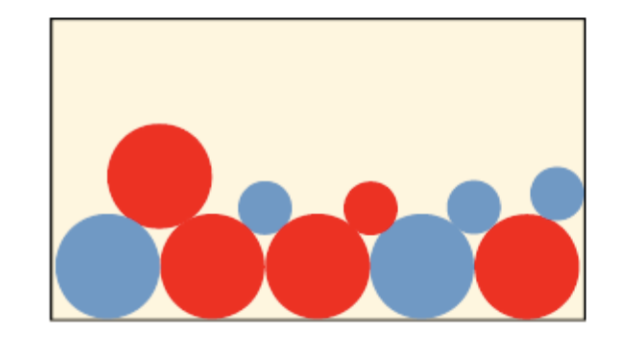
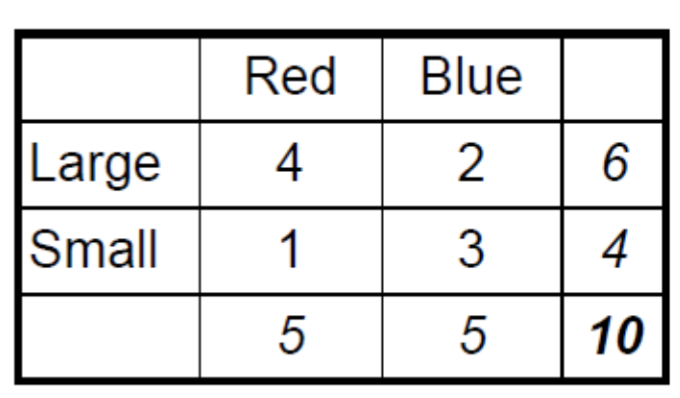
**UC Irvine BANA 273 Assignment 1**

**Q1. Probability**

Recall the example of drawing balls in class. (There are 10 balls in a basket, with 4 large and red, 1 small and red, 2 large and blue and 3 small and blue.) If we repeat the trial twice, in other words, we randomly pick one ball with replacement and then pick the second, and denote possible events as follows:

A1 = Red in the 1st trial, A2 = Red in the 2nd trial

B1 = Large in the 1st trial, B2 = Large in the 2nd trial

1. What are the probabilities of A1, B1, A2, B2?
2. What is the conditional probability of P(A2|A1)?
3. What events do A1C , B2C, A1∩A2, (A1C)∩B2 represent?
4. What are the probabilities of events in c)?

Now consider a different case in which the first ball was picked without replacement, and then the second ball was picked:

1. What is the conditional probability of P(A2|A1)?
2. What are the probabilities of events in c)?
3. Are the results the same as in the first case? Why or why not.

**A1. Probability**

1. The probability of A1 is 0.5

The probability of A2 is 0.5

The probability of B1 is 0.6

The probability of B2 is 0.6

1. The conditional probability of P(A2|A1) is 0.5
2. A1C represents the probability of getting Blue in the first trial.

B2C represents the probability of getting Small in the second trial.

A1∩A2represents the probability of getting Red in the first trial and Red in the second trial.

(A1C)∩B2 represents the probability of getting Blue in the first trial and Large in the second trial.

1. The probability of A1C is 0.5

The probability of B2C is 0.4

The probability of A1∩A2is 0.25

The probability of (A1C)∩B2is 0.5\*0.6 = 0.3

1. The conditional probability of P(A2|A1) is 4/9 = 0.44444
2. The probability of A1C is 0.5

The probability of B2C is 0.4

The probability of A1∩A2is 2/9 = 0.222222

The probability of (A1C)∩B2is 14/45 = 0.3111111

1. The results are the same only for the first 2 cases but not for the last 2 cases.

In the first case A1C, the probability depends on the first trial only. Replacement has not happened and therefore the probability remains the same.

In the second case B2C, during the first trial the probability of getting small is 4/10 and of getting big is 6/10. In the second trial, the probability of getting small again when the first trial is small is 3/9 and of getting small when the first trial is big is 4/9. Therefore the probability of getting small in the second trial when the first trial is small is 4/30 and the probability of getting small in the second trial when the first trial is big is 8/30. Therefore combining the two values we get 2/5 = 0.4

In the third case and fourth case replacement happens. Therefore the probabilities change.

In the third case A1∩A2, the probability of getting A1 is 5/10. Once a red ball is taken there are only 4 red balls left and there are only 9 balls overall. Therefore the probability of taking a red ball again is 4/9. Therefore the probability becomes 2/9 = 0.22222222

In the fourth case (A1C)∩B2, the probability of getting A1C is 5/10 with the probability of blue and large in the first trial being 2/5 and of blue and small in the first trial being 3/5. If a blue and large ball is taken away in the first trial there are only 5 large balls left and the probability becomes 5/9. Multiplying 5/9 with 2/5, we get 2/9 which is the probability of blue and large in the first trial and large in the second trial. If a blue and small ball is taken away in the first trial there are 6 large balls left and the probability becomes 6/9. Multiplying 6/9 with 3/5, we get 2/5 which is the probability of blue and small in the first trial and large in the second trial. So the total probability of (A1C)∩B2 becomes 2/9 + 2/5 which is equal to 14/45.

**Q2. Bayes Rule and Information Gain**

Suppose a medical test is designed to detect one disease with 30% prevalence (it is estimated that 30% of the population has the disease). Among those who have it, the probability of a positive testing result is 0.9. However, the probability of being erroneously tested positive among those who don’t have the disease is 0.2, causing that 41% of the overall population have positive testing results.

(Disease = yes/no, Test = positive/negative)

1. What is the probability of actually having the disease if someone is tested positive?
2. What’s the information gain from the testing information? (*I(Disease; Test)*)
3. What’s the gain ratio of the testing information? (*G(Test; Disease)*)

**A2. Bayes Rule and Information Gain**

1. The probability of actually having the disease if someone is tested positive is 0.65853659
2. The information gain *I(Disease; Test)* from the testing information is 0.322254
3. The gain ratio *G(Test; Disease)* of the testing information is 0.3657820

**Q3. Information, Information Gain Function, & Rank**

The following data is extracted from Weka’s weather.nominal data set. It reports past observations on weather conditions and the decision to Play or not. We are interested in whether the weather conditions can be used to predict the Play or not decision. The excel version of the dataset is available for download on Canvas🡪Assignment 1.

(Excel version: BANA273\_2020\_HW1Q3)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Outlook** | **Temperature** | **Humidity** | **Windy** | **Play** |
| sunny | hot | high | FALSE | no |
| sunny | hot | high | TRUE | no |
| overcast | hot | high | FALSE | yes |
| rainy | mild | high | FALSE | yes |
| rainy | cool | normal | FALSE | yes |
| rainy | cool | normal | TRUE | no |
| overcast | cool | normal | TRUE | yes |
| sunny | mild | high | FALSE | no |
| sunny | cool | normal | FALSE | yes |
| rainy | mild | normal | FALSE | yes |
| sunny | mild | normal | TRUE | yes |
| overcast | mild | high | TRUE | yes |
| overcast | hot | normal | FALSE | yes |
| rainy | mild | high | TRUE | no |

1. Based on the table, what is the entropy of “Play”?
2. Create contingency tables given each of the following attributes:
   1. Temperature
   2. Humidity
   3. Windy
3. Based on the contingency tables, calculate the information gain of each of the above 3 attributes (i.e., Temperature, Humidity, and Windy)
4. Please rank the 3 attributes from the most informative to the least informative for predicting Play

**A3. Information, Information Gain Function, & Rank**

1. The entropy of play is 0.94
2. 1) Temperature

|  |  |  |
| --- | --- | --- |
|  | **Play** |  |
|  |  |  |
| **Temperature** | **no** | **yes** |
| cool | 1 | 3 |
| hot | 2 | 2 |
| mild | 2 | 4 |
| **Grand Total** | **5** | **9** |

2) Humidity

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Play** |  |  |
|  | **no** |  | **yes** |
| **Humidity** |  |  |  |
| high | 4 |  | 3 |
| normal | 1 |  | 6 |
| **Grand Total** | **5** |  | **9** |

3) Windy

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Play** |  |  |
|  | **no** |  | **yes** |
| **Windy** |  |  |  |
| FALSE | 2 |  | 6 |
| TRUE | 3 |  | 3 |
| **Grand Total** | **5** |  | **9** |

1. 1) Temperature

Information gain = 0.029

2) Humidity

Information gain = 0.152

3) Windy

Information gain = 0.048

1. Humidity is the most informative. Windy is the second most informative. Temperature is the third most informative.

Humidity > Windy > Temperature