SEE Regression problems

* $y = mx + c$

$$c = \frac{\Sigma y(\Sigma x^2) - \Sigma x \Sigma xy}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$m = \frac{n * \Sigma xy - \Sigma x \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2}$$

	x	y	xy	x^2	Predicted	Error
	1	2	2	1	2.8	0.8
	2	4	8	4	3.4	0.6
	3	5	15	9	4	1
	4	4	16	16	4.6	0.6
	5	5	25	25	5.2	0.2
Σ	15	20	66	55		

$$c = \frac{20 \times 255 - 15 \times 66}{5 \times 255 - 225}$$

$$c = 2.2$$

$$m = \frac{5 \times 66 - 15 \times 20}{5 \times 55 - 225}$$

$$m = 0.6$$

$$Y = 0.6x + 2.2$$

Q. If the weather is sunny, the player should play or not

- Frequency Table for weather Conditions

	Yes	No
Outlook		
Rainy	2	2
Overcast	5	0
Sunny	3	2
Total	10	4

frequency table for all independent variables (here only)

Likelihood Table

Outlook	Probability	Yes	No
Rainy	4/14 = 0.29	2	2
Overcast	5/14 = 0.35	5	0
Sunny	5/14 = 0.35	3	2
All		10/14 = 0.71	4/14 = 0.29

Applying Bayes Theorem

$$P(Yes/Sunny) = \frac{P(Sunny/Yes) * P(Yes)}{P(Sunny)}$$

$$P(Sunny/Yes) = \frac{3}{10} = 0.3$$

$P(Sunny) = 0.35$

$P(Yes) = 0.71$

$$P(Yes/Sunny) = \frac{0.3 * 0.71}{0.35} = 0.608$$

$$P(No/Sunny) = \frac{P(Sunny/No) * P(No)}{P(Sunny)}$$

$$P(Sunny/No) = \frac{2}{4} = \underline{0.5}$$

$P(No) = 0.29$

$P(Sunny) = 0.35$

$$P(No/Sunny) = \frac{0.5 * 0.29}{0.35}$$

$$= \underline{0.41}$$

From the above calculation
$P(Yes/Sunny) > P(No/Sunny)$
Hence on a Sunny day, player can play the game

Q.

No.	Colour	Legs	Height	Smelly	Species
1	White	3	Short	Yes	M
2	Green	2	Tall	No	M
3	Green	3	Short	Yes	M
4	White	3	Short	Yes	M
5	Green	2	Short	No	H
6	White	2	Tall	No	H
7	White	2	Tall	No	H
8	White	2	Short	Yes	H

New instance: If colour = green, legs = 3, height = tall & smelly = No, Identify which species it belongs to

$P(M) = \frac{4}{8} = 0.5$; $P(H) = 0.5$

Likelihood Frequency table for colour

Colour	M (4)	H (4)	Probability
White	2/4	3/4	5/8 = 0.625
Green	2/4	1/4	3/8 = 0.375
	4/8 = 0.5	4/8 = 0.5	

→ for Legs

Legs	M (4)	H (4)	Prob
3	3/4	0	3/8 = 0.375
2	1/4	4/4	5/8 = 0.625
	4/8 = 0.5	4/8 = 0.5	

→ Height

Height	M	H	Prob
Short	3/4	2/4	5/8
Tall	1/4	2/4	3/8

→ Smelly

Smelly	M	H	Prob
Yes	3/4	1/4	
No	1/4	3/4	

0.0117
0.0156

New instance = (colour = green, legs = 2, Height = Tall, Smelly = No)

P(M / New_instance)

$= \frac{P(M) * P(\text{color} = \text{Green} / M) * P(\text{legs} = 2 / M) * P(\text{Height} = \text{Tall} / M) * P(\text{Smelly} = \text{No} / M)}{P(\text{New instance})}$

$= 0.5 * \frac{2}{4} * \frac{1}{4} * \frac{1}{4} * \frac{1}{4} = 0.0039$

P(H / New instance)

$= P(H) * P(\text{Green} / H) * P(2 / H) * P(\text{Tall} / H) P(\text{No} / H)$

$= 0.5 * 0.25 * 1 * 0.5 * 0.5$

$= 0.0468$

$P(H | \text{New instance}) > P(M | \text{New instance})$

$\therefore$ Species = H

* New inst = (colour = green, legs = 2, Height = Tall, Smelly = Yes)

P(M / New inst)

$= P(M) * P(\text{green} / M) * P(2 / M) * P(\text{Tall} / M) * P(\text{Yes} / M)$

$= 0.5 *$

Pls Species = H

Even though species is M in dataset, there is chance to get H in Naive bayes

Q. Cluster Analysis with the help of classifier

V.IMP

Date
Page

* we can't classify based on shape, size, color. eg (apple, orang & ball are all round but don't fall under same cluster

Cluster is formed based on similarity, distance & centroid

→ Problems with Measuring Distances

→ A more general Formula for Distances

$d_M(x, y) = \sqrt{\sum_{i=1}^{n} d(x_i, y_i)}$

* for continuous attributes $d(x_i, y_i) = (x_i - y_i)^2$
* for discrete

V.IMP Which Cluster Should an Example Belong To?

* $d(x, c_i)$ → distance of x from centroid

eg: $c_1 = (3, 4)$, $c_2 = (4, 6)$ $c_3 = (5, 7)$

$x = (4, 4)$

there are 3 centroids ⇒ there are 3 clusters

$d(x, c_1) = \sqrt{1^2 + 0^2} = 1$

$d(x, c_2) = \sqrt{0^2 + 2^2} = 2$

$d(x, c_3) = \sqrt{10} = 3.$

here the shortest distance is $d(x, c_1)$

∴ x belongs to cluster c_1

* If there are 10 datapoints initially then there are 10 clusters, this can be reduced by using K-means clustering

Q.

Hours Study	Pass(1) / Fail(0)
29	0
15	0
33	1
28	1
39	1

given

Use Logistic regression as classifier to answer the following questions

1. Calculate the probability for the student who studied 33 hrs. given log(odds) = −64 + 2* hours

$p = \frac{1}{1+e^{-x}}$ $\quad s(x) = \frac{1}{1+e^{-x}}$

$x = -64 + 2*33 = -64 + 66 = 2$

$p = \frac{1}{1+e^{-x}} = \underline{0.88}$

There is 88% chance that he will pass

2. At least how many hours student should study that makes he will pass the course with prob of more than 95%.

$p = \frac{1}{1+e^{-x}} = 0.95$

$0.95 * (1 + e^{-x}) = 1$

$0.95 * e^{-z} = 1 - 0.95$

$Z = 2.95$

$\log(\text{odds}) = Z = -64 + 2 * \text{hours}$

$2.94 = -64 + 2 * \text{hours}$

$2 * \text{hours} = 2.94 + 64 = 66.94$

$\text{hours} = \frac{66.94}{2} = 33.47 \text{ hours}$

19-3-24

K-M

*fir

such

clust

*Ite

Example: The table below

	Group-1	Group-2	Group-3
	(2,5)	(4,3)	(1,5)
	(1,4)	(3,1)	(3,1)
	(3,6)	(2,2)	(2,3)
Centroids	(2,5)	(3,4)	(2,3)

* (2,5) (2,5) $d = 0$ → least dist → same cluster
  (2,5) (3,4) $d = \sqrt{1+1} = \sqrt{2}$
  (2,5) (2,3) $d = 2$
* (1,4) (2,5) $d = \sqrt{2}$ => min => right cluster
  (1,4) (3,4) $d = \sqrt{2+0} = \sqrt{2^2} = 2$
  (1,4) (2,3) $d = \sqrt{2}$
* (3,6) (2,5) $d = \sqrt{2}$ => min => right cluster
  (3,6) (3,4) $d = \sqrt{2^2} = 2$
  (3,6) (2,3) $d = \sqrt{1^2 + 3^2} = \sqrt{10}$

Group-2

* (4,3) (2,5) $d = \sqrt{2^2 + 2^2} = \sqrt{8}$
  (4,3) (3,4) $d = \sqrt{1^2 + 1^2} = \sqrt{2}$
  (4,3) (2,3) $d = \sqrt{1^2} = 2$

classmate
Date
Page

~~(2,2), (2,3)~~

* (3, 7) (2, 5) $d = \sqrt{1^2 + 2^2} = \sqrt{5}$ → min
  (3, 7) (3, 4) $d = \sqrt{0 + 3^2} = 3$ wrong cluster
  (3, 7) (2, 3) $d = \sqrt{1^2 + 4^2} = \sqrt{17}$
* (2, 2) (2, 5) $d = \sqrt{0 + 3^2} = 3$
  (2, 2) (3, 4) $d = \sqrt{1 + 2^2} = \sqrt{5}$
  (2, 2) (2, 3) $d = \sqrt{0 + 1^2} = 1$ → min ✓

Group 3

* (1, 5) (2, 5) $d = \sqrt{1^2 + 0} = 1$ → min ✗
  (1, 5) (3, 4) $d = \sqrt{2^2 + 1^2} = \sqrt{5}$
  (1, 5) (2, 3) $d = \sqrt{1 + 2^2} = \sqrt{5}$
* (3, 1) (2, 5) $d = \sqrt{1^2 + 4^2} = \sqrt{17}$
  (3, 1) (3, 4) $d = \sqrt{0 + 3^2} = 3$
  (3, 1) (2, 3) $d = \sqrt{1 + 2^2} = \sqrt{5}$ → min ✓
* (2, 3) (2, 5)
  (2, 3) (3, 4)
  (2, 3) (2, 3)

	Group1	Group2	Group3
	(2,5)	(4, 3)	(2, 2)
	(1,4)		(3, 1)
	(3,6)		(2, 3)
	(3, 7)		
	(1,5)		
centroids	(2, 5.4)	(4, 3)	(2.3, 2)

Q. Apply K-means algorithm to solve the given eg into 3 clusters

	X_i	Y_i	(2,10)	(5,8)	(1,2)	Cluster	(2, 10)	(6,
A1	2	10	0	3.61	8.06	1	0	5.6
A2	2	5	5	4.24	3.16	3	5	
A3	8	4	8.49	5	7.28	2		
~~B1~~	5	8	3.61	0	7.21	2		
B2	7	5	7.07	3.61	6.71	2		
B3	6	4	7.21	4.12	5.39	2		
C1	1	2	8.06	7.21	0	3		
C2	4	9	2.24	1.41	7.62	2		

Updated cent

centroids: (2,10) (5,8) (1,2)

Update centroids

$\frac{8+5+7+6+4}{5} = 6$

$\frac{4+8+5+4+9}{5} = 6$

(New

$\frac{2+4}{2} = 3$

Updated centroids

(2, 10)	(6, 6)	(1.5, 3.5)	New Cluster	(3, 9.5)	(6.5, 5.25)	(1.5, 3.5)
0	5.65	0	1			
5	$\sqrt{4^2+1^2}$		3			
			2			
			2	2.5	3.13	4.87
			2			
			2			
			~~2~~ 3			
			~~3~~ 1			
			~~1~~			

(New Clusters: 1, 3, 2, ~~2~~ 1, 2, 2, 3, ~~2~~ 1)

New

$\frac{2+4}{2} = 3$, $\frac{10+9}{2} = 9.5$ | $8+5+7+6$

X_1	Y_1
2	10
2	5

$A_1, B_1, C_2 \Rightarrow$ Cluster 1
$A_3, B_2, B_3 \Rightarrow$ Cluster 2
$A_2, C_1 \Rightarrow$ Cluster 3

20-3-24

Q.	X	Y	(182, 72)	(~~180~~, 56)	
③	~~168~~	~~60~~	18.43	4.48	→ Cluster 2
④	179	68			→ 1
⑤	182	72			→ 1
⑥	188	77			→ 1
⑦	180	71			→ 1
⑧	180	70			→ 1
⑨	183	84			→ 1
⑩	180	88			→ 1
⑪	180	67			→ 1
⑫	171	76	6.40	21.18	→ 1
①	185	72			→ 1
②	170	56			→ 2

Centroid

Cluster 1	Cluster 2
(179, 68)	(170, 56)
(182, 72)	(168, 60)
(188, 77)	
(180, 71)	
(180, 70)	
(183, 84)	
(180, 88)	
(180, 67)	
(177, 76)	
(185, 72)	
Updated centroids: (181.4, 74.5)	(169, 58)

K1 : 1, 4, 5, 6, 7, 8, 9, 10, 11, 12

K2 : 2, 3

A

$x_1 = (1, 0)$

$x_2 = (2, 2)$

B

$y_1 = (3, 3)$

$y_2 = (4, 4)$

New points: ~~(1, 3) (0, 3)~~

$x_1 y_1 = \sqrt{(3-1)^2 + (3-0)^2} = \sqrt{4+9} = \sqrt{13}$

$x_1 y_2 = \sqrt{(4-1)^2 + (4-0)^2} = \sqrt{3^2 + 4^2} = \sqrt{25} = 5$

$x_2 y_1 = \sqrt{(3-2)^2 + (3-2)^2} = \sqrt{1+1} = \sqrt{2}$

$x_2 y_2 = \sqrt{(4-2)^2 + (4-2)^2} = \sqrt{4+4} = \sqrt{8}$

The smallest of these values is $d(x_2, y_1) = \sqrt{2}$

∴ distance b/w 2 clusters is $d(A, B) = \sqrt{2}$

IMP (3rd)

Normalization

$$x_i = \frac{x_i}{\sqrt{\Sigma x_j}}$$

Q. Consider the following set of 6 one-dimensional ~~set~~ of data points:
18, 22, 25, 42, 27, 43.
Apply aglomerative hierarchical algo to build a dendogram. Merge the clusters using minimum distance & update the proximity matrix accordingly

	18	22	25	27	42	43
18	0	4	7	9	24	25
22	4	0	3	5	20	21
25	7	3	0	2	17	18
27	9	5	2	0	15	16
42	24	20	17	15	0	1
43	25	21	18	16	1	0

* Identify the min(least) value other than 0 = (42,43)

(42,43)

	18	22	25	27	(42,43)
18	0	4	7	9	24
22	4	0	3	5	20
25	7	3	0	2	17
27	9	5	2	0	15
(42,43)	24	20	17	15	0

grouping rows

2 is min

(25,27)

(42,43)(25,27)

	18	22	(25,27)	(42,43)
18	0	4	7	24
22	4	0	3	20
(25,27)	7	3	0	15
(42,43)	24	20	15	0

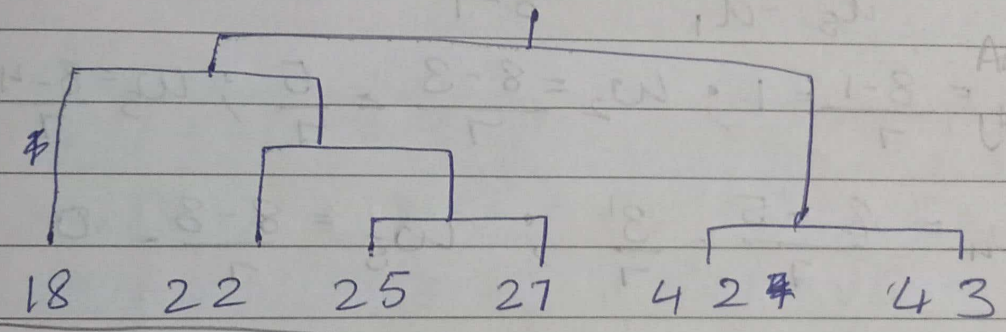
3 is min

(42,43)((25,27),22)

	18	(22,25,27)	(42, 43)
18	0	4	24
(22,25,27)	4	0	15
(42,43)	24	15	0

((42,43),(((25,27),22),18))

	(18,22,25,27)	(42,43)
(18,22,25,27)	0	15
(42,43)	15	0



2-4-24

Weighted Nearest Neighbors

* we are going to assign weights to each of the neighbors
* any neighbor nearest to new instance -> highest weight
* K neighbors ordered according to their distances $d_1, d_2, \ldots d_K$ from x so that d_1 is the smallest & d_K is the greatest

$$w_i = \frac{d_K - d_i}{d_K - d_1}$$

15-4-24

Naive Bayes

Q. The test returns a correct +ve result in only 98% of the cases in which the disease is actually present, and a correct -ve result in only 97% of the cases in which the disease is not present. Furthermore, 0.001 of the entire population have this cancer. Does this patient have cancer or not? How does probability of cancer being +ve compared to probability of not having cancer being +ve.

true +ve

false +ve

- $P(\text{cancer}) = 0.001$

$P(\neg(\text{cancer}) = 1 - P(\text{cancer}) = 1 - 0.001 = 0.999$

(¬ : not having)

$P(\oplus / \text{cancer}) = 0.98, \quad P(\ominus / \text{cancer}) = 0.02$

$P(\ominus / \neg \text{cancer}) = 0.97, \quad P(\oplus / \neg \text{cancer}) = 0.03$

$P(\text{cancer} / \oplus)$ compare with $P(\neg\text{cancer} / \oplus)$

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

$$P(\text{cancer}/\oplus) = \frac{P(\oplus/\text{cancer})\ P(\text{cancer})}{P(\oplus)}$$

$$= \frac{0.98 \times 0.001}{P(\oplus) \rightarrow \text{all +ve cases}}$$

$$= \frac{0.98 \times 0.001}{P(\oplus, \text{cancer}) + P(\oplus, \neg \text{cancer})}$$

$$= \frac{0.98 \times 0.001}{P(\oplus/\text{cancer})\,P(\text{cancer}) + P(\oplus/\neg\text{cancer})\,P(\neg\text{cancer})}$$

$$= \frac{0.98 \times 0.001}{(0.98 \times 0.001) + (0.03 \times 0.999)}$$

$$= 0.031664 = 3.16\%$$

$$P(\text{cancer}/\oplus) = \frac{P(\oplus/\neg\text{cancer}) \cdot P(\text{cancer})}{P(+)}$$

$$= \frac{0.03 \times 0.999}{P(+)}$$

$$= \frac{0.03 \times 0.999}{P(\oplus, \text{cancer}) + P(\oplus, \neg\text{cancer})}$$

$$= 0.02997$$

$$= 0.9683$$

18-4-24

Agglomerative 2-D Problem

Sample	X	Y
P_1	0.40	0.53
P_2	0.22	0.38
P_3	0.35	0.32
P_4	0.26	0.19
P_5	0.08	0.41
P_6	0.45	0.30

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0					
P_2	0.23	0				
P_3	0.22	0.14	0			
P_4	0.37	0.19	0.13	0		
P_5	0.34	0.14	0.28	0.23	0	
P_6	0.24	0.24	0.10	0.22	0.39	0

lowest

(P_3, P_6)

	P_1	P_2	P_3, P_6	P_4	P_5
P_1	0				
P_2	0.23	0			
P_3, P_6	0.22	0.14	0		
P_4	0.37	0.19	0.13	0	
P_5	0.34	0.14	0.28	0.23	0

$((P_3, P_6), P_4)$

	P_1	P_2	P_3, P_4, P_6	P_5
P_1	0			
P_2	0.23	0		
P_3, P_4, P_6	0.22	0.14	0	
P_5	0.34	0.14	0.23	0

$((P_3, P_6) P_4), (P_2, P_5)$

	P_1	$P_2 P_5$	$P_3 P_4 P_6$
P_1	0		
$P_2 P_5$	0.23	0	
P_3, P_4, P_6	0.22	0.14	0

$[(((P_3, P_6), P_4), (P_2, 5)), P_1]$