**CSci 384: Artificial Intelligence Spring, 2020**

**Instructor: Dr. M. E. Kim** **Date: April 23, 2020**

**Due: by the end of the day, May 4rd (Mon.), 2020**

**Assignment 7: Decision Tree Learning & Statistical Learning (184/200 points)**

**Instruction:**

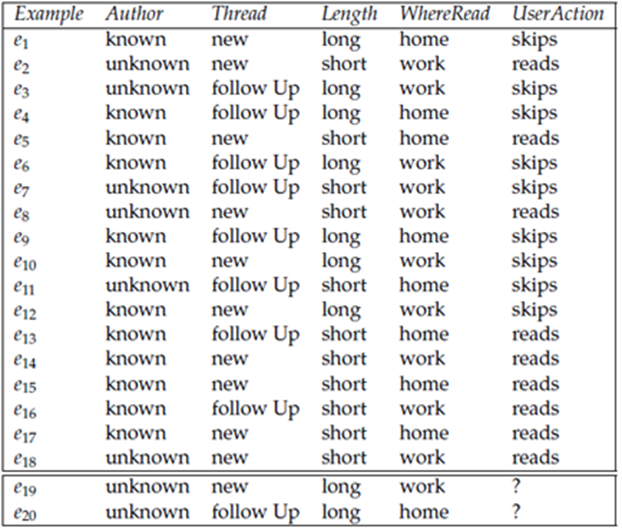
For each question, show the computational steps clearly to compute the entropy and the probability, etc.

If you’d like to, **you can write a short program to compute an entropy, decision tree** and **any probability value in ML/MAP learning,** etc. In such a case, insert the image of output in the corresponding section and submit your source codes + the output images in the compressed folder.

Read the submission instruction and comply with it.

**Q1. [93/100] Inductive Decision Tree Learning**

The table below shows training and test examples typical of a classification task.



The aim is to predict whether a person reads an article posted to a bulletin board given properties of the article.

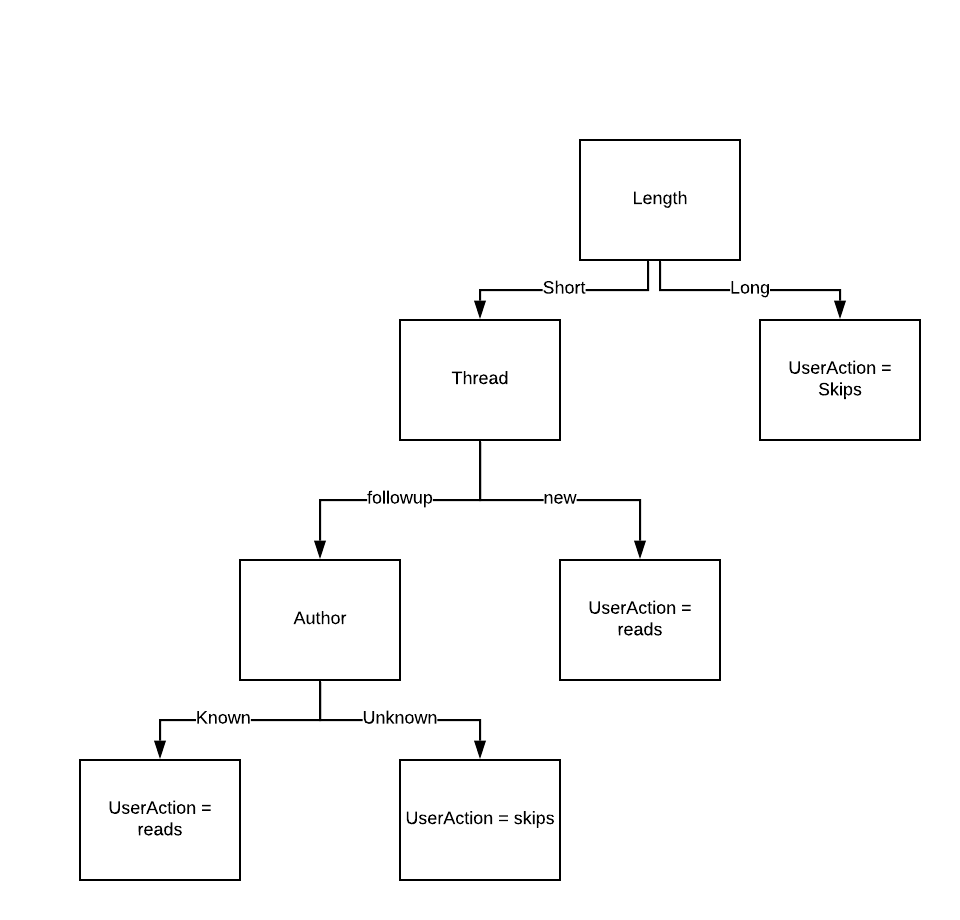
The input features are *Author*, *Thread*, *Length*, and *WhereRead*. There is one target feature, *UserAction*. The domain of *Author* is {known, unknown}, the domain of *Thread* is {new, followup}, and so on.

There are **18 *training examples***, each of which has a value for all of the features.

In this data set, *val*(*e*11, *Author*)=*unknown*, *val*(*e*11, *Thread*)=*followUp*, and *val*(*e*11, *UserAction*)=*skips*.

There are **2 *test examples***, **e19** and **e20**, where the user action is unknown.

The aim is to predict the user action for a new example given its values for the input features.

1. [10/10] Compute the ***initial entropy*** of *UserAction.*
   1. E(S) = -P(reads)log2P(reads) - P(skips)log2P(skips)
   2. = -P(9/18)log2P(9/18) - P(9/18)log2P(9/18)
   3. = 1.0
2. [10/10] From the training examples (e1 – e18), construct an optimal decision tree that classifies the data the best. Show the proper computation of information gain with the entropy of each variable. If there is conflicting description in an example, give the estimated probabilities of each classification using the relative frequencies.
   1. Author
      1. E(author = known) = -(6/12log26/12) – 6/12log26/12 = 1
      2. E(author = unknow) = -(3/6log23/6)- 3/6log23/6 = 1
      3. I(author) = 12/18\*1+ 6/18\*1 = 1
      4. Gain(author) = 1- 1 = 0
   2. Thread
      1. E(thread= new) = -(7/10log27/10) – 3/10log23/10 = .88129
      2. E(thread=followup) = -(2/8log22/8) – 6/8log26/8= .811278
      3. I(thread) =10/18 \* .88129 + 8/18 \* .811278 = 0.850168
      4. Gain = 1- .850168 = .149832
   3. Length
      1. E(length = short) = -(9/11log29/11) – 2/11log22/11 = .668406
      2. E(length = long) = 0
      3. I(length) = 11/18 \* .668406 = 0.418
      4. Gain = 1 - 0.418 = .582
   4. WhereRead
      1. E(WhereRead=home) =-(4/8log24/8) – 4/8log24/8 = 1
      2. E(WhereRead=work) =-(5/10log25/10) – 5/10log25/10 = 1
      3. I(WhereRead) = 8/18\*1 + 10/18\*1 = 1
      4. Gain = 1-1 = 0
   5. 

The rest of the computation must be in the program. See the solution.

1. [6/10] Express the hypothesis generated at 2) in the logical expression.
   1. L = short,L = long, N = new, N = followup, A=known, A = unknown, U = reads, U= skips
   2. (A∧N∧L) ∨(N∧L)  U

x UserAction(x) = reads  (Length(x)=Short  Thread(x) = New)  (Length(x) = Short  Thread(x) = Follow UP  Author(x) = Known) where x is a datum

1. [10/10] For the test data e19 and e20, predict the user’s action based on the hypothesis in 2) – 3), respectively.
   1. From the data e19 it will most probably be skip because the article is long
   2. e20 will also will most likely be skip because the article is long

From the same training examples (e1 – e18),

1. [30/30] Decide the following probabilities that can be derived from the given data of the training examples in the table.

The numeric values are correct, but the way of computation is not what was asked.

What you computed is what is learned from the given data.

What was asked in the question is what can be derived from the given data by counting the frequencies.

1. [4] P( *UserAction* = reads) = 9/18
2. [3] P( *Author* = known | *UserAction* = reads) =
   1. P(*Author* = known | *UserAction* = reads) = P( *Author* = known & *UserAction* = reads) / P(*UserAction* = reads)
   2. = (6/18 )/ (9/18) = 2/3
3. [3] P( *Author* = known | *UserAction* = skips) =
   1. P(*Author* = known | *UserAction* = skips) = P( *Author* = known & *UserAction* = skips) / P(*UserAction* = skips)
   2. = (6/18 )/ (9/18) = 2/3
4. [3] P( *Thread* = new | *UserAction* = reads) =
   1. P(*Thread* = new | *UserAction* = reads) = P( *Thread* = new & *UserAction* = reads) / P(*UserAction* = reads)
   2. = (7/18 )/ (9/18) = 7/9
5. [3] P( *Thread* = new | *UserAction* = skips) =
   1. P(*Thread* = new | *UserAction* = skips) = P( *Thread* = new & *UserAction* = skips) / P(*UserAction* = skips)
   2. = (3/18 )/ (9/18) = 1/3
6. [3] P( *Length* = long | *UserAction* = reads) =
   1. P( *Length* = long & *UserAction* = reads) / P(*UserAction* = reads)
   2. (0/18 )/ (9/18 )= 0
7. [3] P( *Length* = long | *UserAction* = skips) = ?
   1. P( *Length* = long & *UserAction* = skips) / P(*UserAction* = skips)
   2. (7/18 )/ (9/18 )= 7/9
8. [4] P( *WhereRead* = home | *UserAction* = reads) = ?
   1. P( *WhereRead* = home & *UserAction* = reads) / P(*UserAction* = reads)
   2. (4/18)/(9/18) = 4/9
9. [4] P( *WhereRead* = home | *UserAction* = skips) = ?
   1. P( *WhereRead* = home & *UserAction* = skips) / P(*UserAction* = skips)
   2. (4/18)/(9/18) = 4/9
10. [10/10] By means of ***Naïve Bayes Classifier model***, predict the classification, *UserAction*, of the test data **e19**. You have to show the computations of classification probabilities. i.e. decide ***CNB***.
    1. Unkown, new , long, work
    2. ***CNB*** = P( *UserAction* = reads)P(author = unknown| *UserAction* = reads)P(thread = new| *UserAction* = reads)P(length = long| *UserAction* = reads)P(whereread = work| *UserAction* = reads)
       1. 9/18\*~~3/18\*7/9\*0\*4/18~~ = 0
    3. ***CNB*** = P( *UserAction* = skips)P(author = unknown| *UserAction* = skips)P(thread = new| *UserAction* = skips)P(length = long| *UserAction* = skips)P(whereread = work| *UserAction* = skips)
       1. 9/18 \* ~~3/18\*1/3\*5/18~~ = .0077 -- normalize it.
    4. ***CNB*** =(UserAction = skips)

e19 <unknown, new, long, work, ?>

P(UserAction = read) = 9/18 = .5 P(UserAction = skip) = 9/18 = .5

P(UserAction = read | e19)

=α P(read) \* P(unknown | read) \* P(new | read) \* P(long | read) \* P(work | read)

= α 9/18 \* 3/9 \* 7/9 \* 0/9 \* 5/9 = 0

P(UserAction = skips | e19)

= α P(skip) \* P(unknown | skip) \* P(new | skip) \* P(long | skip) \* P(work | skip)

= α 9/18 \* 3/9 \* 3/9 \* 7/9 \* 5/9 = .024α = 1

Since P(UserAction = read | e19) < P(UserAction = skips | e19),

**the Naïve Bayes classifier ( ) for e19 is:**

**i.e. the prediction of userAction for e19 is skips.**

1. [10/10] What is the predicted classification probability of **e19**, i.e. P(***CNB*** | e19), and its prediction?
   1. P(UserAction = skips| Unkown, new , long, work) =~~.0077/(0+.0077)~~
   2. = 1

< P(reads | e19), P(skips | e19) > =  <0, .0265> = <0, 1>, so, P(***CNB*** | e19) = 1.

1. [8/10] What is the predicted classification probability of **e20**, i.e. P(***CNB*** | e20), and its prediction?
   1. Unknown, followup, long, home, skips
      1. 9/18\*~~3/18\*6/18\*4/9~~= ~~0.0123~~
      2. Useraction = no
      3. 0.0123/(0+0.0123) = 1

e20 <unknown, follow-up, long, home>

P(reads| e20)

= P(read) \* P(unknown | read) \* P(follow-up | read) \* P(long | read) \* P(home | read)

= 9/18 \* 3/9 \* 2/9 \* 0/9 \* 4/9 = 0

P(skips | e20)

= P(skip) \* P(unknown | skip) \* P(follow-up | skip) \* P(long | skip) \* P(home | skip)

= 9/18 \* 3/9 \* 6/9 \* 7/9 \* 4/9 = .5 \* .33 \* .67 \* .78 \* .44 = .0379

Thus CNB = (UserAction=skip)

P(UserAction=skip | e20> = .0379 / (.0379+0) = 1.0

**Q2. [16/20] Maximum Likelihood (ML) Learning**

Suppose we toss a thumbtack 4 times and we observe the sequence [heads, tails, heads, heads].

Let *T1*, *T2*, *T3* and *T4* be random variables such that the value of is the outcome on the *ith* toss.

1. [10/10] Give the maximum likelihood (ML) estimate of the probability of heads, P( = head).

Hint: Define P(heads) = *p*. Then, compute the parameter value of *p* using ML.

P(heads) = p P(tails) = t

p log() + t log(1-)

3log + 1log(1-)

 = c/c+l = c/N

 = 3/4

1. [6/10] Suppose that the data is identically independently distributed.

What is the ML estimate of P( = head, = tail, = head, = head) ?

=p3(1-p)

=T’3(1-T’)

P( = head, = tail, = head, = head)

= P(heads) \* P(tails) \* P(heads) \* P(heads)

= .75 \* .25 \* .75 \* .75 = .1055

**Q3. [55/60] Maximum Likelihood (ML) Learning**

The table shows the examples of SPAM and those of HAM messages which are consisted of some words

whose dictionary size is twelve. Suppose that you’ve received SPAM messages for the 1st 3 days, then

HAM messages for the next 5 days, i.e. a data as a sequence of message is <Spam, Spam, Spam, Ham,

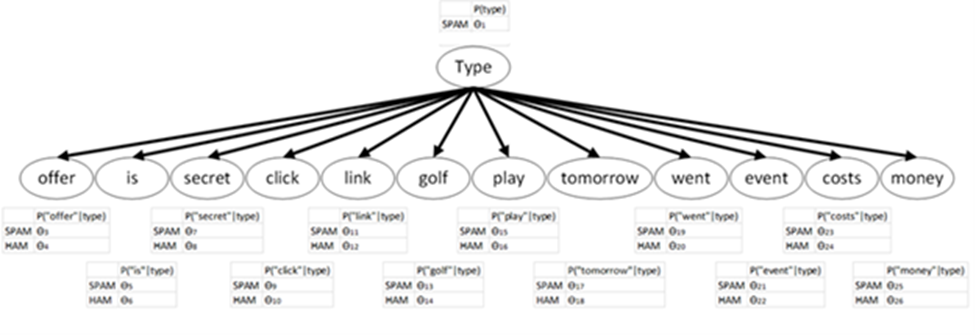
Ham, Ham, Ham, Ham >.

|  |  |
| --- | --- |
| *SPAM* | *HAM* |
| offer is secret | play golf tomorrow |
| click secret link | went play golf |
| secret golf link | secret golf event |
|  | golf is tomorrow |
|  | golf costs money |

1. [10/10] Compute the ***maximum likelihood*** of *SPAM, i.e.* P(*SPAM*)=, using a *log-likelihood.*
   1. 3log + 5log(1-)
   2.  = c/c+l
   3. = 3/3+5
   4.  = 3/8
2. [5/10] In the Bayes net of this ML parameter learning,
3. how many parameters are required?
   1. 2N+1 parameters = 2\*12words + 1 = 25 parameters?

(b) Draw the BN with the CPT of the required parameters (e.g. 1, 2, 3, …..). -- You don’t

yet have to compute the exact values of parameters.



1. [10/10] By ML-learning, compute a *parameter value*, P(“*secret*” | *SPAM*) and P(“*secret*” | *HAM*), respectively, using a log-likelihood.
   1. P("secret"|SPAM)= 3/9 =1/3 = 0.333
   2. P(“*secret*” | *HAM*) = 1/15 = 0.0667
2. [10/10] Now, the new message “*golf*” is received. Compute the *likelihood* of this message is *SPAM*.
   1. P(SPAM|"golf)= [P("golf"|SPAM)\*P(SPAM)]/P("golf")
   2. =((1/9) \* (3/8))/6/24 =1/6 = 0.167

First of all, we have to learn the parameter values to use them for the above computation.

P(SPAM| golf , )

=α P(SPAM| ) P(golf | SPAM, )

=α 11  Similarly, 11  has to be learned by ML learning, =

= α

P(HAM| golf, )

= α P(HAM| )⋅ P(golf | HAM, )

= α (1-θ)⋅θ12  Similarly, θ12 has to be learned by ML learning, =

= α

After normalizing the above, < 1/24, 5/24> = <.1667, .8333 >

1. [10/10] The new message “*secret is secret*” is received. What is the *likelihood* of this message is *SPAM*?
   1. P(“secret is secret”|SPAM) \* P(SPAM)+P(“secret is secret”|HAM)\*P(HAM)
   2. (3/8\* 1\*1/3\*1/(3/8\*1/3+5/8\*1/5\*1/5\*1/5) = 25/26 = 0.962

P(C=SPAM|W=secret, is, secret)

=

= = 

P(C=HAM|W=secret, is, secret)

=

= =

So, < , > = < 5400, 216 > =< 25, 1 >

= < 25/26, 1/26 > = <.9615, .0285>

i.e.

P(C=SPAM|W=secret, is, secret) =

1. [10/10] For a new message, “*tomorrow is secret*”, what is the likelihood of this message is *SPAM* and *HAM*, respectively?
   1. P(C=SPAM|W=tomorrow, is, secret) = 0
   2. P(C=HAM|W=tomorrow, is, secret) = 1 given it won’t be spam

P(C=SPAM|W=tomorrow, is, secret)

=

= = 0 = 0

P(C=HAM|W=tomorrow, is, secret)

=

=

**Q4. [20/20] MAP Learning**

In the data of Candy Example of Chap. 20 (slide #6), 6 candies are unwrapped one by one and a flavor of each candy is as follows: d1 = lime, d2 = cherry, d3= cherry, d4 = lime, d5 = cherry, d6 = cherry.

1. [10/10] By computing the posterior probability of each hypothesis, given the above six data, decide a ***Maximum A Posteriori hypothesis*** (***hMAP***) where each *hi* is defined in the slide #6.

You have to show the essential computational steps if you compute them manually or using an excel sheet. Otherwise, write a program to compute them. – Refer to the handout of Candy Example on the blackboard and follow the submission instruction.

1. lime
2. 𝑃𝑃(ℎ1|d1= lime) = α⋅ P(d1|ℎ1)𝑃𝑃(ℎ1) = α⋅0⋅ 0.1 = 0α ⇒ 0
3. 𝑃𝑃(ℎ2|d1 = lime) = α⋅ P(d1|ℎ2)𝑃𝑃(ℎ2) = α⋅.25⋅ .2 = .05α ⇒ .1
4. 𝑃𝑃(ℎ3|d1 = lime) = α⋅ P(d1|ℎ3)𝑃𝑃(ℎ3) = α⋅.5⋅ .4 = .2α ⇒ .4
5. 𝑃𝑃(ℎ4|d1 = lime) = α⋅ P(d1|ℎ4)𝑃𝑃(ℎ4) = α⋅.75⋅ .2 = .15α ⇒ .3
6. 𝑃𝑃(ℎ5|d1 = lime) = α⋅ P(d1|ℎ5)𝑃𝑃(ℎ5) = α⋅1⋅ .1 = .1α ⇒ .2
7. Lime, cherry
8. 𝑃𝑃(ℎ1|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ1)𝑃𝑃(ℎ1) =α⋅P(d1= lime|ℎ1)⋅P(d2=cherry|ℎ1)𝑃𝑃(ℎ1) = α⋅02⋅⋅ .1 = 0α ⇒ 0
9. 𝑃𝑃(ℎ2|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ2)𝑃𝑃(ℎ2) =α⋅P(d1= lime|ℎ2)⋅P(d2=cherry|ℎ2)𝑃𝑃(ℎ2) = α⋅.25⋅.75\* .2 = .0375α ⇒ .2142
10. 𝑃𝑃(ℎ3|d1 lime, d2= cherry) = α⋅ P(d1, d2|ℎ3)𝑃𝑃(ℎ3) =α⋅P(d1= lime|ℎ3)⋅P(d2=cherry|ℎ3)𝑃𝑃(ℎ3) = α⋅.52 .4 = .1α ⇒ .5714
11. 𝑃𝑃(ℎ4|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ4)𝑃𝑃(ℎ4) =α⋅P(d1= lime|ℎ4)⋅P(d2=cherry|ℎ4)𝑃𝑃(ℎ4) = α⋅.75\*.25\* .2 = .0375α ⇒ .2142
12. 𝑃𝑃(ℎ5|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ5)𝑃𝑃(ℎ5) =α⋅P(d1= lime|ℎ5)⋅P(d2=cherry|ℎ5)𝑃𝑃(ℎ5) = α⋅ 12⋅ .1\*0 = .1α ⇒ 0
13. Lime, cherry, cherry
14. 𝑃𝑃(ℎ1|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ1)𝑃𝑃(ℎ1) =α⋅P(d1= lime|ℎ1)⋅P(d2=cherry|ℎ1)𝑃𝑃(ℎ1) = α⋅02⋅⋅ .1 = 0α ⇒ 0
15. 𝑃𝑃(ℎ2|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ2)𝑃𝑃(ℎ2) =α⋅P(d1= lime|ℎ2)⋅P(d2=cherry|ℎ2)𝑃𝑃(ℎ2) = α⋅.25⋅.75\*.75 .2 = .0281α ⇒ .3211
16. 𝑃𝑃(ℎ3|d1 lime, d2= cherry) = α⋅ P(d1, d2|ℎ3)𝑃𝑃(ℎ3) =α⋅P(d1= lime|ℎ3)⋅P(d2=cherry|ℎ3)𝑃𝑃(ℎ3) = α⋅.53 .4 = .05α ⇒ .5714
17. 𝑃𝑃(ℎ4|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ4)𝑃𝑃(ℎ4) =α⋅P(d1= lime|ℎ4)⋅P(d2=cherry|ℎ4)𝑃𝑃(ℎ4) = α⋅.75\*.25\*.25\* .2 = .0094α ⇒ .1074
18. 𝑃𝑃(ℎ5|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ5)𝑃𝑃(ℎ5) =α⋅P(d1= lime|ℎ5)⋅P(d2=cherry|ℎ5)𝑃𝑃(ℎ5) = α⋅ 12⋅ .1\*0 = .1α ⇒ 0
19. Lime, cherry, cherry, lime
20. 𝑃𝑃(ℎ1|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ1)𝑃𝑃(ℎ1) =α⋅P(d1= lime|ℎ1)⋅P(d2=cherry|ℎ1)𝑃𝑃(ℎ1) = α⋅02⋅⋅ .1 = 0α ⇒ 0
21. 𝑃𝑃(ℎ2|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ2)𝑃𝑃(ℎ2) =α⋅P(d1= lime|ℎ2)⋅P(d2=cherry|ℎ2)𝑃𝑃(ℎ2) = α⋅.25⋅.75\*.75\*.25\* .2 = .007α ⇒ .1795
22. 𝑃𝑃(ℎ3|d1 lime, d2= cherry) = α⋅ P(d1, d2|ℎ3)𝑃𝑃(ℎ3) =α⋅P(d1= lime|ℎ3)⋅P(d2=cherry|ℎ3)𝑃𝑃(ℎ3) = α⋅.54 .4 = .025α ⇒ .6410
23. 𝑃𝑃(ℎ4|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ4)𝑃𝑃(ℎ4) =α⋅P(d1= lime|ℎ4)⋅P(d2=cherry|ℎ4)𝑃𝑃(ℎ4) = α⋅.75\*.25\*.25\*.75\* .2 = .007α ⇒ .1795
24. 𝑃𝑃(ℎ5|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ5)𝑃𝑃(ℎ5) =α⋅P(d1= lime|ℎ5)⋅P(d2=cherry|ℎ5)𝑃𝑃(ℎ5) = α⋅ 12⋅ .1\*0 = .1α ⇒ 0
25. Lime, cherry, cherry, lime, cherry
26. 𝑃𝑃(ℎ1|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ1)𝑃𝑃(ℎ1) =α⋅P(d1= lime|ℎ1)⋅P(d2=cherry|ℎ1)𝑃𝑃(ℎ1) = α⋅02⋅⋅ .1 = 0α ⇒ 0
27. 𝑃𝑃(ℎ2|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ2)𝑃𝑃(ℎ2) =α⋅P(d1= lime|ℎ2)⋅P(d2=cherry|ℎ2)𝑃𝑃(ℎ2) = α⋅.25⋅.75\*.75\*.25\*.75 .2 = .0052α ⇒ .2666
28. 𝑃𝑃(ℎ3|d1 lime, d2= cherry) = α⋅ P(d1, d2|ℎ3)𝑃𝑃(ℎ3) =α⋅P(d1= lime|ℎ3)⋅P(d2=cherry|ℎ3)𝑃𝑃(ℎ3) = α⋅.55 .4 = .0125α ⇒ .6410
29. 𝑃𝑃(ℎ4|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ4)𝑃𝑃(ℎ4) =α⋅P(d1= lime|ℎ4)⋅P(d2=cherry|ℎ4)𝑃𝑃(ℎ4) = α⋅.75\*.25\*.25\*.75\*.25\* .2 = .0018α ⇒ .0923
30. 𝑃𝑃(ℎ5|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ5)𝑃𝑃(ℎ5) =α⋅P(d1= lime|ℎ5)⋅P(d2=cherry|ℎ5)𝑃𝑃(ℎ5) = α⋅ 12⋅ .1\*0 = .1α ⇒ 0
31. Lime, cherry, cherry, lime, cherry, cherry
32. 𝑃𝑃(ℎ1|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ1)𝑃𝑃(ℎ1) =α⋅P(d1= lime|ℎ1)⋅P(d2=cherry|ℎ1)𝑃𝑃(ℎ1) = α⋅02⋅⋅ .1 = 0α ⇒ 0
33. 𝑃𝑃(ℎ2|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ2)𝑃𝑃(ℎ2) =α⋅P(d1= lime|ℎ2)⋅P(d2=cherry|ℎ2)𝑃𝑃(ℎ2) = α⋅.25⋅.75\*.75\*.25\*.75\*.75 .2 = .0039α ⇒ .3696
34. 𝑃𝑃(ℎ3|d1 lime, d2= cherry) = α⋅ P(d1, d2|ℎ3)𝑃𝑃(ℎ3) =α⋅P(d1= lime|ℎ3)⋅P(d2=cherry|ℎ3)𝑃𝑃(ℎ3) = α⋅.56 .4 = .00625α ⇒ .5924
35. 𝑃𝑃(ℎ4|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ4)𝑃𝑃(ℎ4) =α⋅P(d1= lime|ℎ4)⋅P(d2=cherry|ℎ4)𝑃𝑃(ℎ4) = α⋅.75\*.25\*.25\*.75\*.25\*.25\* .2 = .0004α ⇒ .0379
36. 𝑃𝑃(ℎ5|d1= lime, d2= cherry) = α⋅ P(d1, d2|ℎ5)𝑃𝑃(ℎ5) =α⋅P(d1= lime|ℎ5)⋅P(d2=cherry|ℎ5)𝑃𝑃(ℎ5) = α⋅ 12⋅ .1\*0 = .1α ⇒ 0
37. hmap= h3
38. [10/10] The 7th candy is about to be unwrapped. Compute the prediction probability,

<P(d7= cherry | ***hMAP***), P(d7= lime | ***hMAP***) > based on ***hMAP*** in (1), and give a prediction on the flavor of the 7th candy.

<P(d7= cherry|h3), P(d7=lime|h3)> = <0.5,0.5>