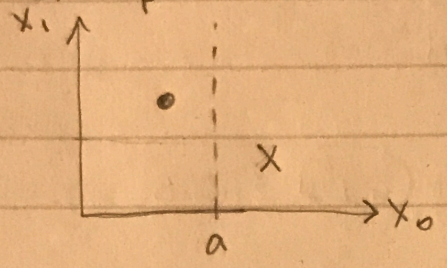
Written Exercises:

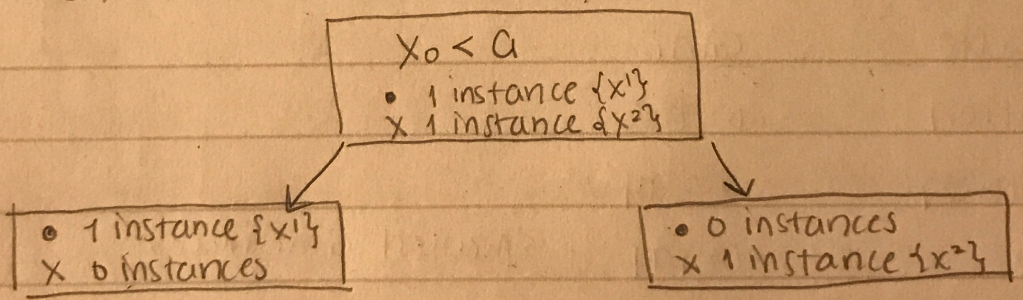
a) Decision trees

ai A) 2D rectangular input space containing points x^1 and x^2. Is it possible to create an example s.t. x^1 and x^2 are <u>not</u> separable by decision stump?

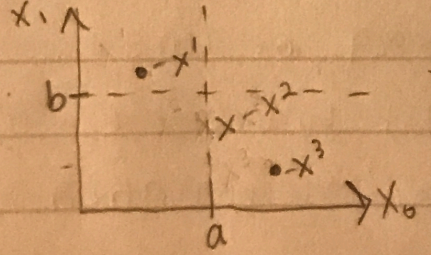


say x^1 classified as •
x^2 classified as ×
and $x^1 = (x^1_0, x^1_1)$
$x^2 = (x^2_0, x^2_1)$

→ No, it is not possible. x^1 and x^2 can always be separable by decision stump, since they are defined on 2-dimensions (2 attributes) and decision stump splits only based on one of them at a time. For example,

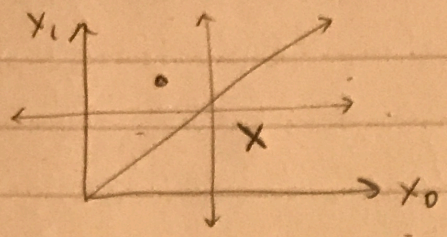


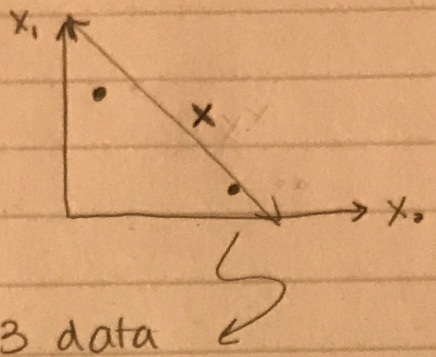
→ Adding another datapoint x^3 does change our answer. For example,



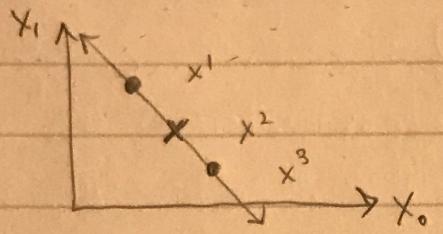
This decision stump (say on the condition $x_0 < a$ again) cannot separate x^1, x^2, x^3 correctly. Same goes for using the other attribute (say $x_1 < b$).

→ In both these cases, you can separate the dataset with a linear classifier.



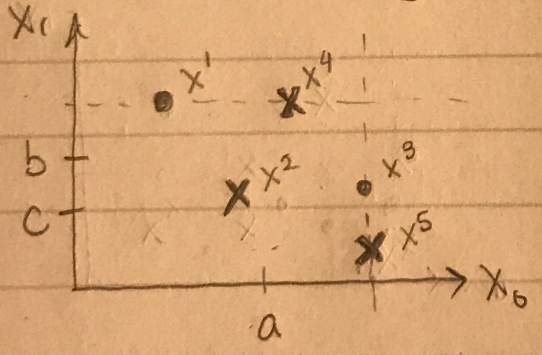


Although in the case of 3 data points, not always. say x^1, x^2, and x^3 all lie on the same line:



There is no linear classifier to correctly separate x^1 and x^3 from x^2.

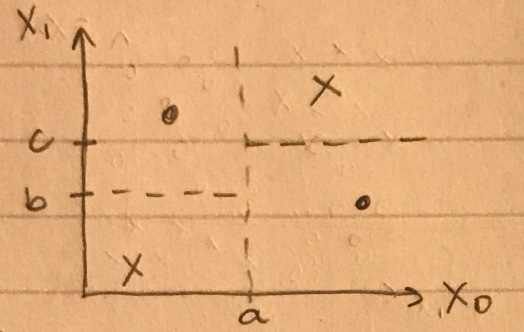
aiB) Is it possible to add more datapoints such that they are not separable by a two-level decision tree?



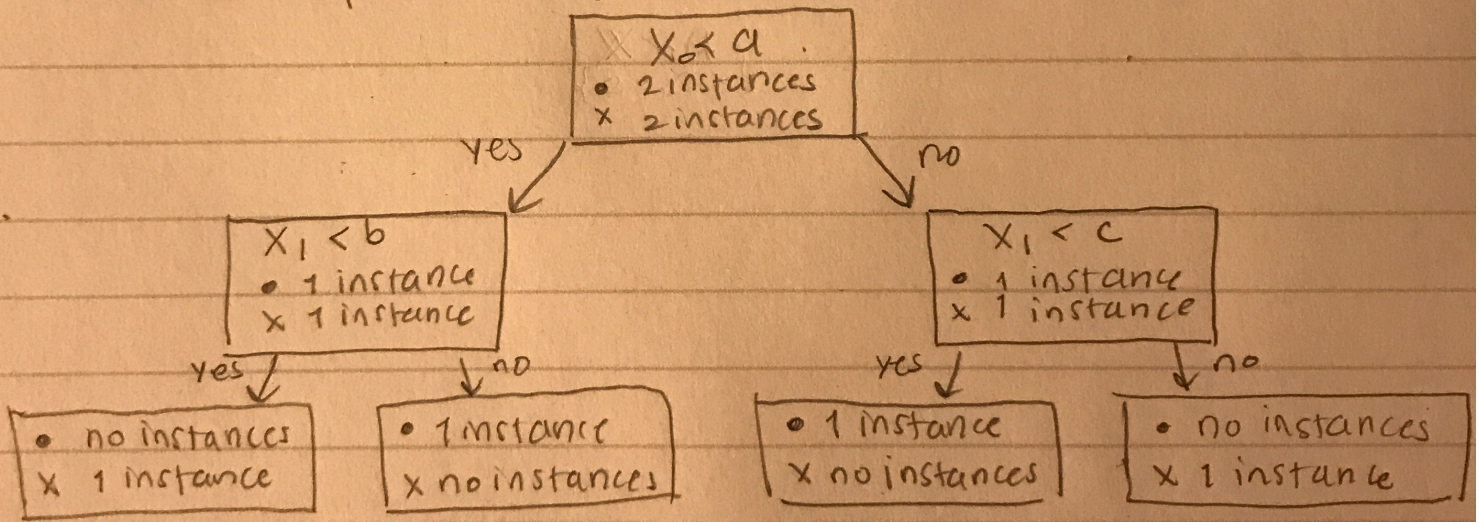
→ Yes, say we have 5 data points as shown above, where x^1 and x^4 have the same X_1 value, and x^3 and x^5 have the same X_0 value.
There is no way to separate on some $X_0 < a$ condition and then $X_1 < b$ and $X_1 < c$ conditions or vice versa (starting with X_1 attribute)

(c) is it possible to construct an example s.t. the data points are separable by a 2-level decision tree but not by a linear classifier?

→ yes, decision trees can fit linearly inseparable data sets. For example, this data set is not separable by a linear classifier.



It is separable by a 2-level decision tree:



This linearly inseparable dataset is separable by a 2-level decision tree.

a ii.) Entropy: $E = -\sum_j p_j \log_2 p_j$

Cross-entropy: $E(T,X) = \sum_{c \in X} P(c) E(c)$

location:	preference: yes:	no:
rural	2	2
urban	2	2
semi-rural	2	1

$E(\text{preference, location}) = P(\text{rural}) E(2,2) + P(\text{urban}) E(2,2) + P(\text{semi-rural}) E(2,1)$

$= \left(\frac{4}{11}\right)\left[-\left(0.5 \log_2(0.5) + 0.5 \log_2(0.5)\right)\right] + \left(\frac{4}{11}\right) + \left(\frac{3}{11}\right)\left[-\left(\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right)\right]$

$= \frac{8}{11} + \left(\frac{3}{11}\right)\left[+(+0.39 + 0.53)\right] = 0.978$

Doing the same calculations for each attribute (in Excel):

	Attribute	Entropy:	info gain (E(T) − E(T,X))
E(T,X)	Location	0.978	0.994 − 0.978 = 0.016
	pollution	0.840	0.153
	Area	0.987	0.007
	windows	0.829	0.165 ← largest gain
E(T)	preference	0.994	

a ii)

Gini index = $1 - \sum_j (p_j^2)$

Location	Yes (preference)	no	total
rural	2	2	4
urban	2	2	4
semi-rural	2	1	3
total:	6	5	11

$P(\text{Location}=\text{rural}) = \frac{4}{11}$

$P(\text{LOC}=\text{urban}) = \frac{4}{11}$

$P(\text{LOC}=\text{semi}) = \frac{3}{11}$

rural : $\text{Gini} = 1 - [(\frac{2}{4})^2 + (\frac{2}{4})^2] = 0.5$

urban : $\text{Gini} = 0.5$

semi : $\text{Gini} = 1 - [(\frac{2}{3})^2 + (\frac{1}{3})^2] = 0.444$

→ weighted Gini index for attribute Location:

$\text{Gini} = (\frac{4}{11})(0.5) + (\frac{4}{11})(0.5) + (\frac{3}{11})(0.444) = 0.485$

computing for all attributes in Excel:

Attribute	Gini
Location	0.485
pollution	0.394
Area	0.491
windows	0.388 ← smallest Gini index
preference	$1 - [(6/11)^2 + (5/11)^2] = 0.496$

conclude: windows should be the root of this decision tree. It has the smallest Gini index and cross-entropy, and the largest information gain.

Entropy Analysis:													
location	yes	no		P(attribute)	p(yes)	log2(p)	p*log2(p)	p(no)	log2(p)	p*log2(p)	E(yes,no)	cross-entropy	info gain
rural	2	2	4	0.364	0.500	-1.000	-0.500	0.500	-1.000	-0.500	1.000	0.364	
urban	2	2	4	0.364	0.500	-1.000	-0.500	0.500	-1.000	-0.500	1.000	0.364	
semi-rural	2	1	3	0.273	0.667	-0.585	-0.390	0.333	-1.585	-0.528	0.918	0.250	
			11									0.978	0.016
pollution	yes	no		P(attribute)	p(yes)	log2(p)	p*log2(p)	p(no)	log2(p)	p*log2(p)	E(yes,no)	cross-entropy	
low	2	1	3	0.273	0.667	-0.585	-0.390	0.333	-1.585	-0.528	0.918	0.250	
med	3	1	4	0.364	0.750	-0.415	-0.311	0.250	-2.000	-0.500	0.811	0.295	
high	1	3	4	0.364	0.250	-2.000	-0.500	0.750	-0.415	-0.311	0.811	0.295	
			11									0.840	0.154
area	yes	no		P(attribute)	p(yes)	log2(p)	p*log2(p)	p(no)	log2(p)	p*log2(p)	E(yes,no)	cross-entropy	
small	3	3	6	0.545	0.500	-1.000	-0.500	0.500	-1.000	-0.500	1.000	0.545	
large	3	2	5	0.455	0.600	-0.737	-0.442	0.400	-1.322	-0.529	0.971	0.441	
			11									0.987	0.007
windows	yes	no		P(attribute)	p(yes)	log2(p)	p*log2(p)	p(no)	log2(p)	p*log2(p)	E(yes,no)	cross-entropy	
small	2	4	6	0.545	0.333	-1.585	-0.528	0.667	-0.585	-0.390	0.918	0.501	
large	4	1	5	0.455	0.800	-0.322	-0.258	0.200	-2.322	-0.464	0.722	0.328	
			11									0.829	0.165
preference	yes	no			p(yes)	log2(p)	p*log2(p)	p(no)	log2(p)	p*log2(p)	E(yes,no)		
	6	5	11		0.545	-0.874	-0.477	0.455	-1.138	-0.517	0.994		

Entropy Analysis:													
location	yes	no		P(attribute)	p(yes)	log2(p)	p*log2(p)	p(no)	log2(p)	p*log2(p)	E(yes,no)	cross-entropy	info gain
rural	2	2	4	=D3/D$6	=B3/D3	=LOG(F3,2)	=F3*G3	=C3/D3	=LOG(I3,2)	=I3*J3	=-SUM(H3,K3)	=E3*L3	
urban	2	2	4	=D4/D$6	=B4/D4	=LOG(F4,2)	=F4*G4	=C4/D4	=LOG(I4,2)	=I4*J4	=-SUM(H4,K4)	=E4*L4	
semi-rural	2	1	3	=D5/D$6	=B5/D5	=LOG(F5,2)	=F5*G5	=C5/D5	=LOG(I5,2)	=I5*J5	=-SUM(H5,K5)	=E5*L5	
			11									=SUM(M3:M5)	=$L$25-M6
pollution	yes	no		P(attribute)	p(yes)	log2(p)	p*log2(p)	p(no)	log2(p)	p*log2(p)	E(yes,no)	cross-entropy	
low	2	1	=SUM(B9:C9)	=D9/D$12	=B9/D9	=LOG(F9,2)	=F9*G9	=C9/D9	=LOG(I9,2)	=I9*J9	=-SUM(H9,K9)	=E9*L9	
med	3	1	=SUM(B10:C10)	=D10/D$12	=B10/D10	=LOG(F10,2)	=F10*G10	=C10/D10	=LOG(I10,2)	=I10*J10	=-SUM(H10,K10)	=E10*L10	
high	1	3	=SUM(B11:C11)	=D11/D$12	=B11/D11	=LOG(F11,2)	=F11*G11	=C11/D11	=LOG(I11,2)	=I11*J11	=-SUM(H11,K11)	=E11*L11	
			=SUM(D9:D11)									=SUM(M9:M11)	=$L$25-M12
area	yes	no		P(attribute)	p(yes)	log2(p)	p*log2(p)	p(no)	log2(p)	p*log2(p)	E(yes,no)	cross-entropy	
small	3	3	=SUM(B15:C15)	=D15/D$17	=B15/D15	=LOG(F15,2)	=F15*G15	=C15/D15	=LOG(I15,2)	=I15*J15	=-SUM(H15,K15)	=E15*L15	
large	3	2	=SUM(B16:C16)	=D16/D$17	=B16/D16	=LOG(F16,2)	=F16*G16	=C16/D16	=LOG(I16,2)	=I16*J16	=-SUM(H16,K16)	=E16*L16	
			=SUM(D15:D16)									=SUM(M15:M16)	=$L$25-M17
windows	yes	no		P(attribute)	p(yes)	log2(p)	p*log2(p)	p(no)	log2(p)	p*log2(p)	E(yes,no)	cross-entropy	
small	2	4	=SUM(B20:C20)	=D20/D$22	=B20/D20	=LOG(F20,2)	=F20*G20	=C20/D20	=LOG(I20,2)	=I20*J20	=-SUM(H20,K20)	=E20*L20	
large	4	1	=SUM(B21:C21)	=D21/D$22	=B21/D21	=LOG(F21,2)	=F21*G21	=C21/D21	=LOG(I21,2)	=I21*J21	=-SUM(H21,K21)	=E21*L21	
			=SUM(D20:D21)									=SUM(M20:M21)	=$L$25-M22
preference	yes	no			p(yes)	log2(p)	p*log2(p)	p(no)	log2(p)	p*log2(p)	E(yes,no)		
	6	5	=SUM(B25:C25)		=B25/D25	=LOG(F25,2)	=F25*G25	=C25/D25	=LOG(I25,2)	=I25*J25	=-SUM(H25,K25)		

Gini Index Analysis:									
location	yes	no		P(loc)	P(yes)	P(no)	P(yes)^2	P(no)^2	Gini
rural	2	2	4	0.364	0.500	0.500	0.250	0.250	0.500
urban	2	2	4	0.364	0.500	0.500	0.250	0.250	0.500
semi-rural	2	1	3	0.273	0.667	0.333	0.444	0.111	0.444
	6	5	11						0.485
pollution	yes	no		P(pollution)	P(yes)	P(no)	P(yes)^2	P(no)^2	Gini
low	2	1	3	0.273	0.667	0.333	0.444	0.111	0.444
med	3	1	4	0.364	0.750	0.250	0.563	0.063	0.375
high	1	3	4	0.364	0.250	0.750	0.063	0.563	0.375
	6	5	11						0.394
area	yes	no		P(area)	P(yes)	P(no)	P(yes)^2	P(no)^2	Gini
small	3	3	6	0.545	0.500	0.500	0.250	0.250	0.500
large	3	2	5	0.455	0.600	0.400	0.360	0.160	0.480
	6	5	11						
									0.491
windows	yes	no		P(windows)	P(yes)	P(no)	P(yes)^2	P(no)^2	Gini
small	2	4	6	0.545	0.333	0.667	0.111	0.444	0.444
large	4	1	5	0.455	0.800	0.200	0.640	0.040	0.320
	6	5	11						
									0.388
preference	yes	no		P(pref)	P(yes)	P(no)	P(yes)^2	P(no)^2	Gini
	6	5	11	1.000	0.545	0.455	0.298	0.207	0.496

Gini Index Analysis:									
location	yes	no		P(loc)	P(yes)	P(no)	P(yes)^2	P(no)^2	Gini
rural	2	2	=SUM(B3:C3)	=D3/D$6	=B3/$D3	=C3/$D3	=F3^2	=G3^2	=1-SUM(H3:I3)
urban	2	2	=SUM(B4:C4)	=D4/D$6	=B4/$D4	=C4/$D4	=F4^2	=G4^2	=1-SUM(H4:I4)
semi-rural	2	1	=SUM(B5:C5)	=D5/D$6	=B5/$D5	=C5/$D5	=F5^2	=G5^2	=1-SUM(H5:I5)
	=SUM(B3:B5)	=SUM(C3:C5)	=SUM(D3:D5)						=SUMPRODUCT(E3:E5,J3:J5)
pollution	yes	no		P(pollution)	P(yes)	P(no)	P(yes)^2	P(no)^2	Gini
low	2	1	=SUM(B10:C10)	=D10/D$6	=B10/$D10	=C10/$D10	=F10^2	=G10^2	=1-SUM(H10:I10)
med	3	1	=SUM(B11:C11)	=D11/D$6	=B11/$D11	=C11/$D11	=F11^2	=G11^2	=1-SUM(H11:I11)
high	1	3	=SUM(B12:C12)	=D12/D$6	=B12/$D12	=C12/$D12	=F12^2	=G12^2	=1-SUM(H12:I12)
	=SUM(B10:B12)	=SUM(C10:C12)	=SUM(D10:D12)						=SUMPRODUCT(E10:E12,J10:J12)
area	yes	no		P(area)	P(yes)	P(no)	P(yes)^2	P(no)^2	Gini
small	3	3	=SUM(B17:C17)	=D17/D$6	=B17/$D17	=C17/$D17	=F17^2	=G17^2	=1-SUM(H17:I17)
large	3	2	=SUM(B18:C18)	=D18/D$6	=B18/$D18	=C18/$D18	=F18^2	=G18^2	=1-SUM(H18:I18)
	=SUM(B17:B18)	=SUM(C17:C18)	=SUM(D17:D18)						
									=SUMPRODUCT(E17:E19,J17:J19)
windows	yes	no		P(windows)	P(yes)	P(no)	P(yes)^2	P(no)^2	Gini
small	2	4	=SUM(B23:C23)	=D23/D$6	=B23/$D23	=C23/$D23	=F23^2	=G23^2	=1-SUM(H23:I23)
large	4	1	=SUM(B24:C24)	=D24/D$6	=B24/$D24	=C24/$D24	=F24^2	=G24^2	=1-SUM(H24:I24)
	=SUM(B23:B24)	=SUM(C23:C24)	=SUM(D23:D24)						
									=SUMPRODUCT(E23:E25,J23:J25)
preference	yes	no		P(pref)	P(yes)	P(no)	P(yes)^2	P(no)^2	Gini
	6	5	=SUM(B29:C29)	=D29/D$6	=B29/$D29	=C29/$D29	=F29^2	=G29^2	=1-SUM(H29:I29)