * P(!F | C) / (P(A) * P(!F | A) + P(B) * P(!F | B) + P(C) * P(!F | C))
   = (0.50 * 0.20) / (0.20 * 0.60 + 0.30 * 0.40 + 0.50 * 0.20) = ~0.294
   (d) See prob2.r. My estimate is 0.1187.
   (e) Yes, this probability applies to this person. The given probabilities of a person belonging to a certain political party, and of a person favoring the particular ballot measure are indicated to be true probabilities, and not mere estimates. So the probability that the person belongs to a particular party based on the fact that they do not favor the measure indeed is applicable.
3. Problem 3 : See prob3.r.
   (a) The department and admit status appear to not be independent. There are differing rates of admittance for each department.

(b) The department and gender of an applicant appear to not be independent. There are differing proportions of Male to Female for each department.

(c) Gender and admit status appear to be independent for department F. The proportion of applicants that were accepted are very close for each of the two genders. Given department F, gender and admit status appear to be conditionally independent.

(d) Run prob3.r to see the table. Here is a recreation of it:

Male	Female
1198	557

4. Problem 4

   (a) See/run prob4.r.

   (b) Petal.Length and Petal.Width seem to be the best two attributes to help distinguish between each of the three species of irises. Using these two features, there are the clearest clusters of species on their scatter plots. There is one clear cluster that does not "touch" the other two at all, and there is minimal overlap (minimal when compared to using other features for classification) between the remaining two species.

5. Problem 5

   (a) See prob5.r.

   (b) $P(x_1 = 2, x_2 = 4, x_3 = 5 \mid A) = 1/6 * 1/6 * 1/6 = 1/216 = \sim 0.00463 \rightarrow$

      $P(A \mid x_1 = 2, x_2 = 4, x_3 = 5) = P(A) * P(x_1 = 2, x_2 = 4, x_3 = 5 \mid A) / (P(A) * P(x_1 = 2, x_2 = 4, x_3 = 5 \mid A) + P(B) * P(x_1 = 2, x_2 = 4, x_3 = 5 \mid B) + P(C) * P(x_1 = 2, x_2 = 4, x_3 = 5 \mid C)) = \sim 0.5294$

   $P(x_1 = 2, x_2 = 4, x_3 = 5 \mid B) = 2/9 * 2/9 * 1/9 = 4/729 = \sim 0.00549 \rightarrow$

      $P(B \mid x_1 = 2, x_2 = 4, x_3 = 5) = P(B) * P(x_1 = 2, x_2 = 4, x_3 = 5 \mid C) / (P(A) * P(x_1 = 2, x_2 = 4, x_3 = 5 \mid A) + P(B) * P(x_1 = 2, x_2 = 4, x_3 = 5 \mid B) + P(C) * P(x_1 = 2, x_2 = 4, x_3 = 5 \mid C)) = \sim 0.3137$

   $P(x_1 = 2, x_2 = 4, x_3 = 5 \mid C) = 2/9 * 1/9 * 1/9 = 2/729 = \sim 0.00274 \rightarrow$

      $P(C \mid x_1 = 2, x_2 = 4, x_3 = 5) = P(C) * P(x_1 = 2, x_2 = 4, x_3 = 5 \mid C) / (P(A) * P(x_1 = 2, x_2 = 4, x_3 = 5 \mid A) + P(B) * P(x_1 = 2, x_2 = 4, x_3 = 5 \mid B) + P(C) * P(x_1 = 2, x_2 = 4, x_3 = 5 \mid C)) = \sim 0.1569$

(c) I would classify this outcome as having been produced by die A. This is because the die with the highest probability for this outcome to have happened is die A, with roughly .53 probability.

(d) My classifier is not giving stellar results: the proportion of correct guesses, or the proportion of times that the classifier gives the best result, is roughly 0.48. When I run the experiment 100000 times, I get a proportion closer to 0.486.