Classification using Machine-Learning Algorithms (MALA)

Math 579
Dr. Jen-Mei Chang

Rodrigo Farnham
Tuyen Ly
Jason Wang

Our Idea

"The placenta is the crystal ball of the baby."

- Dr. Carolyn Salafia

- Extract and identify features of the placenta
- Extract relevant patient/mother data
- Use above data to train a machine-learning system, hopefully to make predictions of a baby's future health

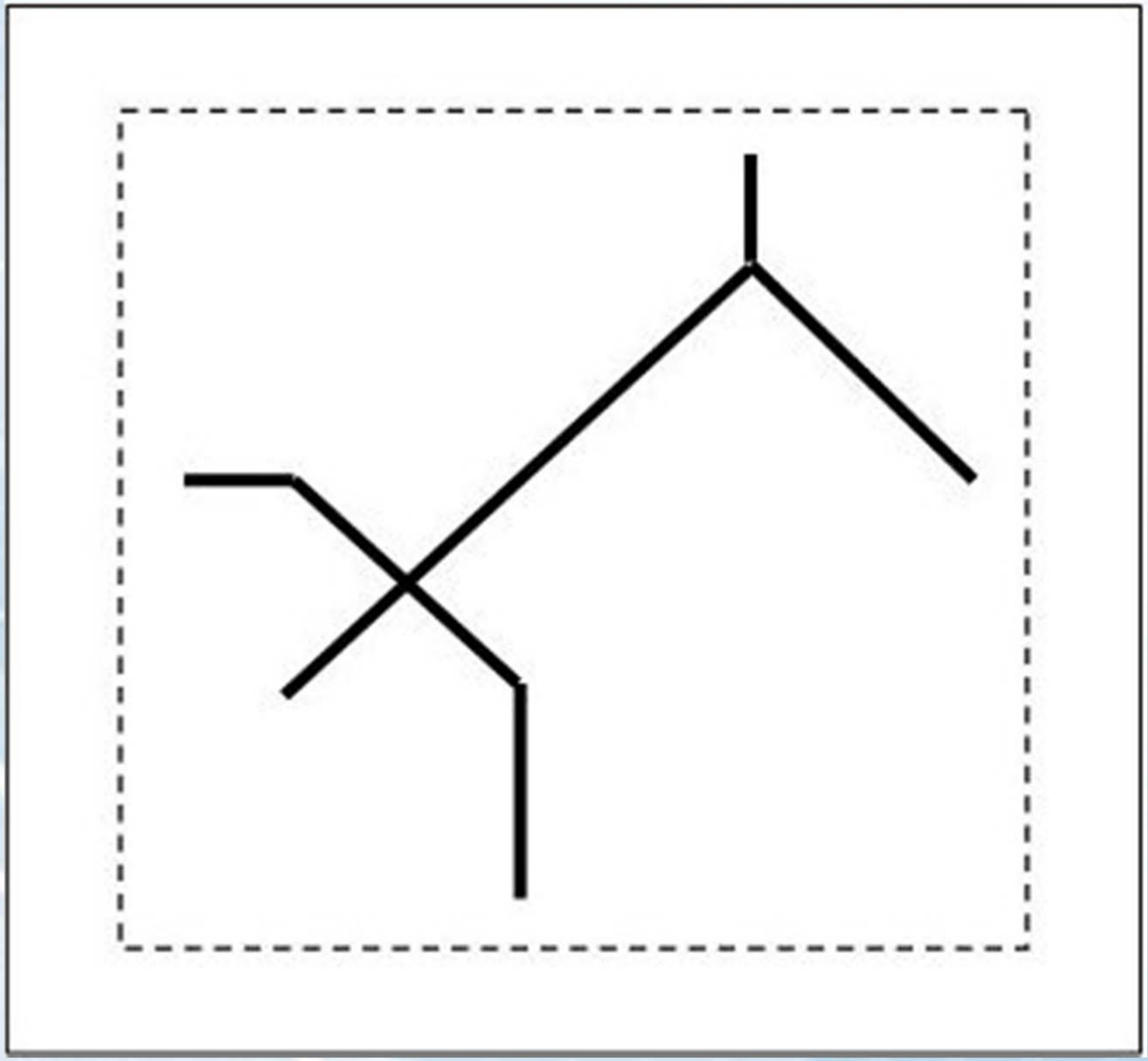
Overview

1) Analyze placental attributes
2) Maternal Attribute Clasification
3) Learning Machine System: Weka

   "*A suite of machine learning software written at the University of Waikato*"

Placental Attribute: Heat Map

Sample Vessel Network



Placental Attribute: Heat Map

Numerical Representation of the Vessel Network (done by hand)

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	1	0	0
0	1	1	0	1	0	0	0	1	0
0	0	0	1	0	0	0	0	0	0
0	0	1	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Placental Attribute: Heat Map

Matlab Output

Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
Inf						0			Inf
Inf						0			Inf
Inf					0		0		Inf
Inf	0	0		0				0	Inf
Inf			0						Inf
Inf		0		0					Inf
Inf				0					Inf
Inf				0					Inf
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf

Placental Attribute: Heat Map

Matlab Output

Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
Inf					1	0	1		Inf
Inf				1.41	1	0	1	1.41	Inf
Inf	1	1	1.41	1	0	1	0	1	Inf
Inf	0	0	1	0	1	1.41	1	0	Inf
Inf	1	1	0	1	1.41		1.41	1	Inf
Inf	1	0	1	0	1				Inf
Inf	1.41	1	1	0	1				Inf
Inf			1	0	1				Inf
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf

Placental Attribute: Heat Map

Matlab Output

Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
Inf			2.83	2	1	0	1	2	Inf
Inf	2	2	2.41	1.41	1	0	1	1.41	Inf
Inf	1	1	1.41	1	0	1	0	1	Inf
Inf	0	0	1	0	1	1.41	1	0	Inf
Inf	1	1	0	1	1.41	2.41	1.41	1	Inf
Inf	1	0	1	0	1	2	2.41	2	Inf
Inf	1.41	1	1	0	1	2			Inf
Inf	2.41	2	1	0	1	2			Inf
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf

Placental Attribute: Heat Map

Matlab Output

Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
Inf	3	3	2.83	2	1	0	1	2	Inf
Inf	2	2	2.41	1.41	1	0	1	1.41	Inf
Inf	1	1	1.41	1	0	1	0	1	Inf
Inf	0	0	1	0	1	1.41	1	0	Inf
Inf	1	1	0	1	1.41	2.41	1.41	1	Inf
Inf	1	0	1	0	1	2	2.41	2	Inf
Inf	1.41	1	1	0	1	2	3	3	Inf
Inf	2.41	2	1	0	1	2	3		Inf
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf

Placental Attribute: Heat Map

Matlab Output

Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
Inf	3	3	2.83	2	1	0	1	2	Inf
Inf	2	2	2.41	1.41	1	0	1	1.41	Inf
Inf	1	1	1.41	1	0	1	0	1	Inf
Inf	0	0	1	0	1	1.41	1	0	Inf
Inf	1	1	0	1	1.41	2.41	1.41	1	Inf
Inf	1	0	1	0	1	2	2.41	2	Inf
Inf	1.41	1	1	0	1	2	3	3	Inf
Inf	2.41	2	1	0	1	2	3	4	Inf
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf

Placental Attribute: Heat Map

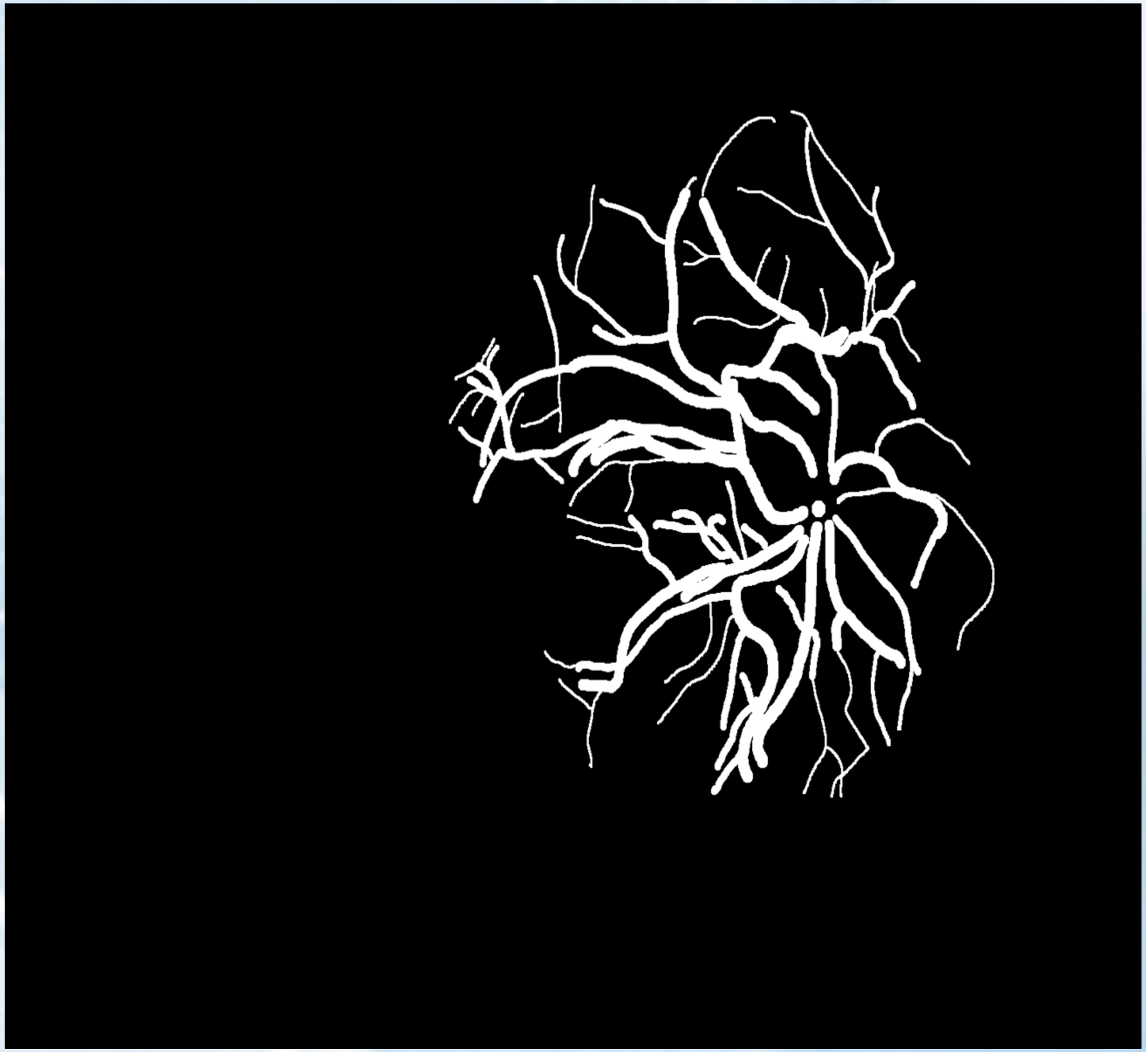
Matlab Output

Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
Inf	3	3	2.83	2	1	0	1	2	Inf
Inf	2	2	2.41	1.41	1	0	1	1.41	Inf
Inf	1	1	1.41	1	0	1	0	1	Inf
Inf	0	0	1	0	1	1.41	1	0	Inf
Inf	1	1	0	1	1.41	2.41	1.41	1	Inf
Inf	1	0	1	0	1	2	2.41	2	Inf
Inf	1.41	1	1	0	1	2	3	3	Inf
Inf	2.41	2	1	0	1	2	3	4	Inf
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf

largest value: (largest distance from vessel network)

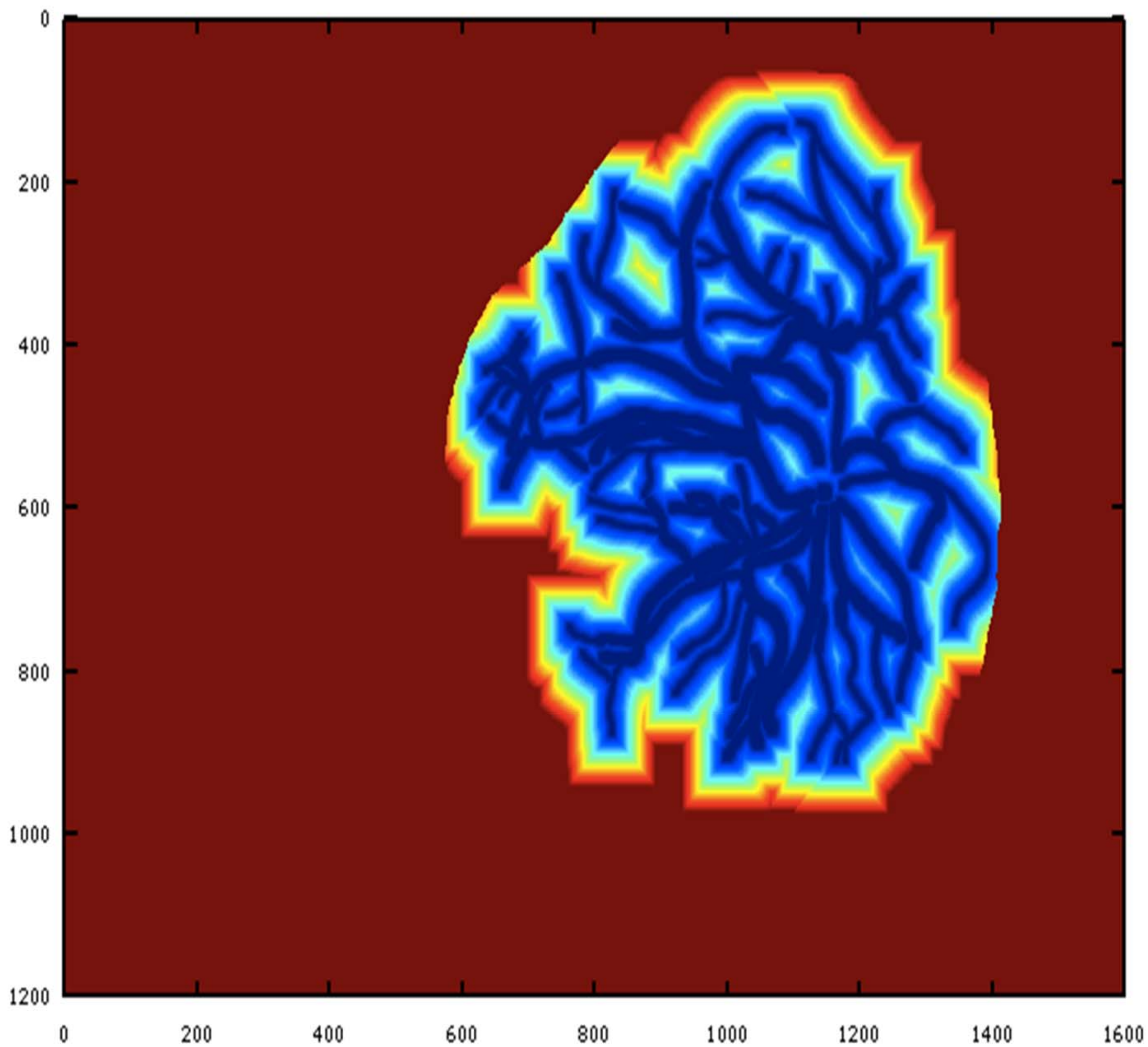
Placental Attribute: Heat Map

Vessel Network



Placental Attribute: Heat Map

Vessel Network



Maternal Attributes

Strength:

•A whole lot of data, total of 209 attributes

- mom's ethnicity
- mom's height
- mom's age at pregnancy start date
- child's birth weight
- mom's total weight gain
- # of previous pregnancies
- # of previous live births
- etc.

Weaknesses:

•*Too much* data

↳A lot of irrelevant data as well as dependent data

Maternal Attributes

	A	B	C	D	E	F	G	H	
1	LABID	BirthWeight	GestationalDays	Weightbeforecuttingcordmembrane	Distancebetweenrupturesiteplacentalmargin	CordLength	CordWeight	Weightaftercuttingcordmembrane	DELV
2	1,353	2,770	261	530	5-10	20	#NULL!	400	
3	1,402	2,692	254	-99	-99	-99	#NULL!	-99	
4	1,433	2,624	262	440	<5	45	#NULL!	350	
5	1,462	4,018	284	420	<5	37	#NULL!	340	
6	1,471	3,489	256	570	<5	51	#NULL!	450	
7	1,472	3,338	278	460	5-10	23	#NULL!	370	
8	1,473	#NULL!	#NULL!	#NULL!		#NULL!	#NULL!	#NULL!	
9	1,474	3,421	286	560	<5	36	#NULL!	450	
10	1,475	#NULL!	#NULL!	#NULL!		#NULL!	#NULL!	#NULL!	
11	1,476	#NULL!	#NULL!	#NULL!		#NULL!	#NULL!	#NULL!	
12	1,478	#NULL!	#NULL!	#NULL!		#NULL!	#NULL!	#NULL!	
13	1,479	2,369	285	370	-99	48	#NULL!	320	
14	1,480	#NULL!	#NULL!	#NULL!		#NULL!	#NULL!	#NULL!	
15	1,482	-999	187	300	<5	19	#NULL!	240	
16	1,489	3,790	274	600	<5	60	#NULL!	460	
17	1,490	#NULL!	#NULL!	#NULL!		#NULL!	#NULL!	#NULL!	
18	1,491	#NULL!	#NULL!	#NULL!		#NULL!	#NULL!	#NULL!	
19	1,492	#NULL!	#NULL!	#NULL!		#NULL!	#NULL!	#NULL!	
20	1,498	#NULL!	#NULL!	#NULL!		#NULL!	#NULL!	#NULL!	
21	1,499	#NULL!	#NULL!	#NULL!		#NULL!	#NULL!	#NULL!	
22	1,501	#NULL!	#NULL!	#NULL!		#NULL!	#NULL!	#NULL!	
23	1,502	3,172	277	470	Unknown	33	#NULL!	400	
24	1,503	3,042	268	480	<5	56	#NULL!	340	
25	1,504	3,657	295	560	5-10	57	#NULL!	460	
26	1,505	#NULL!	#NULL!	#NULL!		#NULL!	#NULL!	#NULL!	#NU
27	1,506	3,185	275	540	5-10	37	#NULL!	450	
28	1,507	3,391	271	690	<5	34	#NULL!	600	
29	1,508	2,556	259	570	<5	61	#NULL!	420	
30	1,509	3,484	266	510	<5	51	#NULL!	420	
31	1,510	#NULL!	#NULL!	#NULL!		#NULL!	#NULL!	#NULL!	
32	1,511	#NULL!	#NULL!	#NULL!		#NULL!	#NULL!	#NULL!	
33	1,512	3,462	281	530	<5	42	#NULL!	430	
34	1,513	2,220	243	440	5-10	15	#NULL!	330	
35	1,514	2,885	270	530	>10	57	#NULL!	440	
36	1,515	2,872	280	500	<5	50	#NULL!	420	
37	1,516	2,827	263	500	<5	53	#NULL!	390	
38	1,517	3,139	264	600	<5	35	#NULL!	470	

jenmei jan 2011

Maternal Attributes

<u>The attributes we chose to look at:</u>

- "Reasonable" attributes
- Attributes whose data is "spread out"
- Results from research papers/publications

I. Mother's total weight gain[1]
II. Number of previous pregnancies
III. Summary Index # of Prenatal Care Adequacy[2]
IV. Gestational Days
V. Family's Poverty Level Index[3]

1 http://www.telegraph.co.uk/health/healthnews/7926233/Putting-on-too-much-weight-in-pregnancy-risk-babys-health.html
2 http://www.sjph.net.sd/files/vol4i4/SJPH-vol4i4-p403-410.pdf
3 http://www.epi.umn.edu/mch/resources/hg/hg_childpoverty.pdf

Birth Weight & Beta Value

The Importance of Birth Weight:

doi: 10.1111/j.1365-3016.2008.00935.x 229

Fetal growth correlates

Placental characteristics and birthweight

Carolyn M. Salafia[a,d], Jun Zhang[e], Adrian K. Charles[f], Michaeline Bresnahan[a,b], Patrick Shrout[c], Wenyu Sun[e] and Elizabeth M. Maas[d]

[a]Department of Epidemiology, Mailman School of Public Health, Columbia University College of Physicians and Surgeons, [b]New York State Psychiatric Institute, [c]Department of Psychology, New York University, New York, NY, [d]EarlyPath Clinical and Research Diagnostics, Larchmont, NY, [e]Epidemiology Branch, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD, USA, and [f]Department of Pathology, Princess Margaret Hospital, Perth, Western Australia

Correspondence:
Carolyn M. Salafia, MD MS, Assistant Professor of Epidemiology, Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, NY 10032, USA.
E-mail: salafiacm@aol.com

Summary

Salafia CM, Zhang J, Charles AK, Bresnahan M, Shrout P, Sun W, Maas EM. Prenatal characteristics and birthweight. *Paediatric and Perinatal Epidemiology* 2008; **22:** 229–239.

Standard gross placental measures capture dimensions relevant to specific placental functions. Our objective was to determine their accountability independent of placental weight for variance in birthweight, an important proxy for intrauterine 'adequacy' in fetal origins studies. The sample consisted of 24 152 singleton liveborn children of the Collaborative Perinatal Project delivered from 34 to 42 completed weeks gestation, with complete data for six placental measures (placental disc shape, umbilical cord length, distance from cord insertion to nearest margin, large diameter, small diameter, placental thickness) and placental weight. Associations between birthweight and placental measures were examined using multiple linear regression. Placental weight alone accounted for 36.6% of birthweight variation; the six other placental measures accounted for 28.1%. Combined, all placental measures accounted for 39.1% of birthweight variation. Seven maternal characteristics (age, height, weight, parity, socio-economic status, cigarette use, and race) were investigated to determine whether their known associations with birthweight were mediated by placental markers. Analysis suggested that the impact of all maternal characteristics except smoking was consistent

Birth Weight & Beta Value

$$\beta = \frac{\log PW^{*}}{\log BW}$$

$$\beta > 0.75$$

ow functional efficiency

$$\beta < 0.75$$

high functional efficiency

Birth Weight & Beta Value

Kleiber's Law:

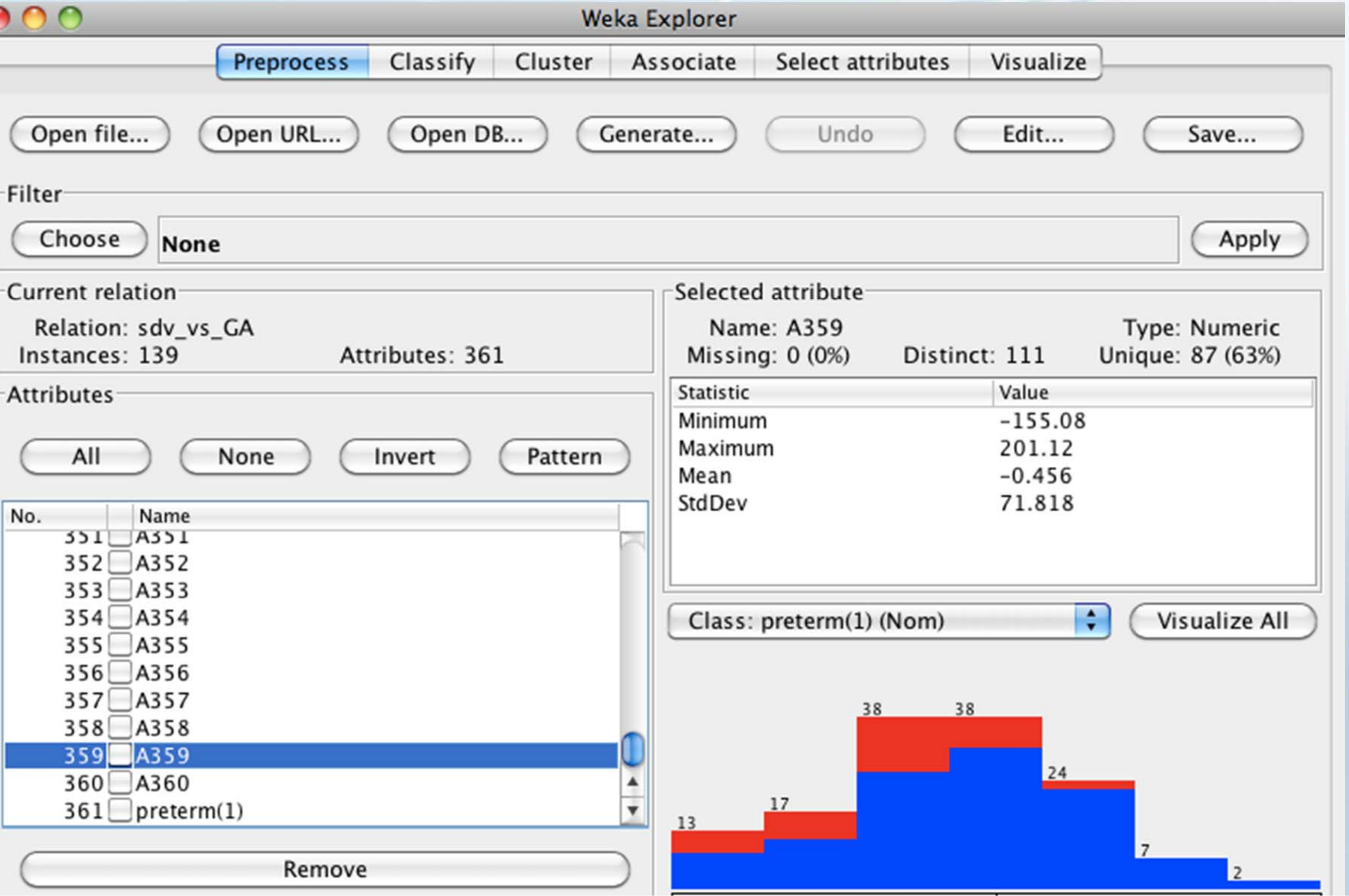
Metabolic rate (q_0) is proportional to body mass (M) raised to ¾ power

$$q_0 \sim M^{3/4}$$

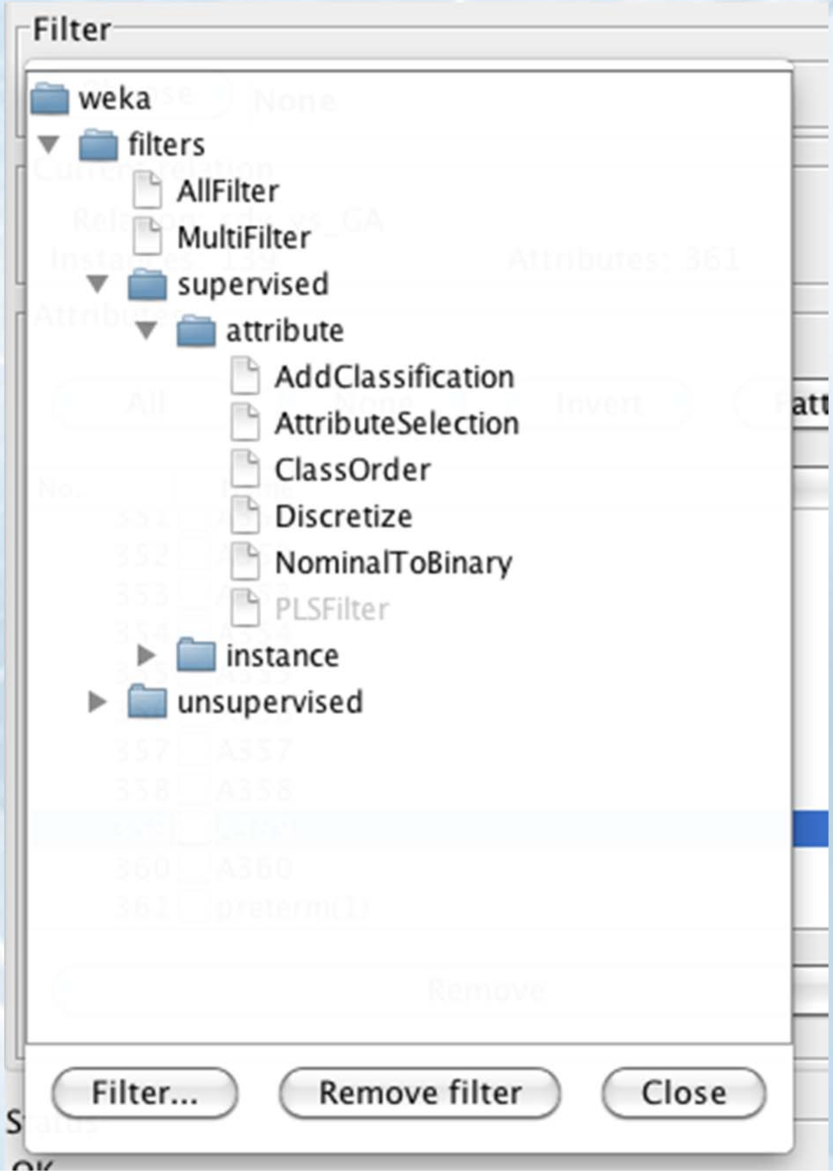
Learning Machine: Weka

Weka: http://www.cs.waikato.ac.nz/ml/weka/

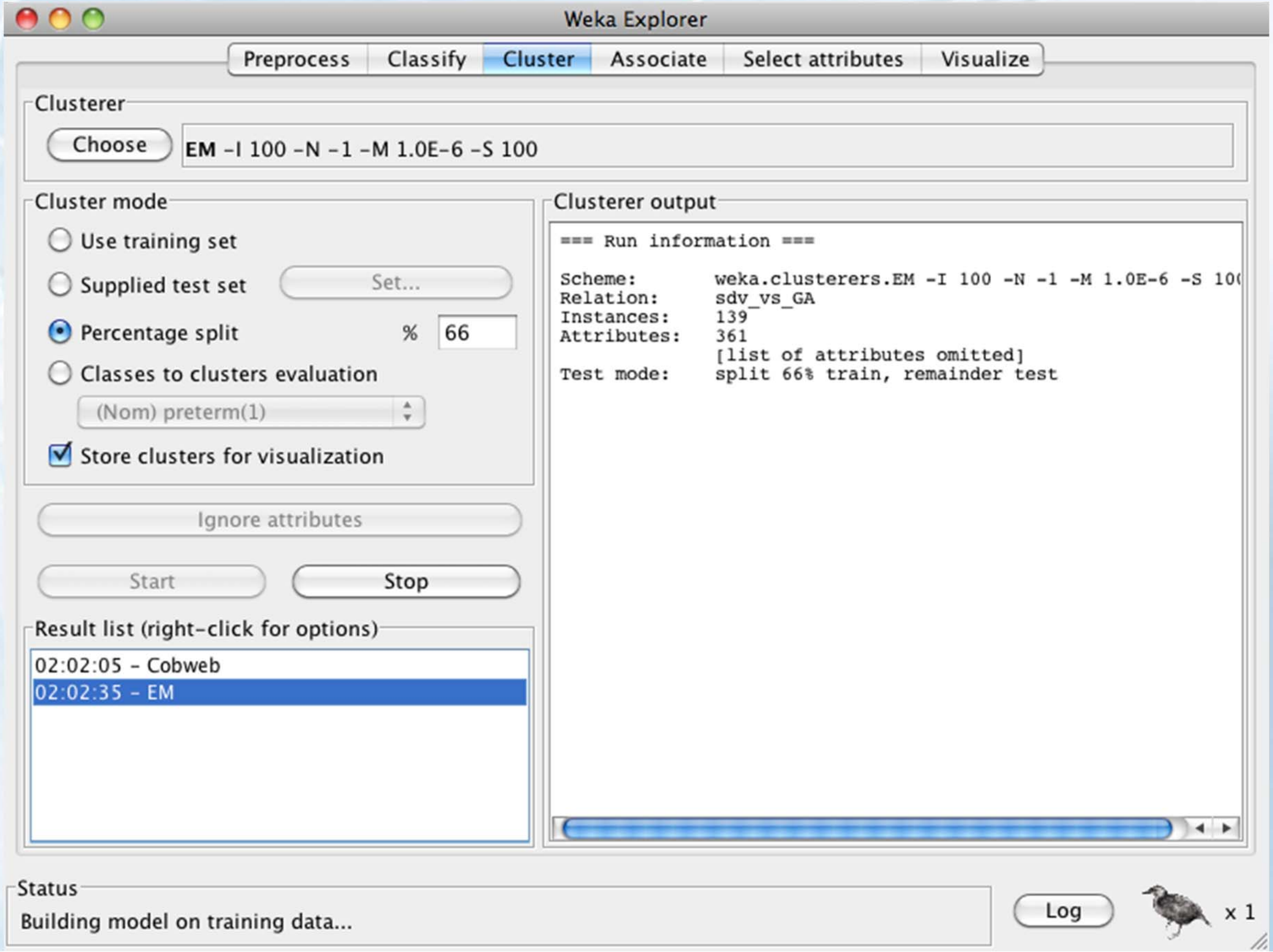
Weka


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: sdv_vs_GA
Instances: 139
Attributes: 361
Selected attribute
Name: A359
Type: Numeric
Missing: 0 (0%)
Distinct: 111
Unique: 87 (63%)
Statistic
Value
Minimum
-155.08
Maximum
201.12
Mean
-0.456
StdDev
71.818
Attributes
All
None
Invert
Pattern
No.
Name
352 A352
353 A353
354 A354
355 A355
356 A356
357 A357
358 A358
359 A359
360 A360
361 preterm(1)
Class: preterm(1) (Nom)
Visualize All
13
17
38
38
24
7
2
Remove

Weka

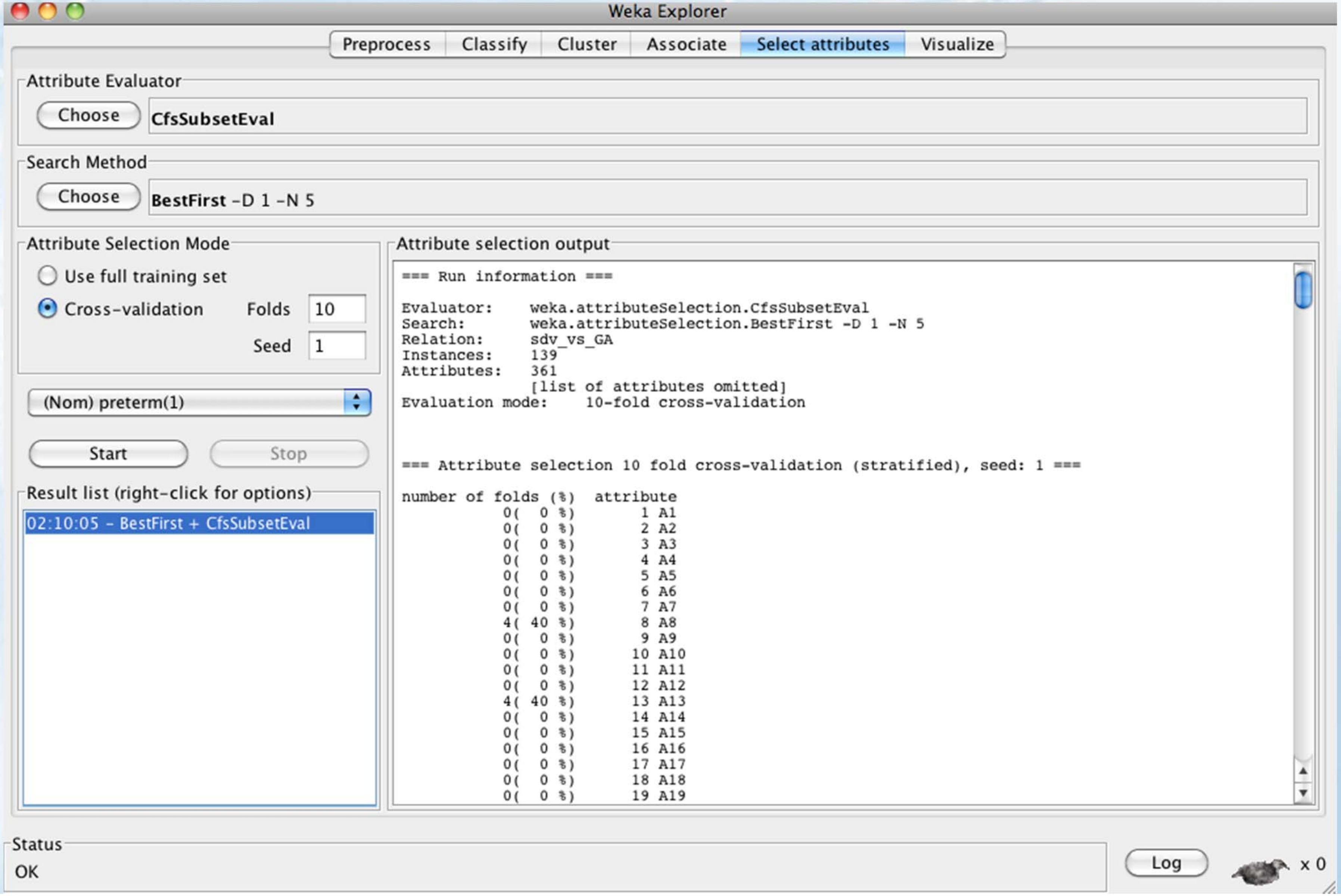

Filter
weka
filters
AllFilter
MultiFilter
supervised
attribute
AddClassification
AttributeSelection
ClassOrder
Discretize
NominalToBinary
PLSFilter
instance
unsupervised
Filter...
Remove filter
Close

Weka

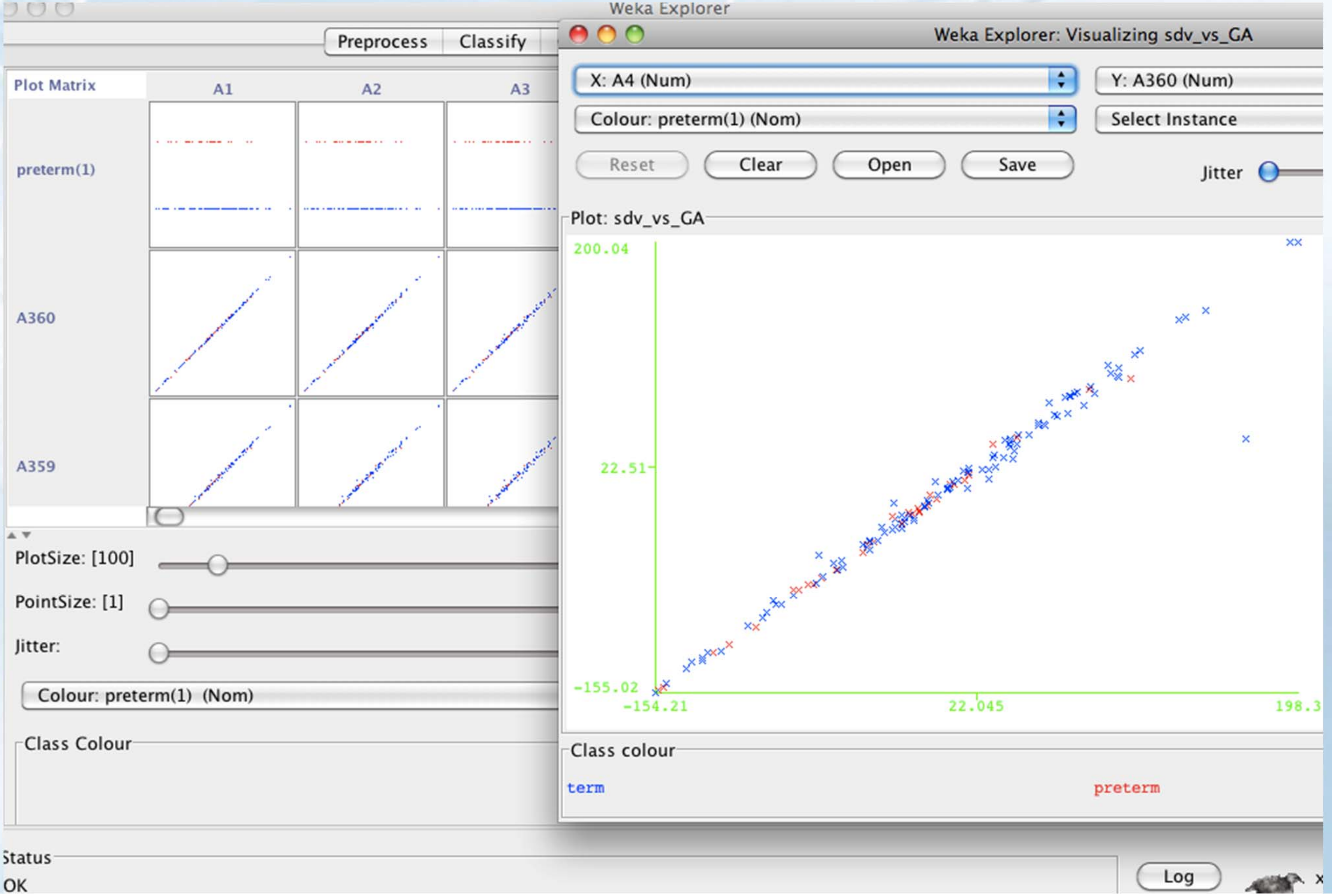


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
EM -I 100 -N -1 -M 1.0E-6 -S 100
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
% 66
Classes to clusters evaluation
(Nom) preterm(1)
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
02:02:05 - Cobweb
02:02:35 - EM
Clusterer output
=== Run information ===
Scheme: weka.clusterers.EM -I 100 -N -1 -M 1.0E-6 -S 10
Relation: sdv_vs_GA
Instances: 139
Attributes: 361
[list of attributes omitted]
Test mode: split 66% train, remainder test
Status
Building model on training data...
Log
x 1

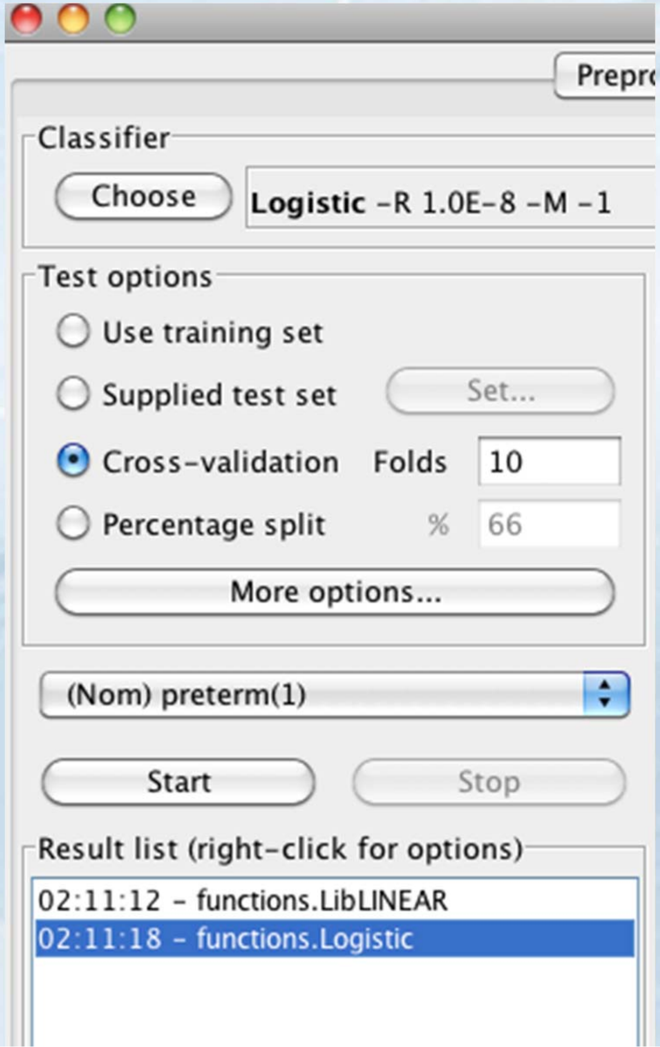
Weka


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
CfsSubsetEval
Search Method
Choose
BestFirst -D 1 -N 5
Attribute Selection Mode
Use full training set
Cross-validation
Folds 10
Seed 1
(Nom) preterm(1)
Start
Stop
Result list (right-click for options)
02:10:05 - BestFirst + CfsSubsetEval
Attribute selection output
=== Run information ===
Evaluator: weka.attributeSelection.CfsSubsetEval
Search: weka.attributeSelection.BestFirst -D 1 -N 5
Relation: sdv_vs_GA
Instances: 139
Attributes: 361
[list of attributes omitted]
Evaluation mode: 10-fold cross-validation
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===
number of folds (%) attribute
0(0 %) 1 A1
0(0 %) 2 A2
0(0 %) 3 A3
0(0 %) 4 A4
0(0 %) 5 A5
0(0 %) 6 A6
0(0 %) 7 A7
4(40 %) 8 A8
0(0 %) 9 A9
0(0 %) 10 A10
0(0 %) 11 A11
0(0 %) 12 A12
4(40 %) 13 A13
0(0 %) 14 A14
0(0 %) 15 A15
0(0 %) 16 A16
0(0 %) 17 A17
0(0 %) 18 A18
0(0 %) 19 A19
Status
OK
Log
x 0

Weka


Weka Explorer
Preprocess
Classify
Plot Matrix
A1
A2
A3
preterm(1)
A360
A359
PlotSize: [100]
PointSize: [1]
Jitter:
Colour: preterm(1) (Nom)
Class Colour
Status
OK
Log
Weka Explorer: Visualizing sdv_vs_GA
X: A4 (Num)
Y: A360 (Num)
Colour: preterm(1) (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: sdv_vs_GA
200.04
22.51
-155.02
-154.21
22.045
198.3
Class colour
term
preterm

Weka


Prepr
Classifier
Choose
Logistic -R 1.0E-8 -M -1
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) preterm(1)
Start
Stop
Result list (right-click for options)
02:11:12 - functions.LibLINEAR
02:11:18 - functions.Logistic

Naïve Bayesian Classifier

sing Conditional Probability using Bayes's Theorem

$$\ldots, F_n) = \frac{p(C)\ p(F_1,\dots,F_n|C)}{p(F_1,\dots,F_n)}.$$

r words:

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

ling the Conditional Probabilities

$$,\dots,F_n)$$

$$(C)\ p(F_1|C)\ p(F_2|C,F_1)\ p(F_3|C,F_1,F_2)\ \ldots p(F_n|C,F_1,F_2,F_3,\dots,F_{n-1}).$$

Naïve Bayesian Classifier

Independence Assumption:

$$(F_i|C, F_j) = p(F_i|C)$$

ying:

$$, F_1, \dots, F_n) = p(C)\, p(F_1|C)\, p(F_2|C)\, p(F_3|C) \cdots$$

$$= p(C) \prod_{i=1}^{n} p(F_i|C).$$

$$F_1, \dots, F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i|C)$$

Normalization constant (1/Z)
is the same for all C,
so we can ignore it

fication Rule

$$y(f_1, \dots, f_n) = \mathrm{argmax}\, p(C = c) \prod^{n} p(F_i = f_i|C = c).$$

Naïve Bayesian Classifier

```
from math import log, exp

class BayesianClassifier(object):
    def __init__(self):
        self.observation_total = 0
        self.observations = {}
        self.observations_dual = {}
        self.labels = {}
        self.feats = {}

    def train(self, event, evidence):
        self.observations.setdefault(event, {})
        self.labels.setdefault(event, 0)
        self.observation_total += 1
        self.labels[event] += 1
        for v in evidence:
            self.observations[event].setdefault(v, 0)
            self.observations_dual.setdefault(v, {})
            self.observations_dual[v].setdefault(event, 0)
            self.feats.setdefault(v, 0)
            self.observations[event][v] += 1
            self.feats[v] += 1
            self.observations_dual[v][event] += 1

    def classify(self, evidence, complement=False):
        estimates = [(self.cond_comp(event, evidence, complement), event)
                     for event in self.labels]
        highest_prob, likely_event = max(estimates)
        estimates = [(exp(prob - highest_prob), event)
                     for prob, event in estimates]
        highest_prob = 1 / sum(prob for prob, event in estimates)
        estimates = [(highest_prob * prob, event)
                     for prob, event in estimates]

        return likely_event, highest_prob

    def cond_comp(self, event, evidence, complement):
        if event not in self.labels:
            return 0
        else:
            #Using log probabilities to prevent underflow
            P_event = log(self.labels[event]) - log(self.observation_total)
            P_condt = 0
            for Bk in evidence:
                if Bk not in self.feats: continue
                P_condt += ( log(self.observations[event].get(Bk,0) + 1) -
                             log(self.labels[event]) )
            if complement:
                for Bk in self.feats:
                    if Bk in evidence: continue
                    P_condt += ( log(self.labels[event] -
                                     self.observations[event].get(Bk, 0) + 1) -
                                 log(self.labels[event]) )
            return P_event + P_condt
```

Questions? Comments?