

A brief overview, comparison and practical applications of machine learning models

HANDS-ON WORKSHOP - LIVE

Dan Tulpan, Assistant Professor
Centre for Genetic Improvement of Livestock
Department of Animal Biosciences
Ontario Agricultural College
University of Guelph
dtulpan@uoguelph.ca



IMPROVE LIFE.

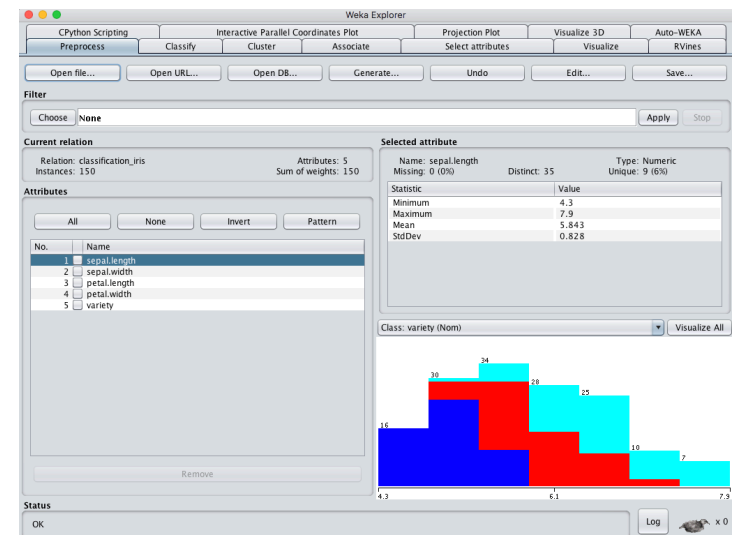
ASAS-NANP Pre-Conference Symposium
ASAS-CSAS-WSASAS Virtual Annual Meeting & Trade Show,
July 19-23, 2020

ONTARIO
AGRICULTURAL COLLEGE
DEPARTMENT OF ANIMAL BIOSCIENCES



Summary

- The Weka Explorer
- Classification problems
- Regression problems
- Data-related artifacts



Classification Problem 1

Data set: `classification_iris.csv`

- Number of data points: 150
- Number of classes: 3
- Number of attributes: 4
 - SL: Sepal length (cm)
 - SW: Sepal width (cm)
 - PL: Petal length (cm)
 - PW: Petal width (cm)

sepal.length	sepal.width	petal.length	petal.width	variety
5.1	3.5	1.4	0.2	Setosa
4.9	3	1.4	0.2	Setosa
4.7	3.2	1.3	0.2	Setosa
7	3.2	4.7	1.4	Versicolor
6.4	3.2	4.5	1.5	Versicolor
6.9	3.1	4.9	1.5	Versicolor
6.3	3.3	6	2.5	Virginica
5.8	2.7	5.1	1.9	Virginica
7.1	3	5.9	2.1	Virginica



Iris setosa



Iris versicolor



Iris virginica

Classification Problem 2

Data set: `classification_random_binary.csv`

- Number of data points: 100
- Number of classes: 2
- Number of attributes: 5
 - Cow_id
 - MS1
 - MS2
 - MS3
 - MS4

Cow_id	Ms1	Ms2	Ms3	Ms4	Pred_phen
1000	0.82670136	3.57660749	11.1282203	11.9935777	no
1004	0.75360992	7.04655697	10.3097353	9.39415003	no
1008	0.05777438	5.72303611	14.2557181	11.7663005	no
1012	0.23848243	5.35097367	13.0912955	12.5702272	yes
1016	0.16725548	4.16588693	9.76702334	12.1613706	yes
1020	0.69281354	6.45123742	12.519231	9.56385006	yes
1024	0.52219519	5.44130996	9.0774877	11.1876432	no
1028	0.636972	3.2613616	10.4232098	11.7134081	yes
1032	0.25784654	3.3175609	15.3697456	11.9128922	yes

Classification Problem 3

Data set: `classification_zoo_dataset.csv`

- Number of data points: 101
- Number of classes: 7
- Number of attributes: 17
 - animal_name
 - Hair
 - Feathers
 - ...

animal_name	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	legs	tail	domestic	catsize	class_type
aardvark	1	0	0	1	0	0	1	1	1	1	0	0	4	0	0	1	Mammal
antelope	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1	Mammal
bass	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	Fish
bear	1	0	0	1	0	0	1	1	1	1	0	0	4	0	0	1	Mammal
boar	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	Mammal
buffalo	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1	Mammal
calf	1	0	0	1	0	0	0	1	1	1	0	0	4	1	1	1	Mammal
carp	0	0	1	0	0	1	0	1	1	0	0	1	0	1	1	0	Fish
catfish	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	Fish
cavy	1	0	0	1	0	0	0	1	1	1	0	0	4	0	1	0	Mammal
cheetah	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	Mammal
chicken	0	1	1	0	1	0	0	0	1	1	0	0	2	1	1	0	Bird
chub	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	Fish
clam	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	Invertebrate

Regression Problem 1

Data set: `regression_linear_x_0_99_err0.csv`

- Number of data points: 100
- Number of attributes: 1
 - x

x	$f(x)=2*x+3$
0	3
1	5
2	7
3	9
4	11
5	13
6	15
7	17
8	19

Regression Problem 2

Data set: `regression_quadratic.csv`

- Number of data points: 100
- Number of attributes: 1
 - x

x	$f(x)=2*x+3$
0	3
1	5
2	7
3	9
4	11
5	13
6	15
7	17
8	19

Regression Problem 3

Data set: `regression_complex_func.csv`

- Number of data points: 100
- Number of attributes: 4
 - x
 - y
 - z
 - t

x	y	z	t	$f(x,y,z,t) = 5x - 2\cos(y) + 3z^2/\sqrt{t}$
1	1	1	1	6.9194
2	2	2	2	27.8029
3	3	3	3	63.7454
4	4	4	4	117.3073
5	5	5	5	192.1378
6	6	6	6	292.6246
7	7	7	7	422.4176
8	8	8	8	583.349
9	9	9	9	775.8223
10	10	10	10	1000.3614
11	11	11	11	1258.9259

Regression Problem 4

Data set: `regression_random.csv`

- Number of data points: 100
- Number of attributes: 5
 - Cow_id
 - Ms1
 - Ms2
 - Ms3
 - Ms4

Cow_id	Ms1	Ms2	Ms3	Ms4	Pred_phen
1000	0.9470999	5.95797774	13.915448	12.9347224	0
1004	0.61073964	5.51292552	10.6797067	11.6490622	0
1008	0.79457971	6.14149345	11.5871536	10.3897927	0
1012	0.26283342	6.98005889	9.47012415	11.9177109	1
1016	0.73682092	5.88477566	9.54904708	12.1499646	1
1020	0.34022992	6.22943088	10.1080563	9.20260174	0
1024	0.31375375	7.23901781	11.2240887	11.5455358	0
1028	0.97597542	5.58866531	11.5971538	10.7777561	0
1032	0.70038486	6.92601426	10.50353	12.2210478	0
1036	0.98200496	5.25294086	13.4670354	9.02896927	1

Data-Related Artifacts

- Formatting errors
 - Special symbols
 - Extra columns
 - Duplicate header labels
- Hidden correlations
 - Unnecessary columns directly correlated with the predictor variable

Formatting Errors

data_error1_column_50_special_symbol.csv

47	1180	0.56906494	5.49645174	12.8512741	10.1804572	yes
48	1184	0.04040781	5.54765022	13.3681585	9.2797879	yes
49	1188	0.12406309	5.77318337	10.8367315	11.6485751	yes
50	1192	0.16223907	5.49"	11.3893505	10.9806007	no
51	1196	0.26181768	4.95858857	10.8131897	9.72609505	yes
52	1200	0.01718791	4.48834486	10.3239951	11.9754723	yes
53	1204	0.54080607	5.08725736	9.90381763	12.3617106	no

data_error2_extra_columns.csv

93	1364	0.9791347	4.5439163	9.817484	9.2959242	yes	
94	1368	0.7994233	7.4402879	10.065688	9.135617	yes	maybe good
95	1372	0.8082822	4.9122384	11.339509	9.072897	no	
96	1376	0.5848974	6.9507931	10.75717	11.55033	no	throw away
97	1380	0.6387242	5.8203253	9.4522329	12.018035	yes	
98	1384	0.5163807	5.8681784	10.936714	9.6650375	no	

Formatting Errors

`header_error1_single_quote.csv`

1	Cow_id	Ms1	Ms2'	Ms3	Ms4	Pred_phen
2	1000	0.54307371	7.09702619	10.2047803	12.1258179	no
3	1004	0.12472931	4.12297535	11.0922358	9.0506221	no
4	1008	0.61846058	5.73718235	11.7349934	9.97657243	no
5	1012	0.6453044	5.0846034	9.76661321	9.6397819	yes
6	1016	0.42973124	6.48659263	11.9666711	10.1245367	yes

`header_error2_duplicate_column_labels.csv`

1	Cow_id	Ms1	Ms4	Ms3	Ms4	Pred_phen
2	1000	0.26211615	6.99414146	12.0456984	9.51521005	no
3	1004	0.27628907	4.80414009	14.1992321	10.8114346	no
4	1008	0.39226337	4.88113945	14.5932792	9.947972	no
5	1012	0.57436216	4.63288931	13.1472392	12.1626214	yes
6	1016	0.97438431	6.60621846	10.9354242	11.5387297	yes
7	1020	0.33248328	3.68439992	11.3911645	10.2929808	yes

Formatting Errors

`header_error3_double_quotes.csv`

1	"Cow_id"	"Ms1"	"Ms2"	"Ms3"	"Ms4"	"Pred_phen"
2	1000	0.9082203	4.6205842	12.304296	10.26865	no
3	1004	0.4561411	4.1029248	10.416078	9.7241627	no
4	1008	0.6286037	5.283989	13.20705	9.7251261	no
5	1012	0.7591403	3.8496747	12.942408	9.5352086	yes
6	1016	0.7561974	4.2597873	12.683271	12.354336	yes
7	1020	0.7930203	7.0657708	10.398488	9.118181	yes
8	1024	0.7171487	6.9522685	10.689412	12.498773	no
9	1028	0.5653197	6.9458774	9.8353356	12.712097	yes

Correlated features

- Lead to over-inflated predictions

`overinflated_acc_ded_column_E_is_dependent_on_A.csv`

animal_id	gen1	gen2	gen3	pred_gen
12	2.291	2.294	103.375	0
12	2.688	1.425	81.123	0
12	2.71	0.395	68.144	0
12	2.02	2.174	102.877	0
17	2.001	2.684	70.788	7
17	2.941	0.047	82.091	7
17	2.289	1.525	58.188	7
17	2.346	2.439	76.152	7
22	2.199	2.478	97.014	14
22	2.583	2.188	74.66	14
22	2.974	2.178	104.732	14
22	2.41	2.801	73.528	14
27	2.952	0.806	110.689	21
27	2.43	0.411	102.791	21
27	2.382	2.03	77.434	21
27	2.447	0.294	62.878	21
32	2.011	2.977	94.383	28