Teresa Liu

CSCI 467 PS 5

1. a) $\frac{n-1}{n} = 1 - \frac{1}{n}$ since there is a $\frac{1}{n}$ chance that the 1st observation is the j^{th} observation from the original sample, so $1 - \frac{1}{n}$ is the negation

b) $\frac{n-1}{n} = 1 - \frac{1}{n}$ for the same reason as part a since the 2nd and 1st observations are independent. (with replacement)

c) $(1 - \frac{1}{n})^n$ Since bootstrap draws observation with replacement, the chance of draving the j^{th} observation from the ongnal sample for any of the n draws in the bootstrap sample is $1 - \frac{1}{n}$. Since there's n observations in the bootstrap sample, and since each observation is independent since there is replacement, the chance of the bootstrap sample not containing the j^{th} observation is $(1 - \frac{1}{n})^n$.

d) $1 - (1 - \frac{1}{5})^5 = 1 - (\frac{4}{5})^5 = 0.6723$

e) $1 - (1 - \frac{1}{100})^{100} = 1 - (\frac{99}{100})^{100} = 0.6340$

f) $1 - (1 - \frac{1}{10000})^{10000} = 1 - (\frac{9999}{10000})^{10000} = 0.6321$

When n is large, the percentage of overlap between two bootstrap samples is around $\boxed{1/3 = 33\%}$ of observations Since each bootstrap sample will contain ~2/3 of the observations, with two samples there will be ~1/3 overlap.

→ $\lim_{n \to \infty} (1 + \frac{\lambda}{n})^n = e^{\lambda}$

→ when $\lambda = -1$, then $\lim_{n \to \infty} (1 - \frac{1}{n})^n = e^{-1} \approx \frac{1}{3}$ which represents the chance that the j^{th} observation will not be in the bootstrap sample of n draws from the n total samples. Therefore, the bootstrap sample will contain around 2/3 of the observations, which will amount to ~1/3 of overlap between 2 bootstrap samples that each contain ~ 2/3 of the observations.

→ bootstrap cannot be used like cross validation in training algorithms to estimate errors since it uses replacement to draw samples, so observation may overlap between validation and training set which you can't have for estimating error

$$= x_{11}^2 + x_{21}^2 + 2x_{11}x_{21}$$
$$= (x_{11} + x_{21})(x_{11} + x_{21})$$

2. $n = 2,\ p = 2,\ x_{11} = x_{12}\quad x_{21} = x_{22}$

$y_1 + y_2 = 0\quad x_{11} + x_{21} = 0\quad x_{12} + x_{22} = 0$ so estimate $\hat{\beta}_0 = 0$

a) minimize:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \hat{\beta}^2_j$$

$$= \boxed{(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)}$$

b) $= y_1^2 - 2y_1\hat{\beta}_1 x_{11} - 2y_1\hat{\beta}_2 x_{12} + \hat{\beta}_1^2 x_{11}^2 + \hat{\beta}_2^2 x_{12}^2 + 2\hat{\beta}_1\hat{\beta}_2 x_{11}x_{12}$
$+ y_2^2 - 2y_2\hat{\beta}_1 x_{21} - 2y_2\hat{\beta}_2 x_{22} + \hat{\beta}_1^2 x_{21}^2 + \hat{\beta}_2^2 x_{22}^2 + 2\hat{\beta}_1\hat{\beta}_2 x_{21}x_{22}$
$+ \lambda\hat{\beta}_1^2 + \lambda\hat{\beta}_2^2$

$\frac{d}{d\beta_1} = -2y_1 x_{11} + 2x_{11}^2\hat{\beta}_1 + 2\hat{\beta}_2 x_{11}x_{12} - 2y_2 x_{21} + 2x_{21}^2\hat{\beta}_1 + 2\hat{\beta}_2 x_{21}x_{22}$
$+ 2\lambda\hat{\beta}_1 = 0$

$\Rightarrow -y_1 x_{11} + x_{11}^2\hat{\beta}_1 + \hat{\beta}_2 x_{11}x_{12} - y_2 x_{21} + x_{21}^2\hat{\beta}_1 + \hat{\beta}_2 x_{21}x_{22} + \lambda\hat{\beta}_1 = 0$

with $x_{11} = x_{12}\quad x_{21} = x_{22}$

$-y_1 x_{11} + x_{11}^2\hat{\beta}_1 + \hat{\beta}_2 x_{11}^2 - y_2 x_{21} + x_{21}^2\hat{\beta}_1 + \hat{\beta}_2 x_{21}^2 + \lambda\hat{\beta}_1$

$\lambda\hat{\beta}_1 + \hat{\beta}_1(x_{11}^2 + x_{21}^2) + \hat{\beta}_2(x_{11}^2 + x_{21}^2) = y_1 x_{11} + y_2 x_{21}$ and

$\Rightarrow \lambda\hat{\beta}_1 + \hat{\beta}_1(x_{11} + x_{21})^2 + \hat{\beta}_2(x_{11} + x_{21})^2 = y_1 x_{11} + y_2 x_{21} + 2\hat{\beta}_1 x_{11}x_{21} + 2\hat{\beta}_2 x_{11}x_{21}$

$\Rightarrow \lambda\hat{\beta}_1 = y_1 x_{11} + y_2 x_{21} + 2\hat{\beta}_1 x_{11}x_{21} + 2\hat{\beta}_2 x_{11}x_{21}$

$\frac{d}{d\beta_2} = -2y_1 x_{12} + 2x_{12}^2\hat{\beta}_2 + 2\hat{\beta}_1 x_{11}x_{12} - 2y_2 x_{22} + 2\hat{\beta}_2 x_{22}^2 + 2\hat{\beta}_2 x_{21}x_{22}$
$+ 2\lambda\hat{\beta}_2 = 0$

$\Rightarrow -y_1 x_{12} + x_{12}^2\hat{\beta}_2 + \hat{\beta}_1 x_{11}x_{12} - y_2 x_{22} + \hat{\beta}_2 x_{22}^2 + \hat{\beta}_1 x_{21}x_{22} + \lambda\hat{\beta}_2 = 0$

with $x_{11} = x_{12}\quad x_{21} = x_{22}$

$-y_1 x_{11} + x_{11}^2\hat{\beta}_2 + \hat{\beta}_1 x_{11}^2 - y_2 x_{21} + \hat{\beta}_2 x_{21}^2 + \hat{\beta}_1 x_{21}^2 + \lambda\hat{\beta}_2 = 0$

$\Rightarrow \lambda\hat{\beta}_2 + \hat{\beta}_2(x_{11}^2 + x_{21}^2) + \hat{\beta}_1(x_{11}^2 + x_{21}^2) = y_1 x_{11} + y_2 x_{21}$

$\Rightarrow \lambda\hat{\beta}_2 = y_1 x_{11} + y_2 x_{21} + 2\hat{\beta}_1 x_{11}x_{21} + 2\hat{\beta}_2 x_{11}x_{21}$

so $\lambda\hat{\beta}_1 = \lambda\hat{\beta}_2$

so $\boxed{\hat{\beta}_1 = \hat{\beta}_2}$

c)

c) lasso : minimize $RSS + \lambda \sum_{j=1}^{p} |\beta_j|$

$= (y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(|\hat{\beta}_1| + |\hat{\beta}_2|)$

d) lasso coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are not unique in both $\hat{\beta}_2$ and $\hat{\beta}_1$ can either be both positive or both negative and produce the same solution.

$\frac{d}{d\beta_1} \lambda|\hat{\beta}_1| = \lambda \frac{|\hat{\beta}_1|}{b}$

$\frac{\lambda|\hat{\beta}_1|}{\hat{\beta}_2} = \frac{\lambda|\hat{\beta}_2|}{\hat{\beta}_2}$ So solutions include:

$\hat{\beta}_1$ and $\hat{\beta}_2$ are both positive

or $\hat{\beta}_1$ and $\hat{\beta}_2$ are both negative.

3. depth 1

DRESS COLOR pink: 5/8 icecream nonpink: 6/8 marshmallow

LIKES SINGING yes: 5/8 icecream no: 6/8 marshmallow

LIKES DANCING: yes: 5/8 icecream no: 6/8 marshmallow

AGE: child: 5/8 icecream adult: 6/8 marshmallow

select DRESS COLOR as 1st node

depth 2

with DRESS COLOR = pink err rate

LIKES SINGING: yes: 4/4 icecream no: 3/4 marshmallow $0 + \frac{1}{4}$

LIKES DANCING: yes: 3/4 icecream no: 2/4 marshmallow $\frac{1}{4} + \frac{1}{2}$

AGE: child: 3/4 icecream adult: 2/4 marshmallow $\frac{1}{4} + \frac{1}{2}$

select LIKES SINGING as 2nd node

with DRESS COLOR = nonpink:

LIKES SINGING yes: 3/4 marshmallow no: 3/4 marshmallow $\frac{1}{4} + \frac{1}{4}$

LIKES DANCING yes: 2/4 marshmallow no: 4/4 marshmallow $\frac{1}{2} + 0$

AGE child: 2/4 icecream adult: 4/4 marshmallow $\frac{1}{2} + 0$

select LIKES SINGING as 2nd node

depth 3

with DRESS COLOR = pink, LIKES SINGING = yes ⇒ classify to icecream

with DRESS COLOR = pink, LIKES SINGING = no:

LIKES DANCING: yes: 1/2 icecream no: 2/2 marshmallow $\frac{1}{2}$

AGE child: 1/2 icecream adult: 2/2 marshmallow $\frac{1}{2}$

select LIKES DANCING as 3rd node

with DRESS COLOR = nonpink, LIKES SINGING = yes,

LIKES DANCING: yes: 1/2 icecream no: 2/2 marshmallow $\frac{1}{2}$

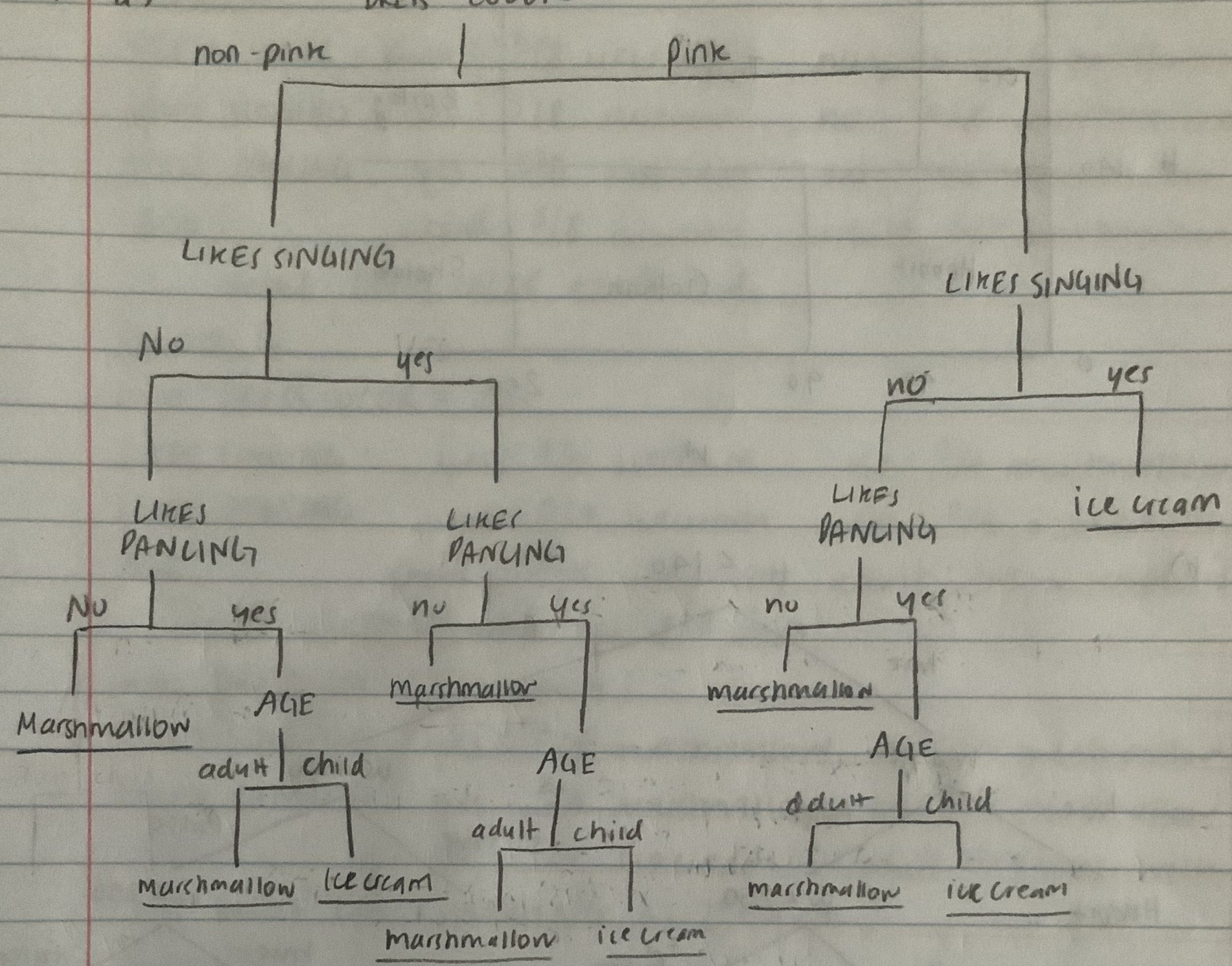
AGE child: 1/2 icecream adult: 2/2 marshmallow $\frac{1}{2}$

select LIKES DANCING as 3rd node

with DRESS COLOR = nonpink, LIKES SINGING = no,

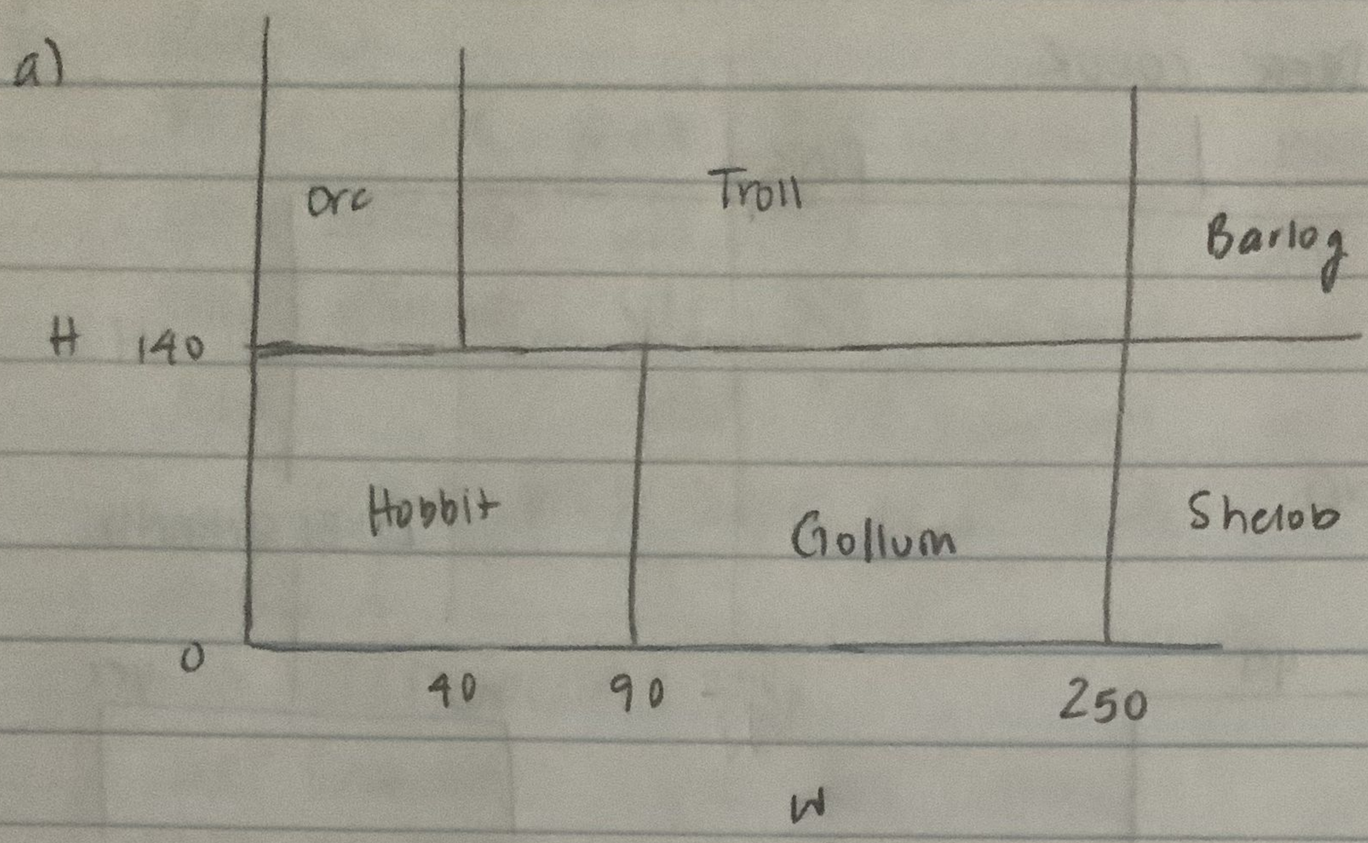
same as above, select LIKES DANCING as 3rd node

3. a)

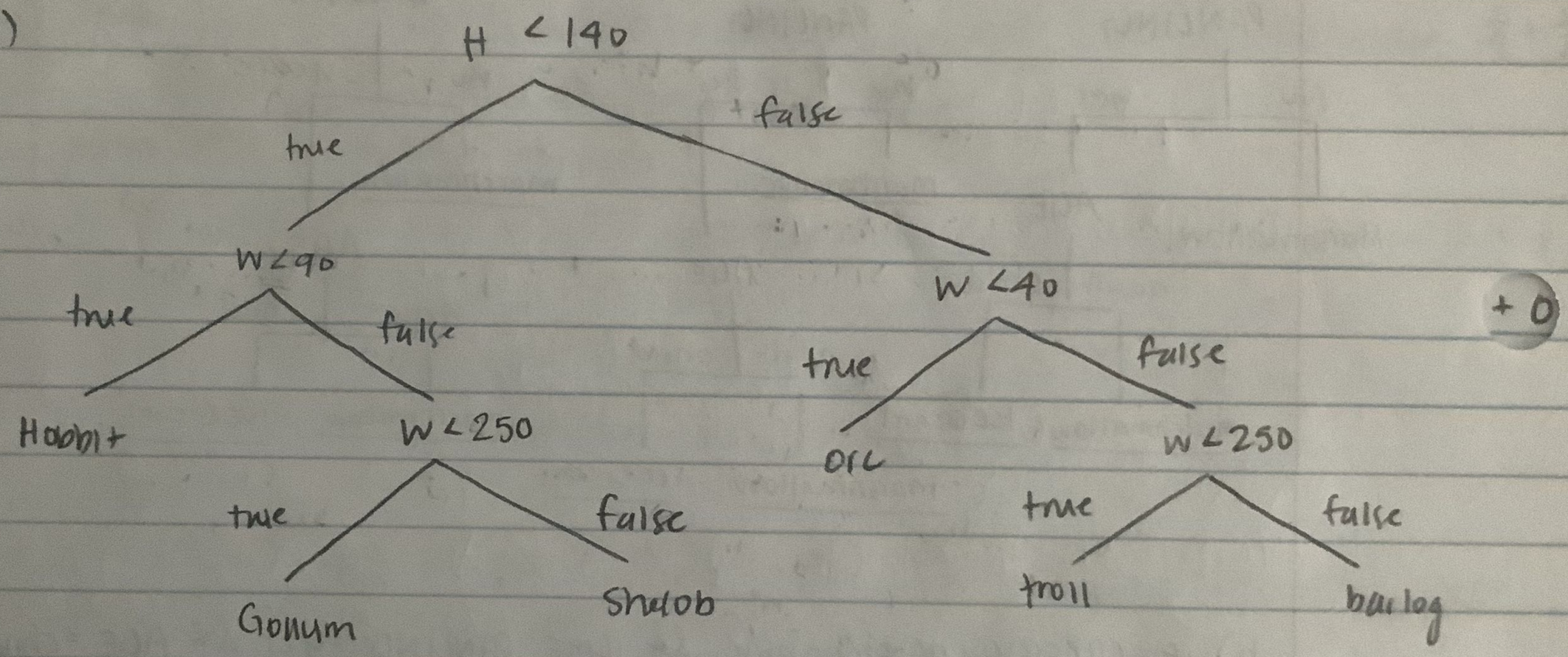


b) IF (DRESS COLOR = non pink && LIKES DANCING = yes && AGE = child
|| DRESS COLOR = pink && LIKES SINGING = yes
|| DRESS COLOR = pink && LIKES DANCING = yes && AGE = child)
THEN DESERT = Ice cream
ELSE
THEN DESERT = marshmallow

4. a)


Orc
Troll
Balrog
H 140
Hobbit
Gollum
Shelob
0
40
90
250
W

b)


H < 140
true
false
W<90
W<40
true
false
true
false
Hobbit
W<250
Orc
W<250
true
false
true
false
Gollum
Shelob
troll
balrog

+0

5. a) use recursive ternary splitting

at each split, top down, make split such that $\{X \mid X_j < s_1\}$, $\{X \mid s_1 \leq x_j < s_2\}$ and $\{X \mid s_2 \leq x\}$ leads to the greatest reduction in RSS where

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 + \sum_{i:x_i \in R_3(j,s)} (y_i - \hat{y}_{R_3})^2$$

where $R_1(j,s) = \{X \mid X_j < s_1\}$

$R_2(j,s) = \{X \mid s_1 \leq X_j < s_2\}$

$R_3(j,s) = \{X \mid s_2 \leq X_j\}$

then split one of the 3 resulting regions, and repeat until stopping criteria.

b) advantages:

- well suited for regions that clearly fit into 3 subregions
- more granularity at each node may lead to a tree of a lesser depth

disadvantages

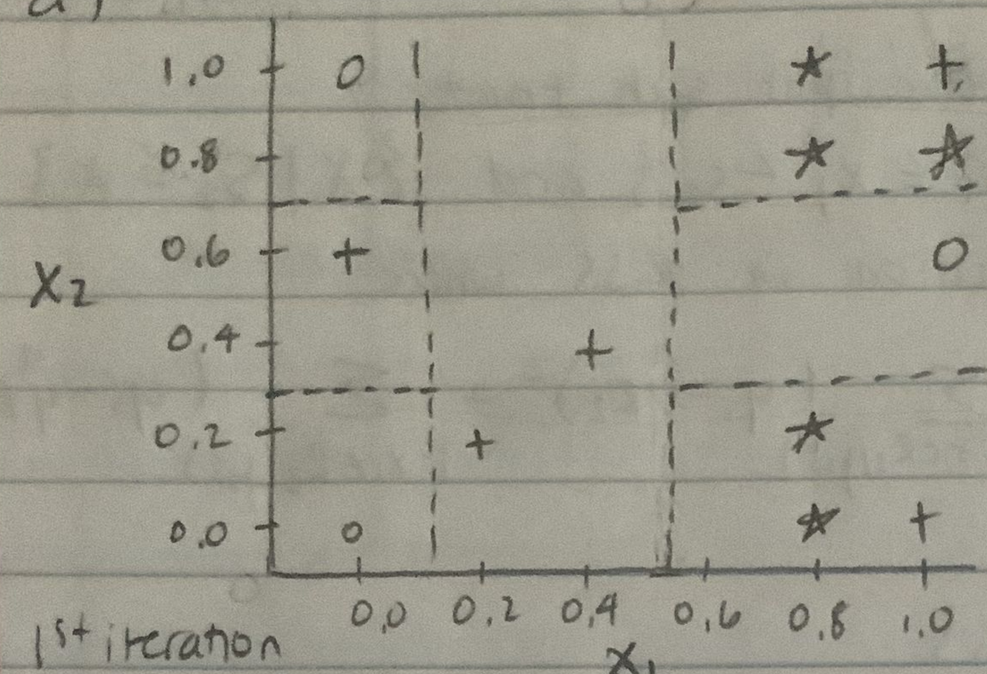
- more computationally difficult to find optimal 3 regions at thresholds s_1 and s_2
- assumes the form that regions split into 3 subregions which won't work well for regions that don't fit assumption

c) you observe the form of the data through graphs to see if regions seem as if they would be better suited to two way splits or three way splits.

Or you can construct both trees and see which produces a model with lower error rates.

since binary splitting is more computationally advantageous. prefer binary splitting if you need better computation efficiency.

d)



1st iteration

$$E = 1 - \max_k (\hat{p}_{mk})$$

$\hat{p}_{mk}$ = proportion of training obs in m^{th} region that are from k^{th} class

X_1: $s_1 = 0.1$ $s_2 = 0.5$ $E = (1 - \frac{2}{3}) + 0 + (1 - \frac{5}{8}) = \frac{1}{3} + \frac{3}{8} = 0.70833$

X_1: $s_1 = 0.5$, $s_2 = 0.9$ $E = (1 - \frac{3}{5}) + 0 + (1 - \frac{1}{2}) = \frac{2}{5} + 0.5 = 0.9$

So split X_1 at $s_1 = 0.1$, $s_2 = 0.5$

2nd iteration: splitting leftmost region, $X_1 < 0.1$

X_2: $s_1 = 0.7$, $s_2 = 0.3$, $E = 0$

3rd iteration: splitting rightmost region $X_1 > 0.5$

X_2: $s_1 = 0.7$, $s_2 = 0.3$ $E = (1 - \frac{3}{4}) + 0 + (1 - \frac{3}{4}) = \frac{1}{4} + \frac{1}{4} = 0.5$

So result:

if $X_1 < 0.1$:
  if $X_2 < 0.3$ classify to "O"
  if $0.3 \leq X_2 < 0.7$ classify to "+"
  if $X_2 \geq 0.7$ classify to "O"
if $0.1 \leq X_1 < 0.5$ classify to "+"
if $0.5 \leq X_1$:
  if $X_2 < 0.3$ classify to "★"
  if $0.3 \leq X_2 < 0.7$ classify to "O"
  if $0.7 \leq X_2$ classify to "★"

6. $Pr(Y=1|X)$: 0.09, 0.13, 0.14, 0.19, 0.53, 0.59, 0.6, 0.63, 0.71

majority polling: Y=0 Y=0 Y=0 Y=0 Y=1 Y=1 Y=1 Y=1 Y=1

since #Y=1's = 6, final classification would be $\boxed{Y=1}$
which is the majority since n=9

average probability: $\frac{0.09+0.13+0.19+0.19+0.53+0.54+0.6+0.63+0.71}{9}$

$= 0.40667$

since average < 0.5, would classify as not $Y=1$,
so $\boxed{Y=0}$