

Bayesian Learning with Sparse Data

Ninan Sajeeth Philip
St.Thomas College

Classification of CRTS detections

Objective: Predict the nature of the object based on available information
About 40 useful different observations may be found.

```
delm12, "FirstDetection_agnitude", "minus", "SecondDetection_agnitude"  
delm13, "FirstDetection_agnitude", "minus", "ThirdDetection_agnitude"  
delm14, "FirstDetection_agnitude", "minus", "FourthDetection_agnitude"  
delm23, "SecondDetection_agnitude", "minus", "ThirdDetection_agnitude"  
delm24, "SecondDetection_agnitude", "minus", "FourthDetection_agnitude"  
delm34, "ThirdDetection_agnitude", "minus", "FourthDetection_agnitude"  
cdm12, "catotFirstDetection_agnitude", "minus", "catotSecondDetection_agnitude"  
cdm13, "catotFirstDetection_agnitude", "minus", "catotThirdDetection_agnitude"  
cdm14, "catotFirstDetection_agnitude", "minus", "catotFourthDetection_agnitude"  
cdm23, "catotSecondDetection_agnitude", "minus", "catotThirdDetection_agnitude"  
cdm24, "catotSecondDetection_agnitude", "minus", "catotFourthDetection_agnitude"  
cdm34, "catotThirdDetection_agnitude", "minus", "catotFourthDetection_agnitude"  
gi, "pphotT_gpr__Magnitude", "minus", "pphotT_ipr__Magnitude"  
gr, "pphotT_gpr__Magnitude", "minus", "pphotT_rpr__Magnitude"  
gz, "pphotT_gpr__Magnitude", "minus", "pphotT_zpr__Magnitude"  
ir, "pphotT_ipr__Magnitude", "minus", "pphotT_rpr__Magnitude"  
iz, "pphotT_ipr__Magnitude", "minus", "pphotT_zpr__Magnitude"  
rz, "pphotT_rpr__Magnitude", "minus", "pphotT_zpr__Magnitude"  
csNusnoD, "cs_nearest_obj_usnobdistance", "", ""  
csNcgraD, "cs_nearest_obj_cgrabsdistance", "", ""  
csNcrateD, "cs_nearest_obj_cratesdistance", "", ""  
csNnedD, "cs_nearest_obj_neddistance", "", ""  
csNveroD, "cs_nearest_obj_verondistance", "", ""  
csNsimbD, "cs_nearest_obj_simbaddistance", "", ""  
csNnvssD, "cs_nearest_obj_nvssdistance", "", ""  
csNradioD, "cs_nearest_obj_radiodistance", "", ""  
csNxyD, "cs_nearest_obj_xydistance", "", ""
```

Skyalert portfolio

.1943000000000002 -0.2018020000000001 -0.007501999999999879 -9999 -9999 -9999 -9999 -9999 -9999 -9999
99838 0.176599 0.1672000000000001 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999
39799999999986 0.1582019999999999 0.2005999999999998 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999
1654990000000001 -0.03549899999999979 0.1300000000000003 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999
2989999999997 0.206899 -0.3973999999999998 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999
.008200000000000043 -0.04010000000000007 -0.04830000000000011 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999
002 0.5975 0.02860099999999983 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999
9 -0.1476000000000001 0.0577009999999998 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999
.0284999999999993 0.01680000000000017 0.0453000000000001 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999
0890009999999997 -0.03699800000000005 -0.125999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999
.006800000000000014 -0.0171999999999999 -0.0239999999999991 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999
-0.2150010000000001 -0.212799 0.002202000000000048 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999
.0323999999999991 -0.003700000000000026 -0.0360999999999994 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999
003 0.1324000000000001 0.1173 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999
5 0.0419 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 0.2 -9999 -9999 -9999
745 0.1418989999999999 -0.1326010000000001 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999
0.01140000000000001 0.00859999999999995 -0.002800000000000058 0.01399999999999993 0.03689999999999993 0
99999 -0.02570000000000005 -0.322201 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999
8860000000000002 0.07400000000000016 -0.4146 -9999 -9999 -9999 -9999 -9999 -9999 -9999 76.0276 0 78.663 -76.
001 -0.10039899999999999 -0.122398 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999
99998 -0.03979799999999976 0.2125020000000001 -9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 0.2829000000000001 0
-0.01859999999999993 -0.002900000000000035 0.01569999999999989 0.02260000000000006 0.04050000000000015 -
0.02190099999999997 0.01790099999999984 -0.004000000000000134 -9999 -9999 -9999 -9999 -9999 -9999 -9999 4.80
000213 -0.258499 -0.25439799999999998 -9999 -9999 -9999 -9999 -9999 -9999 1.5529 1.0345 -9999 -0.5184

Missing values

Class	Total Instances	Missing-Galaxydata	Missing-stardata	Missing-Both
1(CV)	437	376(86.04%)	375(85.81%)	375(85.81%)
2(SN)	546	379(69.41%)	387(70.87%)	377(69.04%)
3(OT)	19	11(57.89%)	11(57.89%)	11(57.89%)
5(BL)	202	151(74.75%)	151(74.75%)	151(74.75%)
6(AGN)	76	43(56.57%)	43(56.57%)	43(56.57%)
7(UVC)	46	43(93.47%)	43(93.47%)	43(93.47%)
9(VS)	64	57(89.06%)	57(89.06%)	57(89.06%)
10(MV)	9	9(100%)	9(100%)	9(100%)
Total	1399	1069(76.41%)	1076(76.91%)	1066(76.19%)

Most ML algorithms crash with so many missing entries.

Questions

- How will we represent the data vector?

Questions

- How will we represent the data vector?
- How will we train the ML algorithm when most entries are missing?

Questions

- How will we represent the data vector?
- How will we train the ML algorithm when most entries are missing?
- How can we quantify the significance of each input feature?

Questions

- How will we represent the data vector?
- How will we train the ML algorithm when most entries are missing?
- How can we quantify the significance of each input feature?
- What algorithms can handle the scenario?

Questions

- How will we represent the data vector?
- How will we train the ML algorithm when most entries are missing?
- How can we quantify the significance of each input feature?
- What algorithms can handle the scenario?
- How can we quantify the reliability of the predictions?

Data Representation

- The nature or class of the object is fixed. What changes is the available input to describe it.

Data Representation

- The nature or class of the object is fixed. What changes is the available input to describe it.
- It was decided to use a binary genetic coding to represent the input vector.
- Each observation has a particular slot in the genetic chain/sequence. A 0 means missing.
- All events with the same genetic code have the same set of observations.
- New observations are appended to the right of the chain, reading from left to right.

Training on Missing Data

- With every input vector, also supply an input mask (genetic code for the input vector) so that the ML algorithm knows what to use for learning and what all is to be ignored.
- Each *species vector* is treated independently by the learning algorithm.
- Update the algorithm (for available inputs) such that the cost function is minimised.
- Test on new data with known nature to estimate the accuracy of the predictions.

```
1 1, "Cataclysmic Variable"  
2 2, "Supernova"  
3 3, "other"  
4 5, "Blazar Outburst"  
5 6, "Active Galactic Nucleus Variability"  
6 7, "UVCeti Variable"  
7 8, "Asteroid"  
8 9, "Variable"  
9 10, "Mira Variable"  
0 11, "High Proper Motion Star"  
1 12, "Comet"  
2 16, "Nova"
```

Results (Training used 80% data)

Results (Training used 80% data)

1	"Cataclysmic Variable"				
2	"Supernova"				
3	"other"				
4	"Blazar Outburst"				
5	"Active Galactic Nucleus Variability"				
6	"UVCeti Variable"				
7	"Asteroid"	"1106111350474117100"	,2,	2.000000,100.000000,	1.000000,0.000000
8	"Variable"	"1106111320694137973"	,1,	1.000000,73.974426,	6.000000,26.025574
9	"Mira Variable"	"1106111260764139610"	,2,	2.000000,98.924429,	1.000000,1.075571
10	"High Proper Moti	"1106111070884131228"	,1,	1.000000,99.999996,	6.000000,0.000004
11	"Comet"	"1106101520474116185"	,2,	2.000000,92.100441,	6.000000,5.611119
12	"Nova"	"1106101380624102629"	,11,	11.000000,99.952249,	2.000000,0.047751
13		"1106101350804125610"	,1,	2.000000,51.359278,	1.000000,48.640722
14		"1106101290854138513"	,1,	1.000000,58.537439,	2.000000,41.462561
15		"1106091460514127317"	,2,	2.000000,100.000000,	1.000000,0.000000
16		"1106091430534110587"	,2,	2.000000,100.000000,	6.000000,0.000000
17		"1106091400584128136"	,2,	2.000000,100.000000,	6.000000,0.000000
18		"1106091210774115315"	,2,	2.000000,100.000000,	6.000000,0.000000
19		"1106081041174137140"	,2,	2.000000,89.355383,	1.000000,10.644617
20		"1106071400654101671"	,6,	6.000000,100.000000,	1.000000,0.000000
21		"1106070070614107631"	,2,	2.000000,100.000000,	1.000000,0.000000
22		"1106070010614108286"	,2,	2.000000,99.957624,	1.000000,0.042230
23		"1106061121124182967"	,1,	1.000000,99.989827,	2.000000,0.010173
24		"1106060010764124342"	,2,	2.000000,100.000000,	6.000000,0.000000
25		"1106051180764143736"	,11,	11.000000,99.999783,	2.000000,0.000217
26		"1106051150794129411"	,2,	2.000000,100.000000,	6.000000,0.000000
27		"1106040230824154629"	,10,	1.000000,58.537439,	2.000000,41.462561
28		"1106040210844128759"	,10,	2.000000,90.926754,	1.000000,9.073243
29		"1106040010704126932"	,2,	2.000000,100.000000,	1.000000,0.000000
30		"1106030120874113978"	,1,	1.000000,76.496431,	2.000000,23.503569

Results

	Real->	[1]	[2]	[3]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[16]	Total
1, "Cataclysmic Variable"														
2, "Supernova"														
3, "other"														
4, "Blazar Outburst"														
5, "Active Galactic Nucleus Variability"														
6, "UVCeti Variable"														
7, "Asteroid"														
8, "Variable"														
9, "Mira Variable"														
10, "High Proper Motior"	Predict													
11, "Comet"	[1]	261	3	6	0	1	0	2	4	3	1	0	0	281
12, "Nova"	[2]	4	398	4	1	2	1	3	2	0	0	1	2	418
	[3]	0	0	28	0	0	0	0	0	0	0	0	0	28
	[5]	0	0	0	66	0	0	0	0	1	0	0	0	67
	[6]	1	0	0	0	131	1	0	0	1	0	0	0	134
	[7]	0	0	0	0	0	28	0	0	0	0	0	0	28
	[8]	0	0	0	0	0	0	5	0	0	0	0	0	5
	[9]	0	0	0	0	0	0	0	18	0	0	0	0	18
	[10]	0	0	0	0	0	0	0	0	10	0	0	0	10
	[11]	0	0	1	0	0	0	0	0	0	40	0	0	41
	[12]	0	0	0	0	0	0	0	0	0	0	3	0	3
	[16]	0	0	0	0	0	0	0	0	0	0	0	0	0
	[]													
	Total	266	401	39	67	134	30	10	24	15	41	4	2	1033

Cross Validation

Real-> [1] , Predict	[2] ,	[3] ,	[5] ,	[6] ,	[7] ,	[8] ,	[9] ,	[10] ,	[11] ,	[12] ,	[16] ,	Total	Real-> [1] , Predict	[2] ,	[3] ,	[5] ,	[6] ,	[7] ,	[8] ,	[9] ,	[10] ,	[11] ,	[12] ,	[16] ,	Total
[1] , 39	12	2	0	1	2	0	1	2	1	0	0	60	[1] , 29	8	2	4	4	0	0	1	1	2	0	0	51
[2] , 16	55	8	1	5	0	1	2	0	4	1	0	93	[2] , 15	53	0	0	10	2	2	1	0	2	0	0	85
[3] , 0	1	0	0	0	0	0	0	0	0	0	0	1	[3] , 1	2	3	0	0	1	0	0	1	0	0	0	8
[5] , 0	0	1	5	0	0	0	0	0	0	0	0	6	[5] , 0	0	0	6	0	0	0	0	1	0	0	0	7
[6] , 6	6	3	3	20	1	0	0	0	0	0	0	39	[6] , 7	10	1	6	13	4	0	1	0	0	0	0	42
[7] , 0	0	0	0	0	1	0	0	0	0	0	0	1	[7] , 0	0	0	0	0	1	0	0	0	0	0	0	1
[8] , 0	0	0	0	0	0	0	0	0	0	0	0	0	[8] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[9] , 0	0	0	0	0	0	0	0	0	0	0	0	0	[9] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[10] , 0	0	0	0	0	0	0	0	0	0	0	0	0	[10] , 2	0	0	0	0	0	0	0	0	0	0	0	2
[11] , 0	0	0	0	0	0	0	0	0	0	0	0	0	[11] , 0	0	0	0	0	0	0	0	0	5	0	0	5
[12] , 0	0	0	0	0	0	0	0	0	0	0	0	0	[12] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[16] , 0	0	0	0	0	0	0	0	0	0	0	0	0	[16] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[]													[]												
Total 61	74	14	9	26	4	1	3	2	5	1	0	200	Total 54	73	6	16	27	8	2	3	3	9	0	0	201
Real-> [1] , Predict	[2] ,	[3] ,	[5] ,	[6] ,	[7] ,	[8] ,	[9] ,	[10] ,	[11] ,	[12] ,	[16] ,	Total	Real-> [1] , Predict	[2] ,	[3] ,	[5] ,	[6] ,	[7] ,	[8] ,	[9] ,	[10] ,	[11] ,	[12] ,	[16] ,	Total
[1] , 24	12	2	2	3	1	0	3	1	4	0	0	52	[1] , 31	10	0	1	3	1	0	0	5	3	0	0	54
[2] , 7	65	1	1	2	2	2	1	1	2	1	1	86	[2] , 12	61	7	1	4	1	1	2	0	10	0	1	100
[3] , 0	1	0	0	1	0	1	0	1	0	0	0	4	[3] , 0	1	0	0	1	0	0	0	0	0	0	0	2
[5] , 1	0	0	11	1	0	0	1	0	0	0	0	14	[5] , 0	0	0	3	2	0	0	0	1	0	0	0	6
[6] , 7	6	1	6	15	2	0	1	0	0	0	0	38	[6] , 4	7	3	4	11	3	0	0	0	0	0	0	32
[7] , 0	0	0	0	0	2	0	0	0	0	0	0	2	[7] , 1	0	0	0	0	1	0	0	0	0	0	0	2
[8] , 0	0	0	0	0	0	0	0	0	0	0	0	0	[8] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[9] , 0	0	0	0	0	0	0	0	0	0	0	0	0	[9] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[10] , 0	0	0	1	0	0	0	1	0	0	0	0	2	[10] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[11] , 0	0	0	0	0	0	0	0	0	2	0	0	2	[11] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[12] , 0	0	0	0	0	0	0	0	0	0	0	0	0	[12] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[16] , 0	0	0	0	0	0	0	0	0	0	0	0	0	[16] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[]													[]												
Total 39	84	4	21	21	8	2	8	2	9	1	1	200	Total 48	79	10	9	21	6	1	2	6	13	0	1	196
Real-> [1] , Predict	[2] ,	[3] ,	[5] ,	[6] ,	[7] ,	[8] ,	[9] ,	[10] ,	[11] ,	[12] ,	[16] ,	Total	Real-> [1] , Predict	[2] ,	[3] ,	[5] ,	[6] ,	[7] ,	[8] ,	[9] ,	[10] ,	[11] ,	[12] ,	[16] ,	Total
[1] , 39	9	0	0	13	2	1	2	0	2	0	0	68	[1] , 6	2	0	0	1	1	0	0	1	0	0	0	11
[2] , 9	57	0	0	7	0	0	1	0	2	2	0	78	[2] , 6	12	0	0	2	0	2	1	0	0	0	0	23
[3] , 0	1	1	0	1	0	0	0	0	0	0	0	3	[3] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[5] , 1	0	0	7	2	0	0	1	0	0	0	0	11	[5] , 0	1	0	1	0	0	0	0	0	0	0	0	2
[6] , 4	8	1	3	12	0	1	2	0	0	0	0	31	[6] , 0	0	1	0	0	0	0	1	0	0	0	0	2
[7] , 0	0	2	0	0	1	0	0	0	0	0	0	3	[7] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[8] , 0	0	0	0	0	0	0	0	0	0	0	0	0	[8] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[9] , 0	0	0	0	0	0	0	0	0	0	0	0	0	[9] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[10] , 1	0	0	0	1	0	0	0	0	1	0	0	3	[10] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[11] , 1	0	0	0	0	0	0	0	0	1	0	0	2	[11] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[12] , 0	0	0	0	0	0	0	0	0	0	0	0	0	[12] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[16] , 0	0	0	0	0	0	0	0	0	0	0	0	0	[16] , 0	0	0	0	0	0	0	0	0	0	0	0	0
[]													[]												
Total 55	75	4	10	36	3	2	6	1	5	2	0	199	Total 12	15	1	1	3	1	2	2	1	0	0	0	38

Continued Learning Model

Humans Learn from every experience. Let us replicate the same in machines too- Continued training and evaluation.

Advantages

- The learning will progress with time – the key features required for correct identification will be exposed by the failing candidates – feed back.
- Prediction accuracy will asymptotically improve until the system replaces human expert.

Latest Status

Real->	[1]	[2]	[3]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[16]	Total
Predict													
[1]	272	4	7	1	1	0	1	4	3	2	0	1	296
[2]	4	404	4	0	4	1	3	2	0	2	1	0	425
[3]	0	1	34	0	0	0	0	0	0	0	0	0	35
[5]	0	0	1	60	0	0	0	0	1	0	0	0	62
[6]	1	0	0	1	126	0	0	0	0	0	0	0	128
[7]	0	0	0	0	0	32	0	0	0	0	0	0	32
[8]	0	0	0	0	0	0	6	0	0	0	0	0	6
[9]	0	0	0	0	0	0	0	18	0	0	0	0	18
[10]	0	0	0	0	0	0	0	0	12	0	0	0	12
[11]	0	0	0	0	0	0	0	0	0	44	0	0	44
[12]	0	0	0	0	0	0	0	0	0	0	5	0	5
[16]	0	0	0	0	0	0	0	0	0	0	0	1	1
[]													
Total	277	409	46	62	131	33	10	24	16	48	6	2	1064

Real->	[1]	[2]	[3]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[16]	Total
Predict													
[1]	1589	0	0	0	0	0	0	0	0	0	0	0	1589
[2]	0	1776	0	0	0	0	0	0	0	0	0	0	1776
[3]	0	0	268	0	0	0	0	0	0	0	0	0	268
[5]	0	0	0	94	0	0	0	0	0	0	0	0	94
[6]	0	0	0	0	603	0	0	0	0	0	0	0	603
[7]	0	0	0	0	0	83	0	0	0	0	0	0	83
[8]	0	0	0	0	0	0	8	0	0	0	0	0	8
[9]	0	0	0	0	0	0	0	24	0	0	0	0	24
[10]	0	0	0	0	0	0	0	0	23	0	0	0	23
[11]	0	0	0	0	0	0	0	0	0	67	0	0	67
[12]	0	0	0	0	0	0	0	0	0	0	14	0	14
[16]	0	0	0	0	0	0	0	0	0	0	0	1	1
[]													
Total	1589	1776	268	94	603	83	8	24	23	67	14	1	4550

* Total Success in Success in Non classified Real success in
 * No. Ist Choice 2nd Choice items two chances
 * 1092 1019 3 20 95.3358 %