1a) $P(Y_1=1) = P(Y=k)\, e^{\beta_1 x_1}$

$$(K)(P(Y=K))(e^{\beta_1 x_i} + \dots + e^{\beta_K x_i}) = 1$$

$$P(Y=K) = \frac{1}{K(e^{\beta_1 x_i} + \dots + e^{\beta_K x_i})} = \frac{1}{K\sum_{i=1}^{k} e^{\beta_i x_i}}$$

$$1 = \sum_{n=1}^{K} P(Y_i = n)$$

$$P(Y_i = K) + \sum_{n=1}^{K-1} P(Y_i = K)\, e^{\beta_n x_i} = 1$$

$$1 = P(Y_i = K)\left(1 + \sum_{n=1}^{K-1} e^{\beta_n x_i}\right)$$

$$P(Y_i = K) = \frac{1}{1 + \sum_{n=1}^{K-1} e^{\beta_j x_i}}$$

1b) Pred. for x_i changes from k to r when $P(Y_i=r) > P(Y_i=k)$

$$P(Y_i = r) = P(Y_i = k)\, e^{\beta_r x_i}$$

$$\frac{e^{\beta_r x_i}}{\sum_{j=1}^{K} e^{x_i \beta_j}} > \frac{e^{\beta_k x_i}}{\sum_{j=1}^{K} e^{x_i \beta_j}}$$

$$e^{\beta_r x_i} > e^{\beta_K x_i}$$

$$\beta_r x_i > \beta_K x_i$$

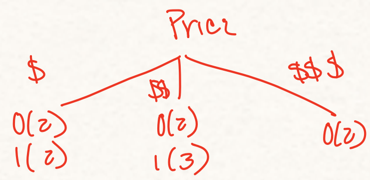
$$x_i(\beta_r - \beta_K) >$$

2a) Type: Basket, Football, Tennis, Tennis, Basketball
Price: \$, \$\$, \$, \$\$, \$\$
location: LA, B, B, LB, LB
Level: I, B, B, I, A

Entropy: 0.994

Feature	Information Gain
Price	$\frac{1}{11}\left[\left(\left(-\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log\frac{2}{5}\right)\cdot 4\right)+\left(5\cdot\left(-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}\right)\right)\right] = 0.8076$
Type	$\frac{1}{11}\left[\left(3\cdot\left(-\frac{1}{3}\log\frac{1}{3}-\frac{2}{3}\log\frac{2}{3}\right)\right)+\left(3\cdot\left(-\frac{2}{3}\log\frac{2}{3}-\frac{1}{3}\log\frac{1}{3}\right)\right)+\left(5\cdot\left(-\frac{2}{5}\log\frac{2}{5}-\frac{3}{5}\log\frac{3}{5}\right)\right)\right] = 0.9422$
Location	$\frac{1}{11}\left[\left(3\cdot\left(-\frac{1}{3}\log\frac{1}{3}-\frac{2}{3}\log\frac{2}{3}\right)\right)+\left(4\cdot\left(\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}\right)\right)+\left(4\cdot\left(\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}\right)\right)\right] = 0.9777$
Level	$\frac{1}{11}\left[\left(4\cdot\left(-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}\right)\right)+\left(3\cdot\left(\frac{1}{3}\log\frac{1}{3}-\frac{2}{3}\log\frac{2}{3}\right)\right)+\left(4\cdot\left(\frac{1}{2}\log\frac{1}{2}\,\frac{1}{2}\log\frac{1}{2}\right)\right)\right] = 0.9777$

Match	Type	Price	Location	Level	OK
R1	Football	\$	LA	Intermediate	0
R2	Tennis	\$\$	Burbank	Advanced	0
R3	Basketball	\$\$	Burbank	Beginner	0
R4	Basketball	\$\$\$	Long Beach	Intermediate	0
R5	Basketball	\$	LA	Intermediate	1
R6	Football	\$\$	Burbank	Beginner	1
R7	Tennis	\$	Burbank	Beginner	1
R8	Basketball	\$	Long Beach	Advanced	0
R9	Football	\$\$\$	LA	Beginner	0
R10	Tennis	\$\$	Long Beach	Intermediate	1
R11	Basketball	\$\$	Long Beach	Advanced	1

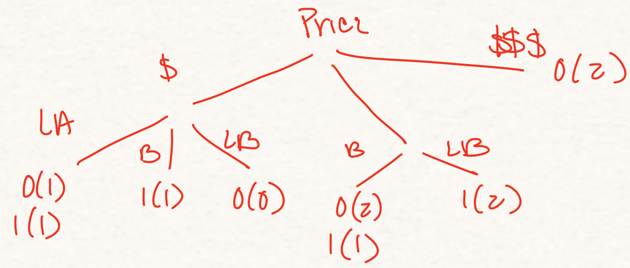


split on price (most info gain)

	Branch	Feature	Info Gain
E=1	\$	Type	$\frac{1}{4}[0+2+0]=0.5$
	\$	location	$\frac{1}{4}[2+0+0]=0.5$
	\$	Level	$\frac{1}{4}[2+0+0]=0.5$
E=0.971	\$\$	Type	$\frac{1}{5}[2\cdot 1+2\cdot 1+0]=\frac{4}{5}$
	\$\$	location	$\frac{1}{5}\left[\left(3\cdot -\frac{1}{3}\log\frac{1}{3}-\frac{2}{3}\log\frac{2}{3}\right)+0+0\right]=0.551$
	\$\$	Level	$\frac{1}{5}[2+2+0]=\frac{4}{5}$

split on location

Match	Type	Price	Location	Level	OK
R1	Football	$	LA	Intermediate	0
R2	Tennis	$$	Burbank	Advanced	0
R3	Basketball	$$	Burbank	Beginner	0
R4	Basketball	$$$	Long Beach	Intermediate	0
R5	Basketball	$	LA	Intermediate	1
R6	Football	$$	Burbank	Beginner	1
R7	Tennis	$	Burbank	Beginner	1
R8	Basketball	$	Long Beach	Advanced	0
R9	Football	$$$	LA	Beginner	0
R10	Tennis	$$	Long Beach	Intermediate	1
R11	Basketball	$$	Long Beach	Advanced	1



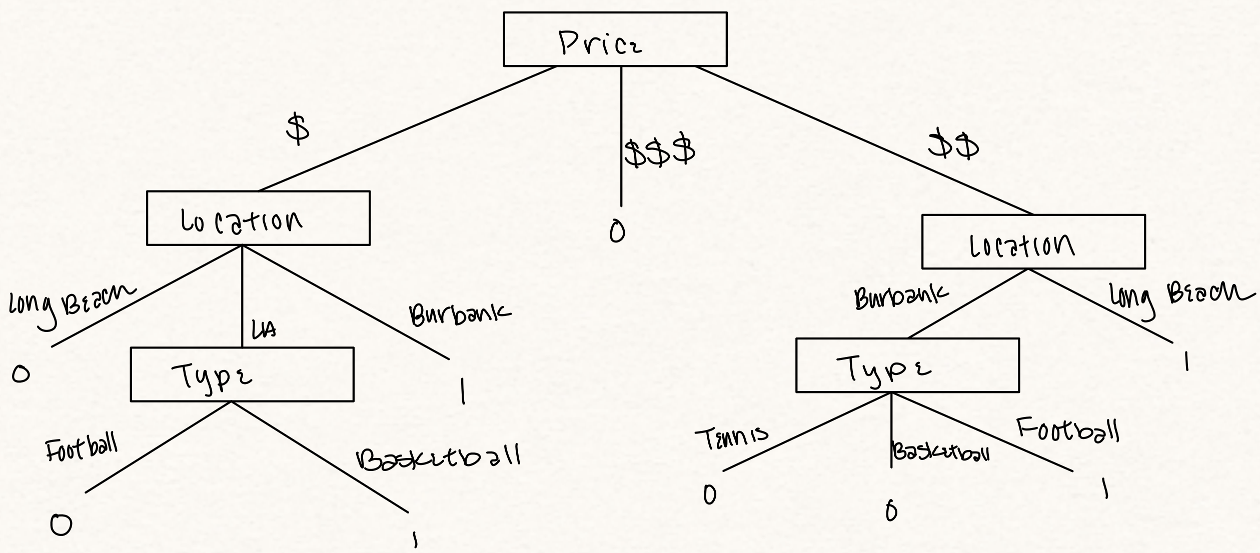
Split on LA

Fea	Info Gain
Type	$\frac{1}{2}[0 + 1\log 1 + 0\log 0] = 0$
Level	$\frac{1}{2}[(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2})\cdot 2] = 1$

We split on type

Split on B

Fea	Info Gain
Type	0
Level	$\frac{1}{3}[0 + (-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2})\cdot 2] = \frac{2}{3}$



2b) Our error is $0/11 = 0$; it classifies all our training data correctly

2c) R12, R13, R16

2d)

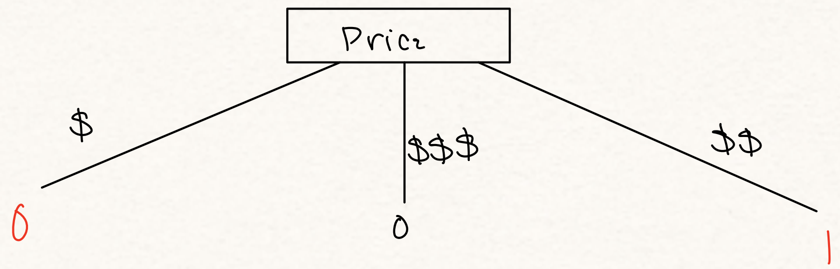
Match	OK	Pred
R12	0	1
R13	1	1
R14	0	0
R15	1	0
R16	0	1

$3/5$ error, F1: $\frac{1}{1+\frac{1}{2}(2+1)} = 0.4$

bad, misclassified over half the points

2e)

Tree	Error	Leaves	$\frac{Error(T_0) - Error(T_1)}{\lvert T_0\rvert - \lvert T_1\rvert}$
Full	0.6	9	—
LA/Type	0.6	8	0
Burbank/Type	0.6	6	0
$/location	0.2	4	$\frac{0.4}{2} = 0.2$
$$/location	0.2	3	0
Price	0.4	1	-0.1



3a) It's more likely to overfit on the non-linearly separable data. This is b/c not being linearly separable can be indicative of many features that differentiate the classes. The tree would try to account for all of these & overfit. I would use a logistic regression model

3b) There is no need to, as decision trees aren't sensitive to the magnitude of variables

3c) Yes, because they split on certain values. E.g., if you have [1, 2, 3, 4, 100000], & you split on 2.5, the 10000 doesn't matter as much.

4a) $R^2 = 0.661$

4b) I would increase the depth & inc. number of features we split on.
max_depth = 12, max_features = 5 gives us $R^2 = 0.91$

4c) python

4d) python
The feature importance for the two classifiers are nearly identical.

5a) $\Delta w_i = c(t-z) * x_i$

$w_1 = w_2 = w_c = 1$, $c = 2$

Point	Pred	Δw_1	Δw_2	Δw_c
(2,-3)	0	4 => 5	-6 => -5	2 => 3
(4,4)	1	-8 => -3	-8 => -13	-6 => -3
(2,-3)	1	-4 => -7	6 => -5	6 => 3

$(2)(1)+(-3)(1)+1 = 0$

$(4)(5)+(4)(-5)+3 = 3 \Rightarrow 1$

$(2)(-3)+(-3)(-13)-3 = 30 \Rightarrow 1$

Point	Pred	Δw_1	Δw_2	Δw_c
(2,-3)	1	—	—	—
(4,4)	0	—	—	—
(2,-3)	1	-4 => -11	6 => 1	-6 => -3

$(2)(-7)+(-3)(-5)+3 = 4 \Rightarrow 1$

$(4)(-7)+(4)(-5)+3 = -45 \Rightarrow 0$

$(2)(-7)+(-3)(-5)+3 = 3 \Rightarrow 1$

5b) It hasn't converged, b/c it is still misclassifying points. In this case, I don't expect it to converge, b/c (2,-3) has both outputs 0 and 1.

6a) $w_1 = w_2 = -1,\ w_0 = 1$

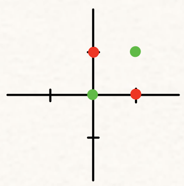
$(0)(-1) + (0)(-1) + 1 = 1 \Rightarrow +1$

$(0)(-1) + (1)(-1) + 1 = 0 \Rightarrow +1$

$(1)(-1) + (0)(-1) + 1 = 0 \Rightarrow +1$

$(1)(-1) + (1)(-1) + 1 = -1 \Rightarrow -1$

With $w_1 = w_2 = -1$ and $w_0 = 1$, the dataset can be perfectly classified using a single activation unit

6b) 

As we can see from plotting the points, the data is not linearly separable, thus we will n

6c) Yes, you should be able to, by first separating (1,1) from the other three, then separating (0,0) from the others. These two activation units & their weights can be combined to form a neural network

7a) 4 - linear, smaller C is closer dist

7b) 3 - linear, larger C is further points from line

7c) 5 - quadratic

7d) 1 - smaller σ, tighter boundary

7e) 6 - larger σ, loose boundary