Bayesian Network for Project 1

Contents

Summary

Through the data collected by students playing Pokémon GO, we aim to draw relations between the various variables ranging from the players' level and the Pokémon CP they would most likely be able to obtain. Using a Bayesian network is a tool, we could effectively answer questions that many might ask. For example, in our second research question we have found that it is easier to catch Pokémon in the daytime.

Data Processing

Given a raw data set, we used Microsoft Excel to manually eliminated invalid entries by assessing and considering what was realistic and what was not possible (void data). Subsequently, the first letter of every word was edited to uppercase to ensure consistency. Following that, we used the "Convert text to columns wizard" to separate Pokemon types for those which had 2 types. Lastly, we rounded off the "waiting time" column to the nearest integer for better calculation. After which, we used the function `textscan` and `table` function to import the data into $MATLAB$.

```
disp(consolidatedtable(7:13,:));
```

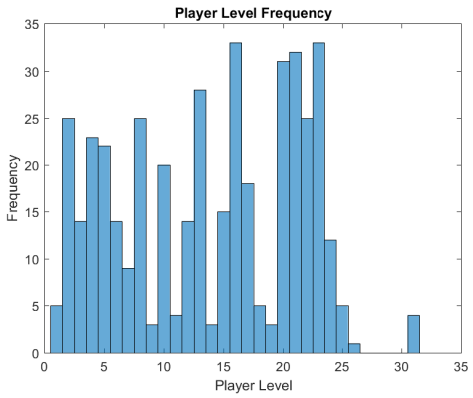
pokeType1	pokeType2	pokeCP	playerLvl	waitingTime	pokeLoc	tofDay	pokeBalls	razzBerry
'Normal'	''	210	15	1	'school'	'day'	1	'no'
'Normal'	''	135	16	3	'school'	'day'	3	'yes'
'Bug'	'Poison'	98	16	3	'school'	'day'	3	'no'
'Flying'	'Normal'	112	16	3	'school'	'day'	4	'no'
'Normal'	''	114	16	0	'school'	'day'	1	'no'
'Normal'	''	82	16	6	'school'	'day'	1	'no'
'Water'	''	171	20	5	'forest (around NTU)'	'night'	1	'no'

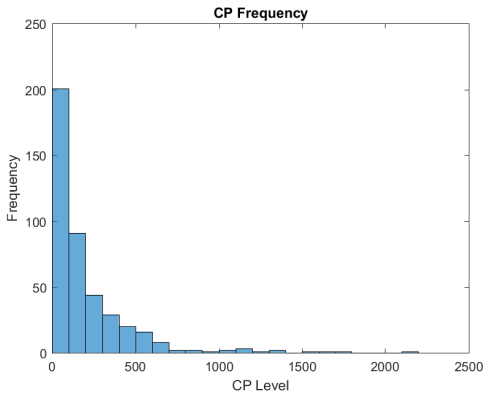
Data Exploration

Using the extracted data, we have found that the player levels were spread mainly between the 0 and 25 range. We have created a function `hist(playerLvl)` and `hist(CP)` to plot the frequency for player level and Pokémon CP in 2 separate histograms. The code for the function can be found within the appendix.

Player level and Pokémon are distributed according to their respective histograms:

```
hist(playerLvl);
hist(pokeCP);
```





We also assigned each random variable with a True/False indicator. For the following example, the indicator is true using the following assumptions:

- Razz: When Razz Berry was used
- Location: When the Pokémon was caught in water
- Time: The Pokémon was caught during daytime
- Level: Within the range 1∼10
- Waiting Time: Between 0∼5min
- Type: Pokémon type is grass
- CP: Pokémon CP is 100∼400
- Pokeball: A Pokémon was caught within the use of 1∼3 Pokéballs.

Here are some simple examples on computing probabilities.

1. Given that the player level is > 10, the probability that the CP for the Pokémon caught is < 100 or > 400 is

$$\begin{aligned} P(CP'|Level') &= 1 - P(CP|Level') \\ &= 1 - \frac{138}{266} \\ &= 0.481 \end{aligned}$$

   $\therefore$ The probability is calculated to be approximately half.

2. The joint probability that Razz Berry was used and player level is between 1 and 10 is

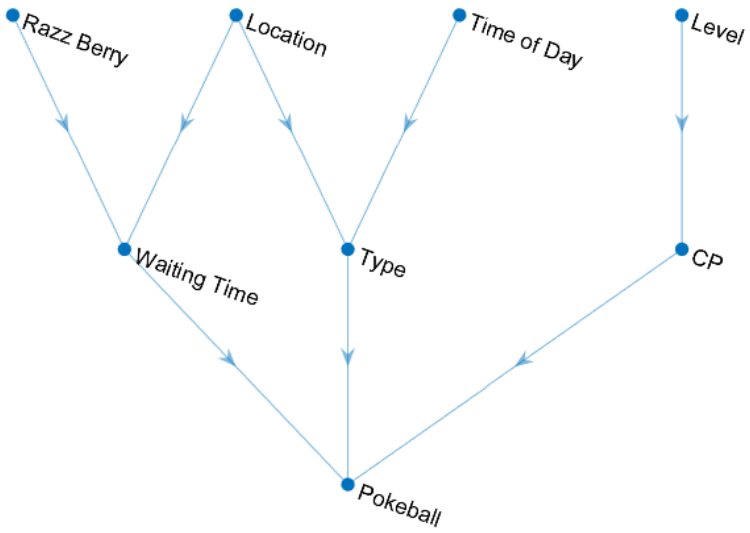
$$\begin{aligned} P(Razz, Level) &= P(Razz) * P(Level) \\ &= (1 - 0.838) * 0.376 \\ &= 0.0609 \end{aligned}$$

   $\therefore$ The probability is close to zero.

Network structure

We constructed a `graph` function using the `plot(digraph())` function in the latest version of *MATLAB* to construct the following Bayesian Network. The network was based on our understanding and drawing of perceived relations between each of the nodes. This results in directed arrows linking the various nodes together. The code for the function `graph()` can be found within the appendix.

```
graph1();
```



Probability Tables

For storing of the probability tables, we divide the table into 2 columns, "$True$" and "$False$". For the condition "$True$", we calculate the frequency by using a `for` loop to comb through the entire table and finding the data that satisfies the condition, then divide by the total amount of data (rows). The probability for "$False$" can be calculated by the formula $P(True) = 1 - P(False)$. The code to calculate the probability tables for *Player Level* and *Pokéballs* are attached in the appendix.

Using this method, we implemented multiple `for` loops to calculate the probability for each condition for every single node. As a result, we will have probability tables for each node that is featured in the directed graph.

To store the probability tables for each of the eight nodes (as per our Bayesian network), we save each table in a separate matrix and store these matrices into separate files (.mat) outside of $MATLAB$.

Research Questions

Q1.

Find the joint probability of using Razz berry and Waiting Time being less than or equal to 3 minutes.

We apply BNT(Bayesian Networks Toolbox) and found $P(Razz = true, WaitT = true)$ to be 0.0871. We may also compute the probability by hand:

$$\begin{aligned}
P(Razz, WaitT) &= P(Razz, WaitT|Razz, Loca) * P(Razz, Loca) + \\
&\quad P(Razz, WaitT|Razz, Loca') * P(Razz, Loca') + \\
&\quad P(Razz, WaitT|Razz', Loca) * P(Razz', Loca) + \\
&\quad P(Razz, WaitT|Razz', Loca') * P(Razz', Loca') \\
&\quad = P(WaitT|Razz, Loca) * P(Razz) * P(Loca) + \\
&\quad P(WaitT|Razz, Loca') * P(Razz) * P(Loca') \\
&\quad = 0.5 * (1 - 0.838) * 0.0211 + \frac{35}{65} * (1 - 0.838) * (1 - 0.0211) \\
&\quad = 0.871
\end{aligned}$$

The result is the same as when using BNT.

Q2.

Are 'Grass' type Pokémon easier to catch in the daytime?

We apply BNT(Bayesian Networks Toolbox) and find that $P(Time = true \mid Type = true)$ to be 0.764. We may also compute the probability by hand:

$$\begin{aligned}
P(Time, Type) &= P(Type|Time, Loca) * P(Time) * P(Loca) + \\
&\quad P(Type|Time, Loca') * P(Time) * P(Loca') \\
&\quad = 0 + \frac{13}{340} * 0.815 * (1 - 0.0211) \\
&\quad = 0.0305
\end{aligned} \tag{1}$$

$$\begin{aligned}
P(Type) &= P(Type|Time, Loca) * P(Time) * P(Loca) + \\
&\quad P(Type|Time, Loca') * P(Time) * P(Loca') + \\
&\quad P(Type|Time', Loca) * P(Time') * P(Loca) + \\
&\quad P(Type|Time', Loca') * P(Time') * P(Loca') \\
&\quad = 0 + 0 + \frac{13}{340} * 0.815 * (1 - 0.0211) + \frac{4}{77} * (1 - 0.815) * (1 - 0.0211) \\
&\quad = 0.0399
\end{aligned} \tag{2}$$

$$\begin{aligned}
P(Time|Type) &= \frac{P(Time, Type)}{P(Type)} \\
&= \frac{0.0305}{0.0399} \\
&= 0.764
\end{aligned}$$

Noticed that the result is the same as when using BNT.

Since 0.764 (Daytime) is larger than $1 - 0.764$ (Nighttime), grass type Pokémon are caught more easily in the day than at night.

Q3.

Given player level to be above 10, and the Pokémon was caught using less than 3 pokéballs, what is the probablity that the pokémon CP is between 100 and 400?

Apply BNT(Bayesian Networks Toolbox) and find P(CP=True|Pokeball=True,Level=False) to be 0.5616. This probability is quite hard to calculate by hand, since its answer will depend on all nodes. However, the idea is similar to Q1 and Q2, so we will truncate the calculation of the formula.

$$P(CP|Pokeball, Level) = \frac{P(CP, Pokeball, Level)}{P(Pokeball, Level)}$$

$$\begin{aligned}
P(CP, Pokeball, Level) = {} & P(CP, Pokeball, Level|WaitT, Type, CP, Level) * P(WaitT, Type, CP, Level)+ \\
& P(CP, Pokeball, Level|WaitT', Type, CP, Level) * P(WaitT', Type, CP, Level)+ \\
& P(CP, Pokeball, Level|WaitT, Type', CP, Level) * P(WaitT, Type', CP, Level)+ \\
& P(CP, Pokeball, Level|WaitT, Type, CP', Level) * P(WaitT, Type, CP', Level)+ \\
& P(CP, Pokeball, Level|WaitT, Type, CP, Level') * P(WaitT, Type, CP, Level')+ \\
& P(CP, Pokeball, Level|WaitT', Type', CP, Level) * P(WaitT', Type', CP, Level)+ \\
& P(CP, Pokeball, Level|WaitT', Type, CP', Level') * P(WaitT', Type, CP', Level)+ \\
& P(CP, Pokeball, Level|WaitT', Type, CP, Level') * P(WaitT', Type, CP, Level')+ \\
& P(CP, Pokeball, Level|WaitT, Type', CP', Level) * P(WaitT, Type', CP', Level)+ \\
& P(CP, Pokeball, Level|WaitT, Type', CP, Level') * P(WaitT, Type', CP, Level')+ \\
& P(CP, Pokeball, Level|WaitT, Type, CP', Level') * P(WaitT, Type, CP', Level')+ \\
& P(CP, Pokeball, Level|WaitT', Type', CP', Level) * P(WaitT', Type', CP', Level)+ \\
& P(CP, Pokeball, Level|WaitT', Type', CP, Level') * P(WaitT', Type', CP, Level')+ \\
& P(CP, Pokeball, Level|WaitT', Type, CP', Level') * P(WaitT', Type, CP', Level')+ \\
& P(CP, Pokeball, Level|WaitT, Type', CP', Level') * P(WaitT, Type', CP', Level')+ \\
& P(CP, Pokeball, Level|WaitT', Type', CP', Level') * P(WaitT', Type', CP', Level')+ \\
& ...
\end{aligned}$$

$P(Pokeball, Level) = ...$ calculation is similar

Appendix

We will list our code for the functions mentioned in the report over here in the appendix for easier reference.

Contents

Code to import filtered CSV data into *MATLAB*

```
function [timeStamp,pokeName,pokeType1,pokeType2,pokeCP,playerLvl,
    playerTeam,waitingTime,pokeLoc,weather,tofDay,lure,pokeStops,
    pokeBalls,razzBerry]=import();
input = textscan(fopen('Pokemon.csv'),'%s %s %s %s %f %f %s %f %s %s
    %s %s %f %f %s','Headerlines',1,'Delimiter',',');
timeStamp = input{1,1};
pokeName = input{1,2};
pokeType1 = input{1,3};
pokeType2 = input{1,4};
pokeCP = input{1,5};
playerLvl = input{1,6};
playerTeam = input{1,7};
waitingTime = input{1,8};
pokeLoc = input{1,9};
weather = input{1,10};
tofDay = input{1,11};
lure = input{1,12};
pokeStops = input{1,13};
pokeBalls = input{1,14};
razzBerry = input{1,15};
end
```

Code to print Bayesian Network

```
function graph1()
s=[1 1 2 3 4 5 6 7];
t=[2 5 8 2 5 8 7 8];
u={'Location' 'Waiting Time' 'Razz Berry' 'Time of Day' 'Type' '
    Level' 'CP' 'Pokeball'};
G = digraph(s,t);
plot(G,'NodeLabel',u);
axis off;
end
```

Code to print *PlayerLevel* histogram

```
function hist(data)
figure(2);
histogram(data);
title('Player Level Frequency');
xlabel('Player Level');
ylabel('Frequency');
end
```

Code to print *Pokémon* CP histogram

```
function hist(data)
figure(3);
histogram(data);
title('CP Frequency');
xlabel('CP Level');
ylabel('Frequency');
end
```

Code to print probability table for *PlayerLevel*

```
playerLvltab=zeros(max(playerLvl),2);
for i=1:max(playerLvl)
    for j=1:size(consolidatedtable,1)
        if playerLvl(j)==i;
            playerLvltab(i,1)=playerLvltab(i,1)+1;
        end
    end
    playerLvltab(i,1)=playerLvltab(i,1)/size(playerLvl,1);
    playerLvltab(i,2)=1-playerLvltab(i,1);
end
v=[1:max(playerLvl)]';
probtableplayerLvl = table(v,playerLvltab(:,1),playerLvltab(:,2),...
    'VariableNames',{'PlayerLevel','True','False'});
disp(probtableplayerLvl);
```

PlayerLevel	True	False
1	0.011737	0.98826
2	0.058685	0.94131
3	0.032864	0.96714
4	0.053991	0.94601
5	0.051643	0.94836
6	0.032864	0.96714
7	0.021127	0.97887
8	0.058685	0.94131
9	0.0070423	0.99296
10	0.046948	0.95305
11	0.0093897	0.99061
12	0.032864	0.96714
13	0.065728	0.93427
14	0.0070423	0.99296
15	0.035211	0.96479
16	0.077465	0.92254
17	0.042254	0.95775
18	0.011737	0.98826
19	0.0070423	0.99296
20	0.07277	0.92723
21	0.075117	0.92488
22	0.058685	0.94131
23	0.077465	0.92254
24	0.028169	0.97183
25	0.011737	0.98826
26	0.0023474	0.99765
27	0	1
28	0	1
29	0	1
30	0	1
31	0.0093897	0.99061

Code to print probability table for *Pokéballs* used

```
pokeBallstab=zeros(max(pokeBalls),2);
for i=1:max(pokeBalls)
    for j=1:size(consolidatedtable,1)
        if pokeBalls(j)==i;
            pokeBallstab(i,1)=pokeBallstab(i,1)+1;
        end
    end
    pokeBallstab(i,1)=pokeBallstab(i,1)/size(pokeBalls,1);
    pokeBallstab(i,2)=1-pokeBallstab(i,1);
end
v=[1:max(pokeBalls)]';
probtablepokeBallstab = table(v,pokeBallstab(:,1),pokeBallstab(:,2)
    ,...
    'VariableNames',{'Pokeballs','True','False'});
disp(probtablepokeBallstab);
```

Pokeballs	True	False
1	0.48826	0.51174
2	0.223	0.777
3	0.093897	0.9061
4	0.042254	0.95775
5	0.035211	0.96479
6	0.035211	0.96479
7	0.018779	0.98122
8	0.016432	0.98357
9	0.0023474	0.99765
10	0.0046948	0.99531
11	0.0046948	0.99531
12	0.0070423	0.99296
13	0.0046948	0.99531
14	0.0070423	0.99296
15	0	1
16	0	1
17	0.0070423	0.99296
18	0	1
19	0	1
20	0.0023474	0.99765
21	0	1
22	0	1
23	0	1
24	0	1
25	0	1
26	0	1
27	0.0023474	0.99765
28	0	1
29	0	1
30	0	1
31	0	1
32	0	1
33	0	1
34	0	1
35	0	1
36	0	1
37	0	1
38	0	1
39	0	1
40	0	1
41	0	1
42	0	1
43	0	1
44	0	1
45	0	1
46	0	1
47	0	1
48	0	1
49	0	1
50	0.0023474	0.99765
51	0	1

52	0	1
53	0	1
54	0	1
55	0	1
56	0	1
57	0	1
58	0	1
59	0	1
60	0.0023474	0.99765

Code to setup Bayesian Network for BNT

```
N=8;
dag=zeros(N,N);
Razz=1;Loca=2;Time=3;Level=4;WaitT=5;Type=6;CP=7;Pokeball=8;
dag([Razz Loca],WaitT)=1;
dag([Loca Time],Type)=1;
dag(Level,CP)=1;
dag([WaitT Type CP],Pokeball)=1;
discrete_nodes=1:N;
node_sizes=2*ones(1,N);
bnet=mk_bnet(dag,node_sizes,'names',{'Razz','Loca','Time','Level'
    ,...
    'WaitT','Type','CP','Pokeball'},'discrete',discrete_nodes);
```

We assume the following conditions to simplify our calculation:

- Razz=True : Razz Berry was used
- Loca=True : Pokémon was caught in water
- Time=True : Daytime
- Level=True : 1∼10
- WaitT=True : 0∼5
- Type=True : grass
- CP=True :CP 100∼400
- Pokéball=True : 1∼3

Then we assign all conditional probabilities with its respective values:

```
bnet.CPD{Razz}=tabular_CPD(bnet,Razz,[0.838,1-0.838]);
bnet.CPD{Loca}=tabular_CPD(bnet,Loca,[1-0.0211,0.0211]);
bnet.CPD{Time}=tabular_CPD(bnet,Time,[1-0.815,0.815]);
bnet.CPD{Level}=tabular_CPD(bnet,Level,[1-0.376,0.376]);
bnet.CPD{WaitT}=tabular_CPD(bnet,WaitT
    ,[1-290/365,1-35/65,1/5,0.5,290/365,...
    35/65,4/5,0.5]);
bnet.CPD{Type}=tabular_CPD(bnet,Type
    ,[1-4/77,1,1-13/340,1,4/77,0,13/340,0]);
bnet.CPD{CP}=tabular_CPD(bnet,CP
    ,[1-138/266,1-26/160,138/266,26/160]);
bnet.CPD{Pokeball}=tabular_CPD(bnet,Pokeball
    ,[1-25/55,1-167/195,1-1/4,...
    1-6/8,1-21/28,1-120/125,1-2/3,1-1/2,25/55,167/195,1/4,6/8,21/28,...
    120/125,2/3,1/2]);
draw_graph(dag,{'Razz Berry','Location','Time of Day','Level',...
    'Waiting Time','Type','CP','Pokeball'});
```

Q1.

Find the joint probability of using Razz Berry and Waiting Time to be less than or equal to 3 minutes.

The following lines of code compute $P(Razz = true, WaitT = true)$ in BNT:

```
engine=jtree_inf_engine(bnet);
evidence=cell(1,N);
[engine,loglik]=enter_evidence(engine,evidence);
marg=marginal_nodes(engine,[Razz WaitT]);
marg.T(2,2)
```

```
ans =

    0.0871
```

Q2.

Are Pokémon with type 'Grass' easier caught in the daytime?

```
engine=jtree_inf_engine(bnet);
evidence=cell(1,N);
evidence{Type}=2; % Noted here 1=False 2=True
[engine,loglik]=enter_evidence(engine,evidence);
marg=marginal_nodes(engine,Time);
marg.T()
fprintf('Since %f larger than %f\n',marg.T(2),marg.T(1));
fprintf('Then grass type is easier to caught in day than in night.\n
    ');
```

```
ans =

    0.2357
    0.7643

Since 0.764290 is larger than 0.235710
Grass type Pokémon are easier to catch in the day than at night.
```

Q3.

Given a player level to be above 10, and the Pokémon was caught using less than 3 Pokéballs. What is the probablity that Pokémon CP is between 100 and 400?

The following codes compute $P(CP = True \mid Pokeball = True, Level = False)$.

```
engine=jtree_inf_engine(bnet);
evidence=cell(1,N);
evidence{Level}=1;
evidence{Pokeball}=2;
[engine,loglik]=enter_evidence(engine,evidence);
marg=marginal_nodes(engine,CP);
marg.T(2)
```

```
ans =

    0.5616
```

Miscellaneous Data

- P(Razz) = $1 - 0.838$
- P(Loca) = 0.0211
- P(Time) = 0.815
- P(Level) = 0.376
- P(WaitT | Razz,Loca) = 0.5
- P(WaitT | Razz',Loca) = $\frac{4}{5}$
- P(WaitT | Razz,Loca') = $\frac{35}{65}$
- P(WaitT | Razz',Loca') = $\frac{290}{365}$
- P(Type | Loca,Time) = 0
- P(Type | Loca', Time) = $\frac{13}{340}$
- P(Type | Loca,Time') = 0
- P(Type | Loca',Time') = $\frac{4}{77}$
- P(CP | Level) = $\frac{26}{160}$
- P(CP | Level') = $\frac{138}{266}$
- P(Pokeball | type,CP) = $\frac{3}{5}$
- P(Pokeball | type,CP') = $\frac{7}{14}$
- P(Pokeball | type',CP) = $\frac{141}{155}$
- P(Pokeball | type',CP') = $\frac{192}{268}$
- P(Pokeball | CP,type,WaitT) = $\frac{1}{2}$
- P(Pokeball | CP,type,WaitT') = $\frac{2}{3}$
- P(Pokeball | CP,type',WaitT) = $\frac{120}{125}$
- P(Pokeball | CP,type',WaitT') = $\frac{21}{28}$
- P(Pokeball | CP',type,WaitT) = $\frac{6}{8}$
- P(Pokeball | CP',type,WaitT') = $\frac{1}{4}$
- P(Pokeball | CP',type',WaitT) = $\frac{167}{195}$
- P(Pokeball | CP',type',WaitT') = $\frac{25}{55}$