Data at Scale HW3 — Lorenzo Scaturchio

1. $\text{Entropy}(V_1) = E(3,3) = -\left(\frac{3}{6}\log_2\left(\frac{3}{6}\right) + \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right) = 1$ $\quad E\,V_1\text{yes}(3,0) = 0 \quad E\,V_1\text{no}(1,-2) = 0.92$

$\text{Entropy}(V_2) = E(4,2) = -\left(\frac{4}{6}\log_2\left(\frac{4}{6}\right) + \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right) = 0.92$ $\quad E\,V_2\text{yes}(2,2) = 1 \quad E\,V_2\text{no}(2,0) = 0$

$\text{Entropy}(V_3) = E(4,2) = -\left(\frac{4}{6}\log_2\left(\frac{4}{6}\right) + \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right) = 0.92.$ $\quad E\,V_3\text{yes}(3,1) = 0.81 \quad E\,V_3\text{no}(1,-1) = 1$

$\text{Entropy}(V_4) = (4,2) = -\left(\frac{4}{6}\log_2\left(\frac{4}{6}\right) + \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right) = 0.92$ $\quad E\,V_4\text{yes}(2,2) = 1 \quad E\,V_4\text{no}(2,0) = 0$

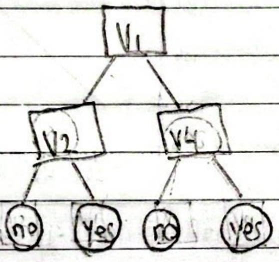
$\text{Entropy}(C) = (4,2) = -\left(\frac{4}{6}\log_2\left(\frac{4}{6}\right) + \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right) = 0.92$

$G(S,C) = \text{Entropy}(V_1) - (\text{Entropy}(V_1\text{yes}) + \text{Entropy}(V_1\text{no})) \rightarrow V_1 = 1 - \left(\frac{3}{6}(0) + \frac{3}{6}(0.92)\right) = \boxed{0.46}$

$G(S,C) = \text{Entropy}(V_2) - (\text{Entropy}(V_2\text{yes}) + \text{Entropy}(V_2\text{no})) \rightarrow V_2 = 0.92 - \left(\frac{4}{6}(1) + 0\right) = \boxed{0.25}$

$G(S,C) = \text{Entropy}(V_3) - (\text{Entropy}(V_3\text{yes}) + \text{Entropy}(V_3\text{no})) \rightarrow V_3 = 0.92 - \left(\frac{4}{6}(0.81) + \frac{2}{6}(1)\right) = \boxed{0.04}$

$G(S,C) = \text{Entropy}(V_4) - (\text{Entropy}(V_1\text{yes}) + \text{Entropy}(V_1\text{no})) \rightarrow V_4 = 0.92 - \left(\frac{4}{6}(1) + 0\right) = \boxed{0.25}$



2. $P(A>50) = 0.01$, $P(A>50^c) = 0.99 \rightarrow P(X|A) = 0.9$ and $P(X|A^c) = 0.8 \rightarrow P(A|X) = \frac{P(A\cap X)}{P(A\cap X) + P(A^c \cap X)} \rightarrow \frac{P(A)\cdot P(X|A)}{P(A)P(X|A) + P(A^c) P(X|A^c)} \rightarrow \frac{0.01\times 0.9}{(0.01\times 0.9) + (0.99\times 0.8)} \rightarrow \frac{0.009}{0.009 + 0.792} \rightarrow \frac{0.009}{0.801} = 11.23\%$ $\therefore \boxed{P = 0.11}$

3.

a. $P(\text{Kim}|\text{Name}) = 0.33$, $P(\text{female}|\text{Gender}) = 0.66$, $P(\text{male}|\text{Gender}) = 0.33 \rightarrow P(\text{male}|\text{Kim}) = \frac{P(\text{male}\cap\text{kim})}{P(\text{male}\cap\text{kim}) + P(\text{male}^c \cap \text{kim})} \rightarrow \frac{P(\text{male})\cdot P(\text{kim}|\text{male})}{P(\text{male})P(\text{kim}|\text{male}) + P(\text{male}^c)} \rightarrow \frac{0.33\cdot 0.33}{(0.33\cdot 0.33) + (0.5\cdot 0.68)} = \boxed{P(\text{male}|\text{kim}) = 0.25}$

$P(\text{Female}|\text{Kim}) = \frac{P(\text{Female}\cap\text{Kim})}{P(\text{female}\cap\text{kim}) + P(\text{male}^c \cap \text{kim})} \rightarrow \frac{P(\text{female})\cdot P(\text{Kim}|\text{female})}{P(\text{female})\cdot P(\text{Kim}|\text{female}) + P(\text{female}^c)\cdot P(\text{Kim}^c|\text{female}^c)} \rightarrow \frac{0.66\cdot 0.5}{(0.66\cdot 0.5) + (0.33\cdot 0.5)} \rightarrow \boxed{P(\text{female}|\text{Kim}) = 0.75}$

b. $P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$

Attrib	No	Yes	Prob
over 170	4	4	0.5
long hair	2	6	0.25
blue eyes	3	5	0.38
male	5	3	0.63
Kim	5	3	0.63

All	No	Yes	Prob
All	19/40	21/40	
	0.47	0.53	

$P(\text{male}\ldots|\text{kim}) = 0.5\times 0.25\times 0.38\times 0.63\times 0.63 = \boxed{0.019}$

$P(\text{female}\ldots|\text{kim}) = 0.5\times 0.25\times 0.38\times 0.37\times 0.63 = \boxed{0.048}$

4.

a. $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ $\quad \beta_1 = \Sigma[(x_i - x)(y_i - y)] / \Sigma[(x_i - x)^2]$ $\quad \beta_0 = y - \beta_1 x$ $\quad$ Sales $= y_i$ $\quad$ AD $= x_i$

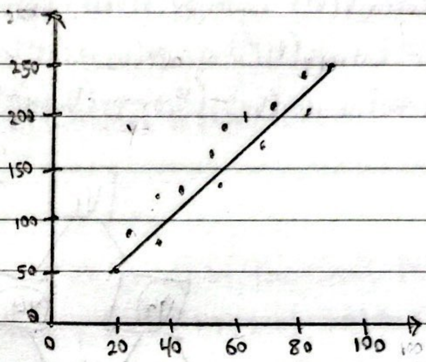
	x	y	x-avg(x)	y-avg(y)	(x-avg(x))²	Sum((col D, (col E))
1	22	64	-22.2	-56.2	492.84	1247.64
2	25	74	-19.2	-46.2	368.84	887.04
3	29	82	-15.2	-38.2	231.04	580.64
4	35	90	-9.2	-30.2	84.64	277.84
5	38	100	-6.2	-20.2	38.44	125.24
6	42	120	-2.2	-0.2	4.84	0.44
7	46	120	1.8	-0.2	3.24	-0.36
8	52	142	7.8	21.8	60.84	170.04
9	65	180	20.8	59.8	432.64	1243.84
10	88	230	43.8	109.8	1918.44	4809.24

$\beta_1 = (9341.6)/(3635.6) = 2.5695$

$\beta_0 = (1202) - (2.5695)(442) = 6.629$

$\boxed{\therefore y = 2.5695x + 6.629}$

b.



c. $r^2 = \dfrac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n \cdot (\Sigma x^2 - (\Sigma x)^2] \times [n \cdot (\Sigma y^2 - (\Sigma y)^2]}} \rightarrow (442 \cdot 1202) - (442)(1202) / \sqrt{10[(442)^2 - 442^2] \cdot 10[(1202)^2 - 1202^2]} = \boxed{0.9927}$

d. $y = 2.5695(50{,}000) + 6.629 = \boxed{\$135{,}104 \text{ prediction}}$

5. Single linkage

	1	(2,3)	4	5	6	7
1	0	0.664	0.962	0.750	1.033	0.25
(2,3)		0	0.605	0.702	0.44	0.58
4			0	0.37	0.39	0.73
5				0	0.7	0.5
6					0	0.86
7						0

→

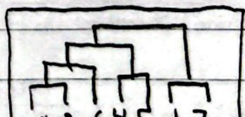
	(1,7)	(2,3)	4	5	6
(1,7)	0	0.58	0.73	0.5	0.87
(2,3)		0	0.61	0.7	0.44
4			0	0.37	0.39
5				0	0.7
6					0

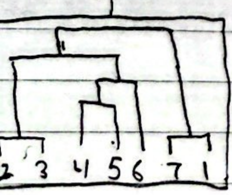
→

	(1,7)	(2,3)	(4,5)	6
(1,7)	0	0.58	0.73	0.87
(2,3)		0	0.61	0.44
(4,5)			0	0.39
6				0

→

	(1,7)	(2,3)	(4,5,6)
(1,7)	0	0.58	0.5
(2,3)		0	0.44
(4,5,6)			0





Complete linkage

	1	(2,3)	4	5	6	7
1	0	0.72	0.96	0.75	1.03	0.25
(2,3)		0	0.71	0.82	0.49	0.6
4			0	0.37	0.39	0.73
5				0	0.7	0.5
6					0	0.87
7						0

→

	(1,7)	(2,3)	4	5	6
(1,7)	0	0.72	0.96	0.75	1.03
(2,3)		0	0.71	0.82	0.49
4			0	0.37	0.39
5				0	0.7
6					0

→

	(1,7)	(2,3)	(4,5)	6
(1,7)	0	0.72	0.96	1.03
(2,3)		0	0.82	0.49
(4,5)			0	0.7
6				0

→

	(1,7)	(2,3)	(4,5,6)
(1,7)	0	1.033	0.96
(2,3)		0	0.82
(4,5,6)			0

Topic