CLINICAL INTERPRETATION OF OMICS CLUSTERING RESULTS

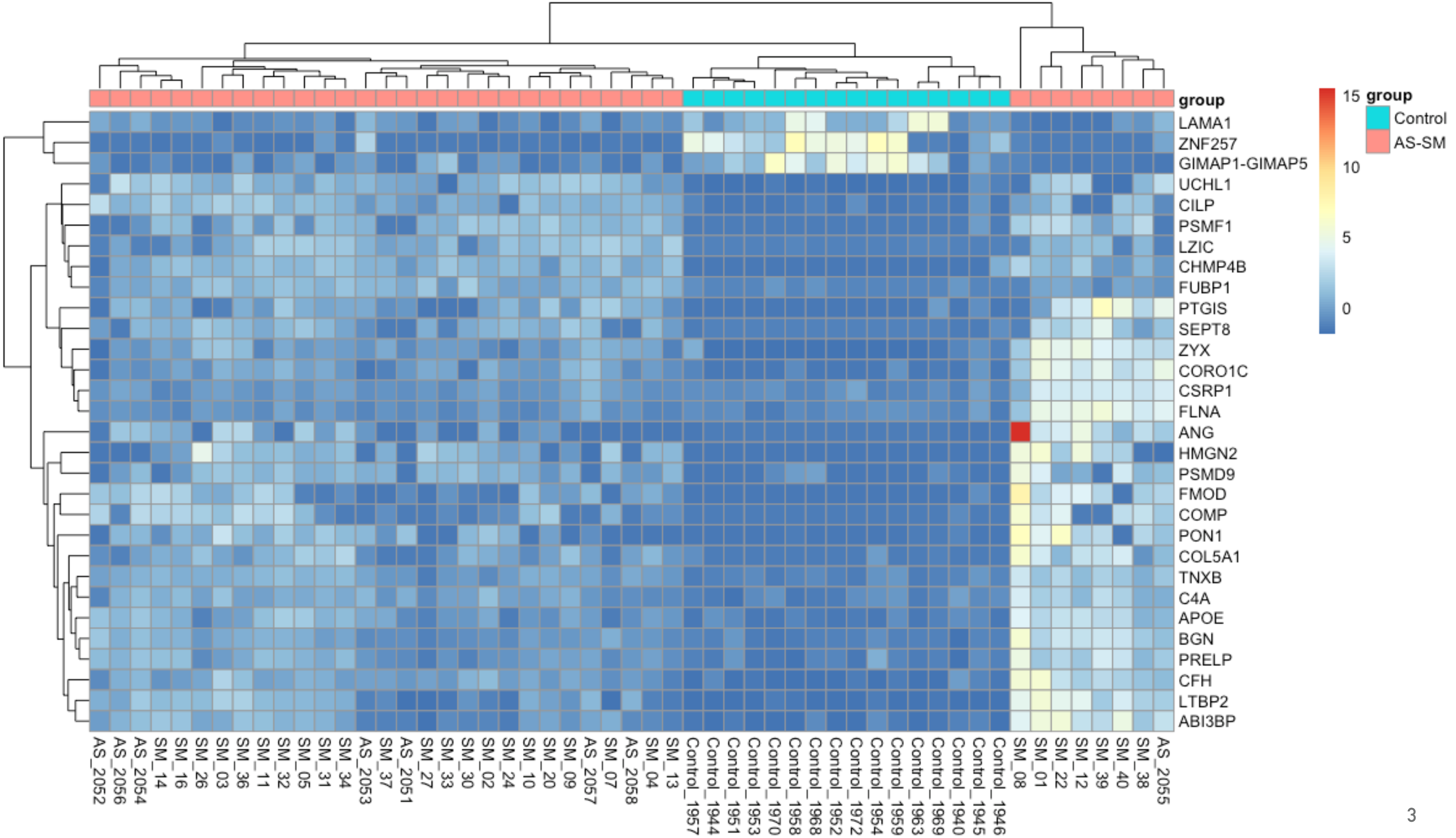
Trends in Bioinformatics WS 17/18

Supervisor: Milena Kraus

Ajay Kesar

AGENDA

- Task
- Data
- Feature Selection Models
  - Wrapper / Embedded / Filter
- Wrapper: Decision Tree
- Embedded: normal Logistic Regression with L1 & L2 regularization, Randomized Logistic Regression with L1 regularization
- Filter: Pearson's Correlation
- Future Work


group
LAMA1
ZNF257
GIMAP1-GIMAP5
UCHL1
CILP
PSMF1
LZIC
CHMP4B
FUBP1
PTGIS
SEPT8
ZYX
CORO1C
CSRP1
FLNA
ANG
HMGN2
PSMD9
FMOD
COMP
PON1
COL5A1
TNXB
C4A
APOE
BGN
PRELP
CFH
LTBP2
ABI3BP
15
10
5
0
group
Control
AS-SM
AS_2052
AS_2056
AS_2054
SM_14
SM_16
SM_26
SM_03
SM_36
SM_11
SM_32
SM_05
SM_31
SM_34
AS_2053
SM_37
AS_2051
SM_27
SM_33
SM_30
SM_02
SM_24
SM_10
SM_20
SM_09
AS_2057
SM_07
AS_2058
SM_04
SM_13
Control_1957
Control_1944
Control_1951
Control_1953
Control_1970
Control_1958
Control_1968
Control_1952
Control_1972
Control_1954
Control_1959
Control_1963
Control_1969
Control_1940
Control_1945
Control_1946
SM_08
SM_01
SM_22
SM_12
SM_39
SM_40
SM_38
AS_2055

A	B	C	D	E	F	G	H
	PatientID	Category	ParameterName	Event	Value	Unit	Unnamed: 6
0	SM 01	Demographics / Clinical Parameters	Gender	General	female		
1	SM 01	Demographics / Clinical Parameters	Study group	General	No cardiac medication		
2	SM 01	General	Time between VISIT 1 and VISIT 2	General	223		
3	SM 01	General	Time between VISIT 1 and surgery	General	3		
4	SM 01	General	Time between surgery and VISIT 2	General	220		
5	SM 01	Surgery Parameters	Age at Surgery	Surgery	69	years	
6	SM 01	Surgery Parameters	Aortic Valve Replacement	Surgery	yes		
7	SM 01	Surgery Parameters	MIC	Surgery	yes		
8	SM 01	Surgery Parameters	Additional Surgery	Surgery	no		
9	SM 01	Surgery Parameters	CABG Surgery	Surgery	no		
10	SM 01	Surgery Parameters	Other Surgery	Surgery	no		
11	SM 01	Surgery Parameters	Aortic Valve Size	Surgery	23	mm	
12	SM 01	Surgery Parameters	Biological Aortic Valve?	Surgery	yes		
13	SM 01	Surgery Parameters	Modell of Aortic Valve	Surgery	CE Perimount Magna Ease 23mm, 3300 TFX		
14	SM 01	Surgery Parameters	Aortic Clamping Time	Surgery	87	min	
15	SM 01	Surgery Parameters	Perfusion Time	Surgery	122	min	
16	SM 01	Surgery Parameters	Reperfusion Time	Surgery	18	min	
17	SM 01	Surgery Parameters	Cardiac Arrest Time	Surgery	87	min	
18	SM 01	Surgery Parameters	Biopsy	Surgery	yes		
19	SM 01	Surgery Parameters	Postoperative Pacemaker	Surgery	no		
20	SM 01	Catheter Measurement	Left ventricular systolic pressure	General	230	mm[Hg]	2015-03-18
21	SM 01	Catheter Measurement	Left ventricular end diastolic pressure	General	15	mm[Hg]	2015-03-18
22	SM 01	Catheter Measurement	Pressure in the ascending aorta	General	180	mm[Hg]	2015-03-18
23	SM 01	Demographics / Clinical Parameters	Height	VISIT 1	164	cm	
24	SM 01	Demographics / Clinical Parameters	Weight	VISIT 1	71	kg	
25	SM 01	Demographics / Clinical Parameters	NYHA stage	VISIT 1	2		
26	SM 01	Demographics / Clinical Parameters	Blood pressure systolic right arm	VISIT 1	164	mm[Hg]	
27	SM 01	Demographics / Clinical Parameters	Blood pressure diastolic right arm	VISIT 1	80	mm[Hg]	
28	SM 01	Demographics / Clinical Parameters	Blood pressure mean right arm	VISIT 1	122	mm[Hg]	
29	SM 01	Demographics / Clinical Parameters	Blood pressure systolic left arm	VISIT 1	164	mm[Hg]	
30	SM 01	Demographics / Clinical Parameters	Blood pressure diastolic left arm	VISIT 1	80	mm[Hg]	
31	SM 01	Demographics / Clinical Parameters	Blood pressure mean left arm	VISIT 1	122	mm[Hg]	
32	SM 01	Demographics / Clinical Parameters	Time of blood pressure	VISIT 1	5	pm	
33	SM 01	Demographics / Clinical Parameters	Heart rate	VISIT 1	76	per minute	
45	SM 01	Diagnoses	Diagnoses		AS III		
46	SM 01	Diagnoses	Diagnoses		Hypercholesterinemia, (fam.)		
47	SM 01	Diagnoses	Diagnoses		Gastritis (Autoimmune)		
48	SM 01	Diagnoses	Diagnoses		Hysterectomy 1990		
49	SM 01	Diagnoses	Diagnoses		Ovarectomy 2010		
50	SM 01	Diagnoses	Diagnoses		Vitamin B 12 anemia		
51	SM 01	Diagnoses	Diagnoses		hypertension		
52	SM 01	Diagnoses	Diagnoses		Arterial hypertension		
53	SM 01	Diagnoses	Diagnoses		Bicuspid aortic valve		

DATA

- 29 Patients
  - 33 common features
  - 199 different features in total

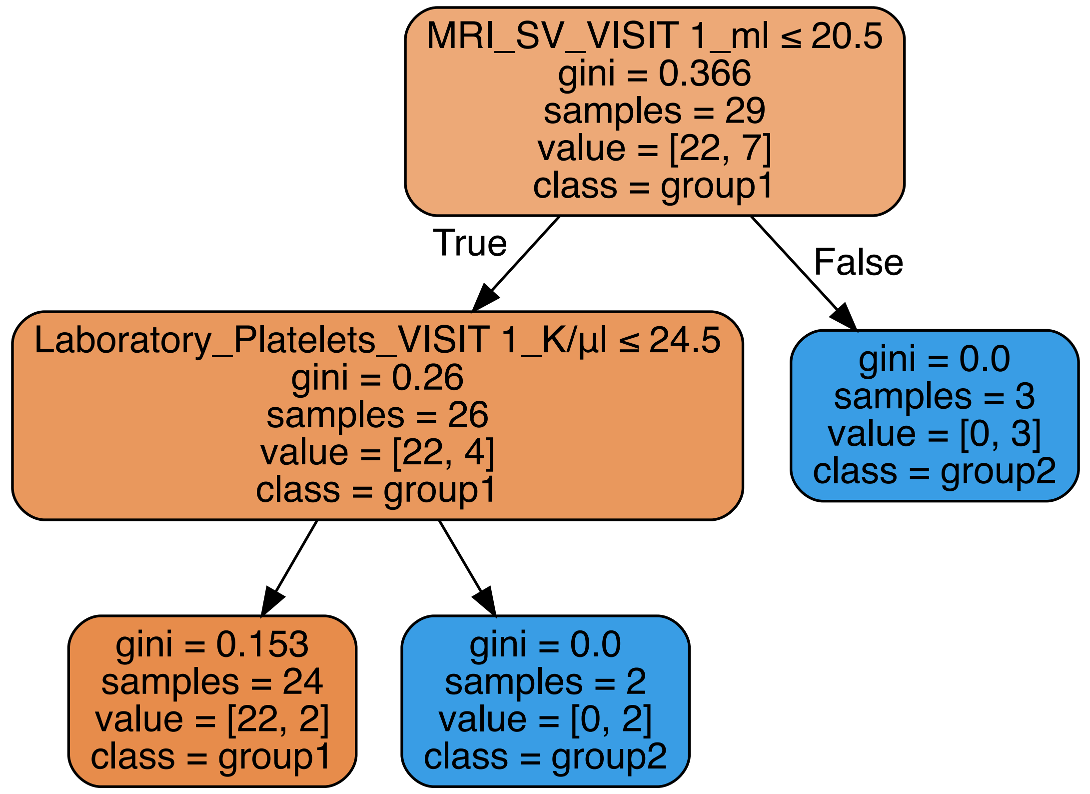
- Feature selection for heterogeneous data to support interpretation

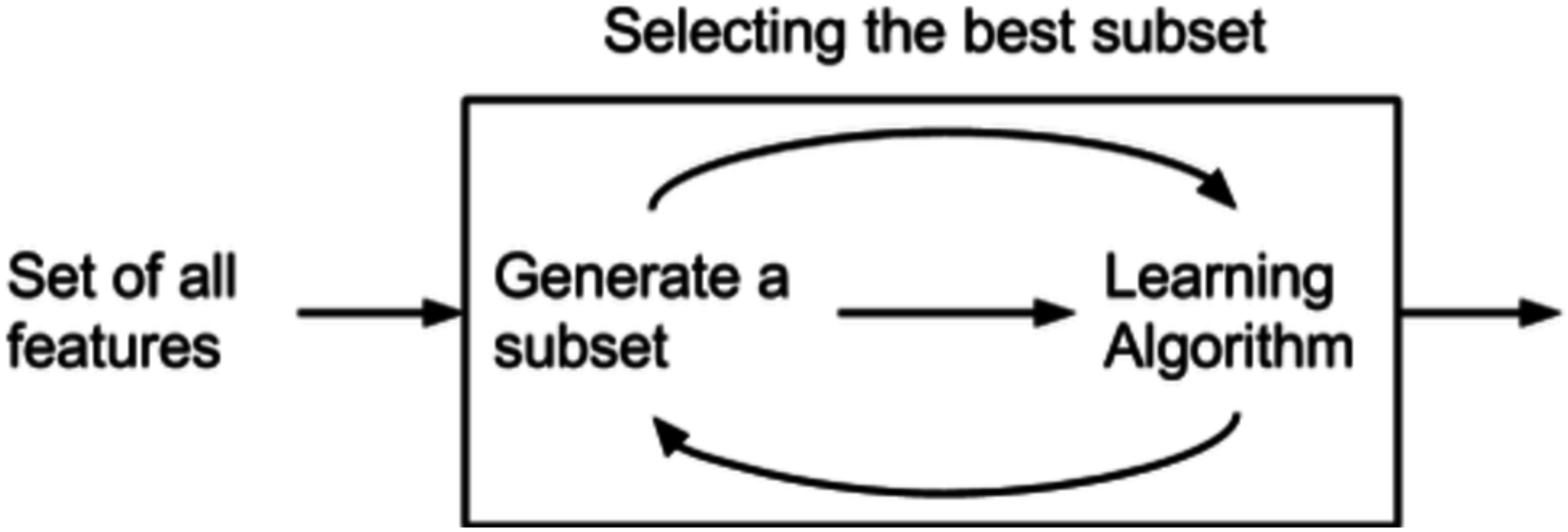
FEATURE SELECTION MODELS

1. Wrapper
2. Embedded
3. Filter

FEATURE SELECTION MODELS - WRAPPER

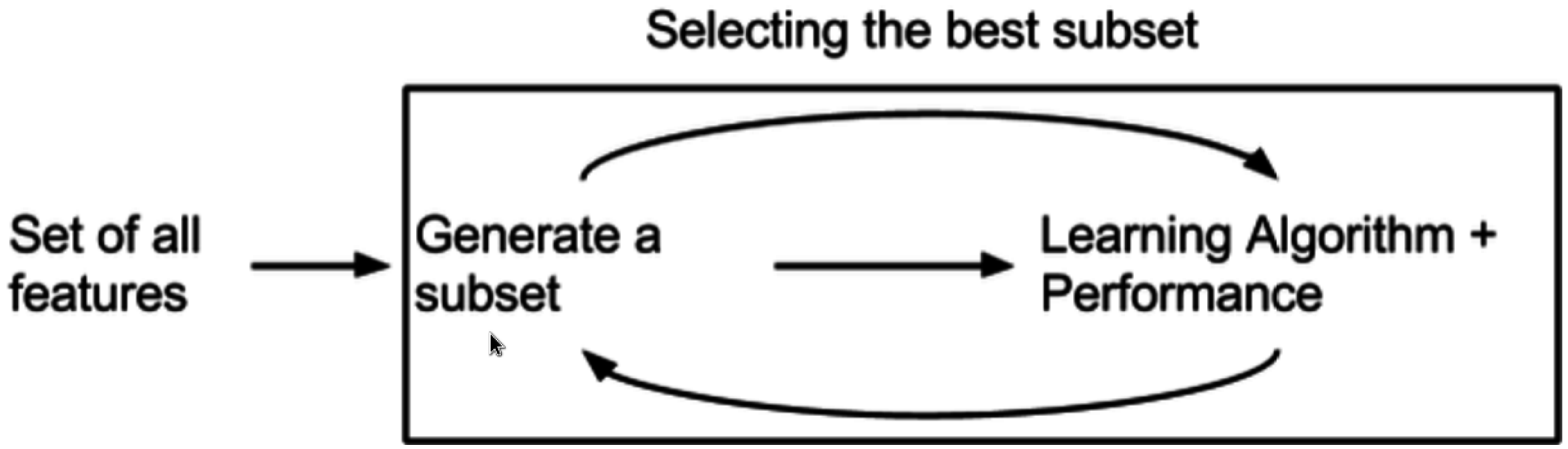
- Use a predictive model to evaluate the relative usefulness of parameter subsets
- a) how to search the space of all possible parameter subsets
- b) how to evaluate the prediction performance
- c) predictive model
- Tend to be computationally expensive
- Decision Tree





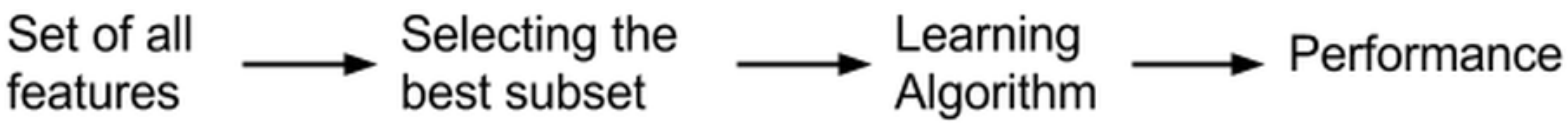
FEATURE SELECTION MODELS - EMBEDDED

- Select features using the information gained from training a learning algorithm instead of treating it as a black box
  - Lasso & Ridge regression
  - Normal Logistic Regression with L1 & L2 regularization, Randomized Logistic Regression with L1 regularization



FEATURE SELECTION MODELS - FILTER

- Filters select features based on criteria independent of any supervised learning algorithm, performance of filters may not be optimal for a chosen learning algorithm
  - Pearson’s Correlation
  - ANOVA: Analysis of variance



WRAPPER

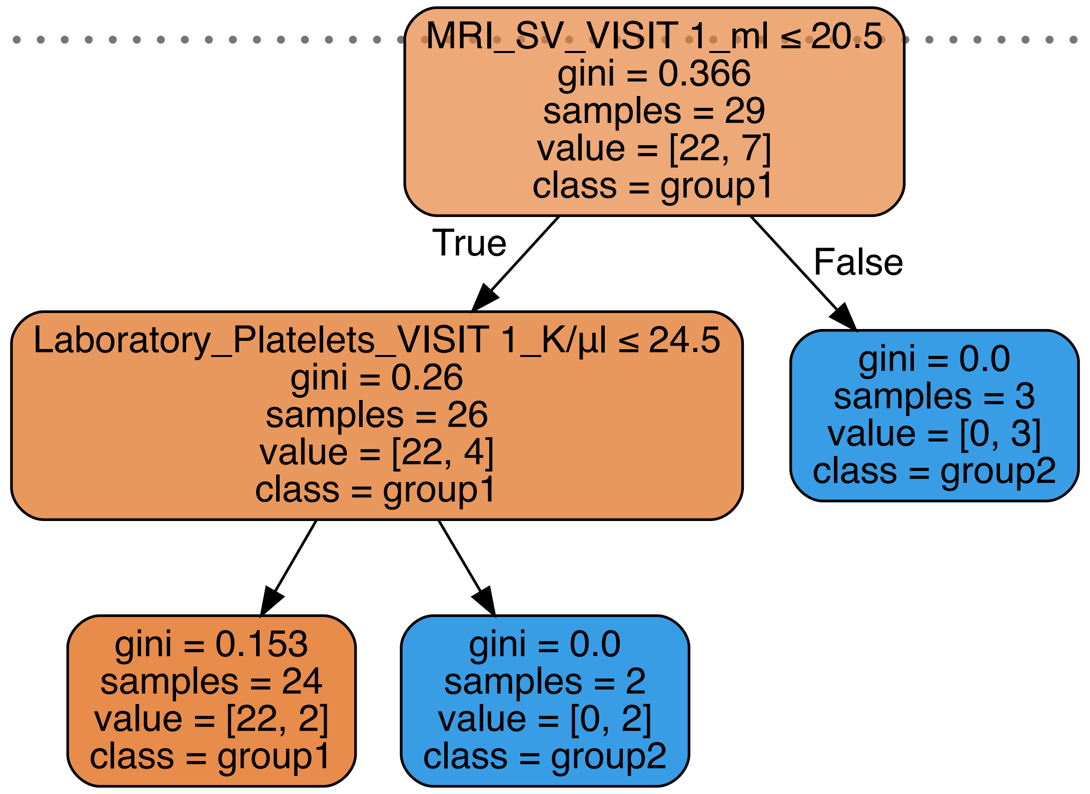
DECISION TREES

- can handle:
  - heterogeneous data
  - missing values
  - different parameter scales
  - nonlinearities
- easily interpretable

SCIKIT-LEARN - DATA PREPARATION

- Problem: estimators assume all values to be numerical
- encode categorical values
- discard entire incomplete categories, focus on complete ones
- potential imputation strategies:
    - mean
    - median
    - most frequent value in row/column
    - regression

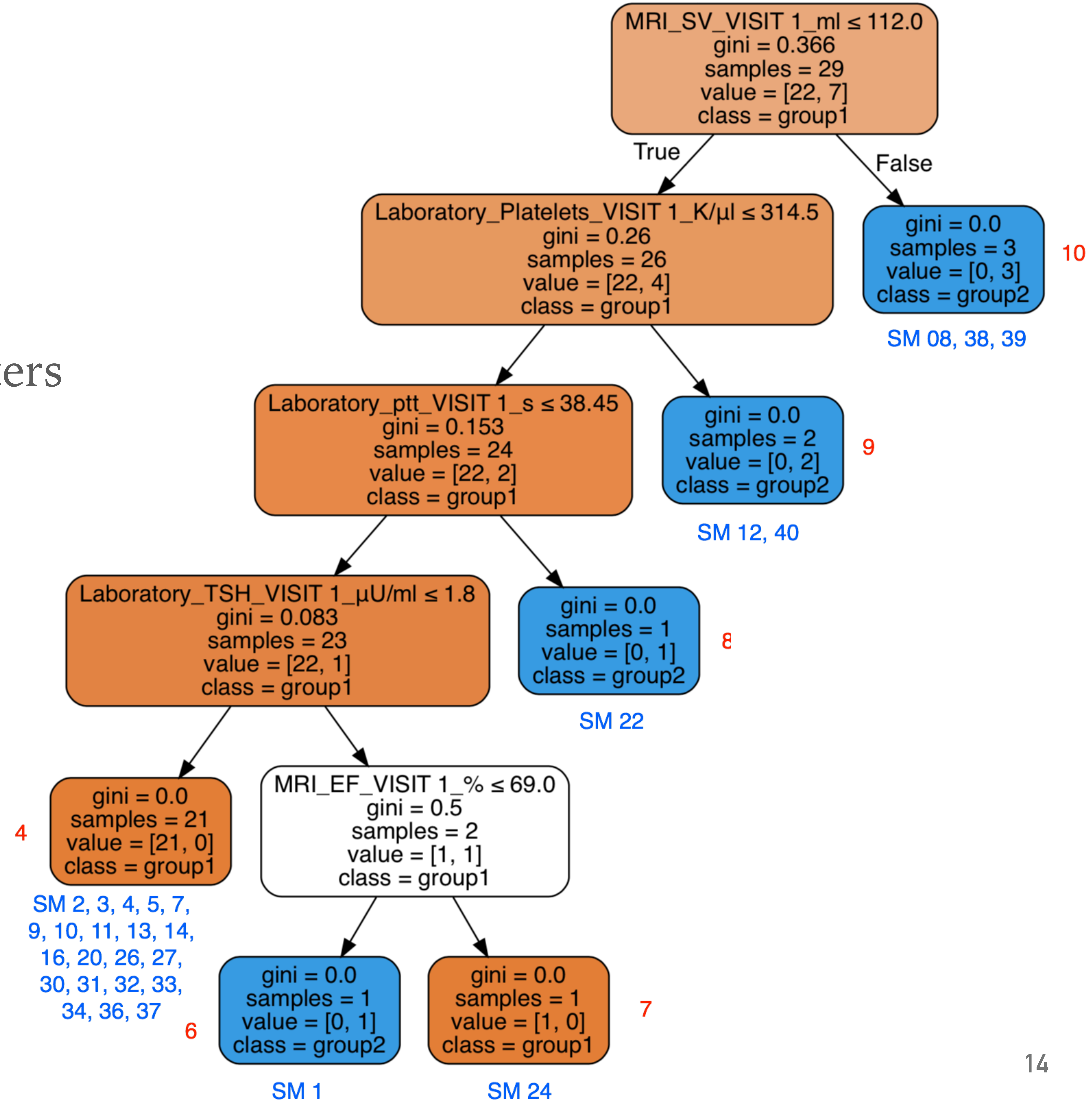
DECISION TREES



- in scikit-learn: CART algorithm
- a non-parametric DT learning technique
- DT are formed by collection of rules:
  - Rules are selected to get the best split to differentiate observations
  - Once a rule is selected & splits a node into two, same process is applied to each child node (recursive)
  - Splitting stops when no further gain can be made or some pre-set stopping rules are met
  - Alternatively, data are split as much as possible & then the tree is later pruned

DECISION TREE RESULTS

- Gini criterion
- Fits perfectly with molecular clusters



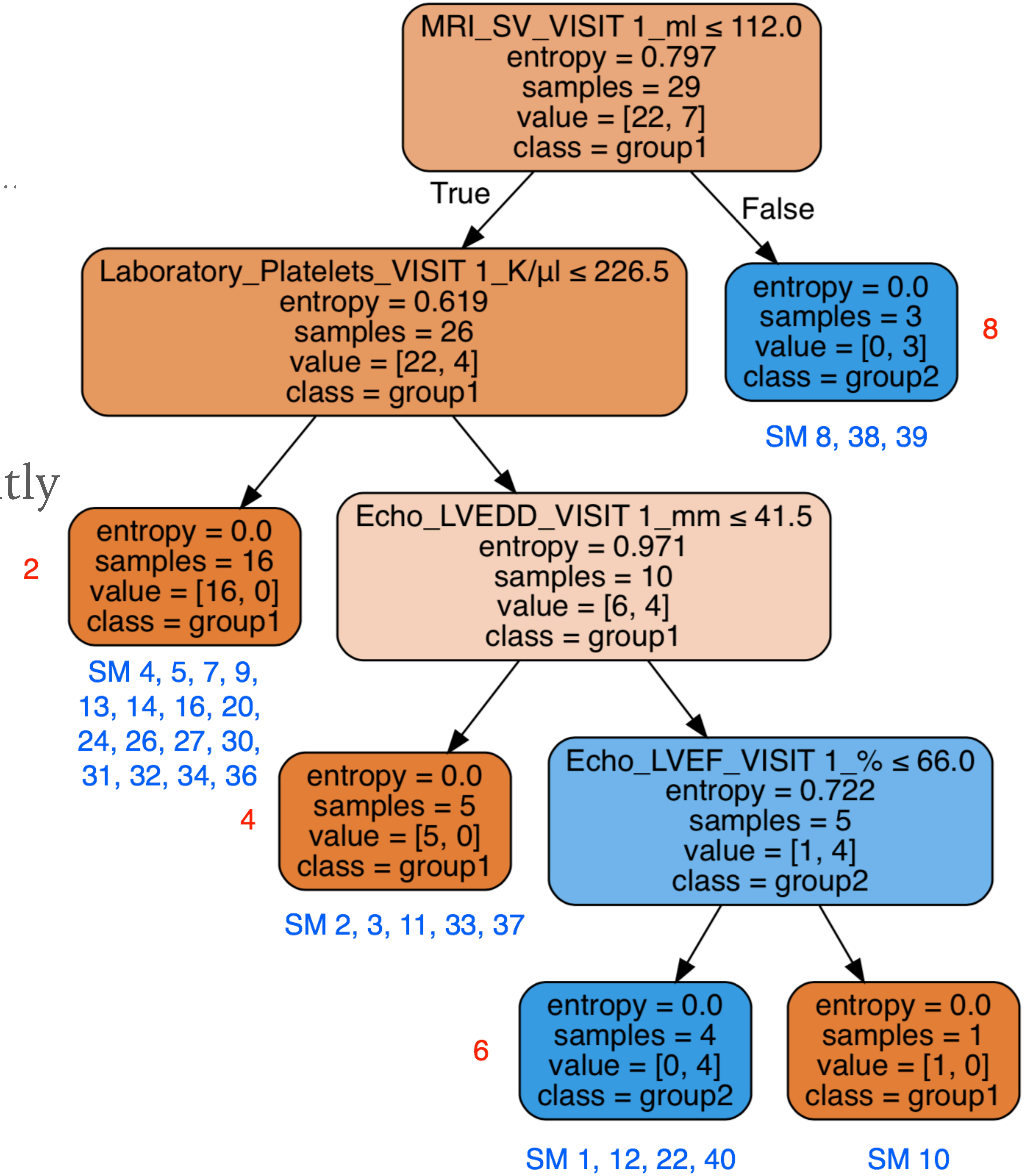
DECISION TREE FEATURE IMPORTANCE

- Out of 33 categories
- Only 5 non-zero

Feature	Feature importance
Laboratory_TSH_VISIT 1_μU/ml	0.08596838
MRI_EF_VISIT 1_%	0.09415584
Laboratory_ptt_VISIT 1_s	0.16511387
Laboratory_Platelets_VISIT 1_K/μl	0.29212454
MRI_SV_VISIT 1_ml	0.36263736

DECISION TREE RESULTS

- Entropy criterion
- fits perfectly with molecular clusters
- Split node only if
  information gain increases significantly
- Lower redundancy
- Higher efficiency

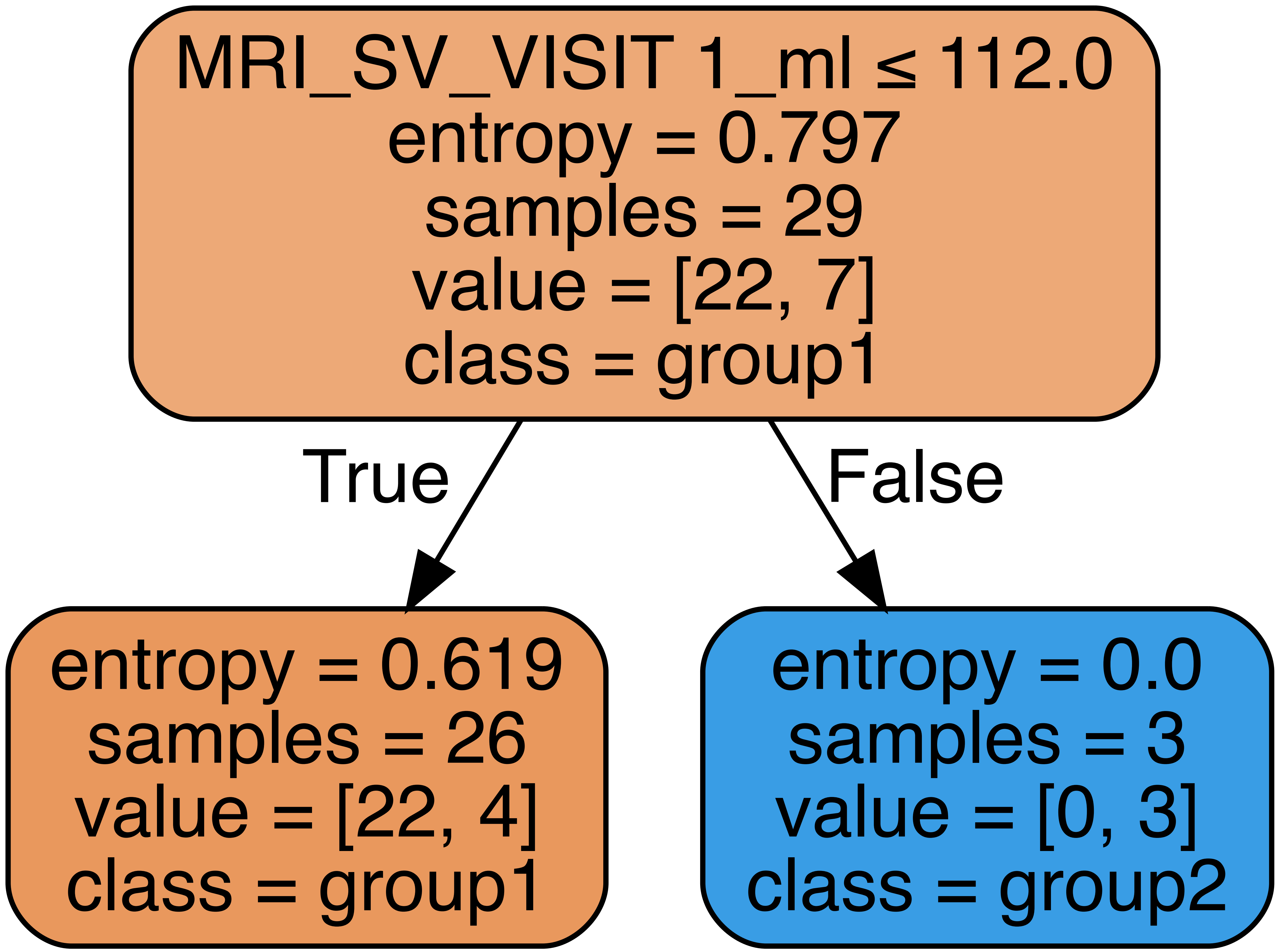


DECISION TREE FEATURE IMPORTANCE

- out of 33 categories
- only 4 non-zero

Feature	Feature importance
Echo_LVEF_VISIT 1_%	0.15610965
Echo_LVEDD_VISIT 1_mm	0.26380684
Laboratory_Platelets_VISIT 1_K/μl	0.27654621
MRI_SV_VISIT 1_ml	0.3035373

DECISION TREE



EMBEDDED

LOGISTIC REGRESSION WITH REGULARIZATION

- Logistic Regression is a linear prediction model for classification
- Optimization problem trying to fit the data
- Regularization controls and reduces overfitting
- Regularization adds bias towards particular values e.g. near zero
- Helps to have a small error and small parameter values
- $y = m*x + b + e$

L1 & L2 REGULARIZATION

- used to reduce model overfitting for better predictions
- L1 weight regularization penalizes feature weight values by adding the sum of their absolute values to error term: e = $Sum(Abs(w))$
- L1 regularization has built-in feature selection, prunes unneeded features by setting associated weights to zero

- L2 weight regularization penalizes feature weight values by adding the sum of their squared values to error term: e = $Sum(w \hat{} 2)$
- L2 regularization works with all forms of training algorithms, doesn't provide feature selection

NORMAL LOGISTIC REGRESSION WITH L1 REGULARIZATION

➤ out of 33 categories

Feature	Feature coefficients
Laboratory_Platelets_VISIT 1_K/µl	0.11179517
Echo_LVEDD_VISIT 1_mm	0.11617795
MRI_EDV_VISIT 1_ml	0.15542398
Surgery Parameters_Age at Surgery_Surgery_years	0.31858142
MRI_SV_VISIT 1_ml	0.37391707

Feature	Feature coefficients
Laboratory_tpz_VISIT 1_%	-0.22774189
Demographics / Clinical Parameters_Height_VISIT 1_cm	-0.2506049
Laboratory_White cell blood count_VISIT 1_K/µl	-0.34042031
MRI_EDVi_VISIT 1_ml/m²	-0.54855701

NORMAL LOGISTIC REGRESSION WITH L2 REGULARIZATION

- out of 33 categories
- 8 significant

Feature	Feature importance
Echo_LVEDD_VISIT 1_mm	0.3479861
MRI_SV_VISIT 1_ml	0.40306596
Surgery Parameters_Age at Surgery_Surgery_years	0.43264012

Feature	Feature importance
Laboratory_ptt_VISIT 1_s	-0.26775122
Laboratory_White cell blood count_VISIT 1_K/µl	-0.3192058
MRI_EDVi_VISIT 1_ml/m²	-0.35991696
Laboratory_tpz_VISIT 1_%	-0.36770301
Demographics / Clinical Parameters_Height_VISIT 1_cm	-0.38427853

RANDOMIZED LOGISTIC REGRESSION WITH L1 REGULARIZATION

- Stability Selection
- Feature selection on different subsets of data & features
- After several runs, the selection results are aggregated and features are ranked by frequency
- Using scikit-learn's Randomized Logistic Regression, implements stability selection by subsampling the training data and fitting a L1-penalized logistic regression model
- the method assigns high scores to features that are repeatedly selected across randomizations
- In short, features selected more often are considered good features

RANDOMIZED LOGISTIC REGRESSION WITH L1 REGULARIZATION

- out of 33 categories
- only 5 non-zero

Feature	Feature importance
MRI_CI_VISIT 1_l/min/m²	0,01
Laboratory_Platelets_VISIT 1_K/µl	0,015
MRI_CO_VISIT 1_l/min	0,015
MRI_SVi_VISIT 1_ml/m²	0,015
MRI_SV_VISIT 1_ml	0,0125

FILTER

PEARSON CORRELATION

- Correlation quantifies the relationship between two variables
- Pearson's correlation coefficient is a measure of the strength of the relationship between two continuous variables
- For linear relationships
- range from -1 to +1, 0 meaning no correlation
- significant if <-0.25 or > 0.25
- detect which feature pairs are strongly connected for patients

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

FUTURE WORK

- Pearson Correlation
- Evaluate all applied feature importance results
- Evaluation of results with clinician

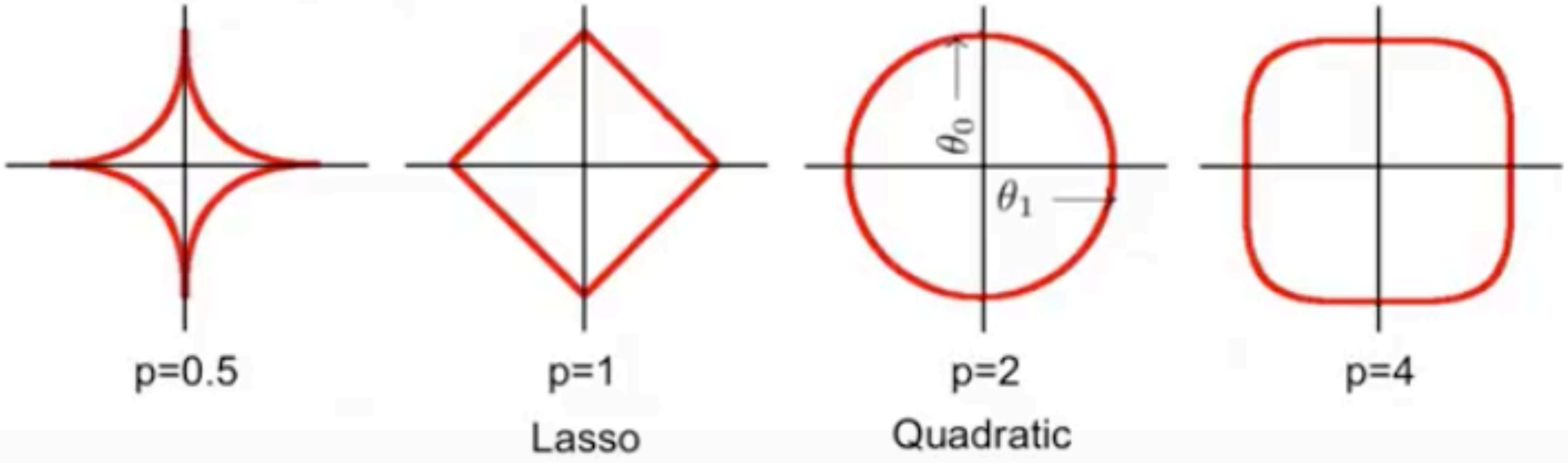
REFERENCES

- Trees: http://scikit-learn.org/stable/modules/tree.html
- Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." Journal of machine learning research 3.Mar (2003): 1157-1182.
- Deng, Houtao, and George Runger. "Feature selection via regularized trees." Neural Networks (IJCNN), The 2012 International Joint Conference on. IEEE, 2012.
- http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RandomizedLogisticRegression.html
- http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- https://msdn.microsoft.com/en-us/magazine/dn904675.aspx

REGULARIZATION

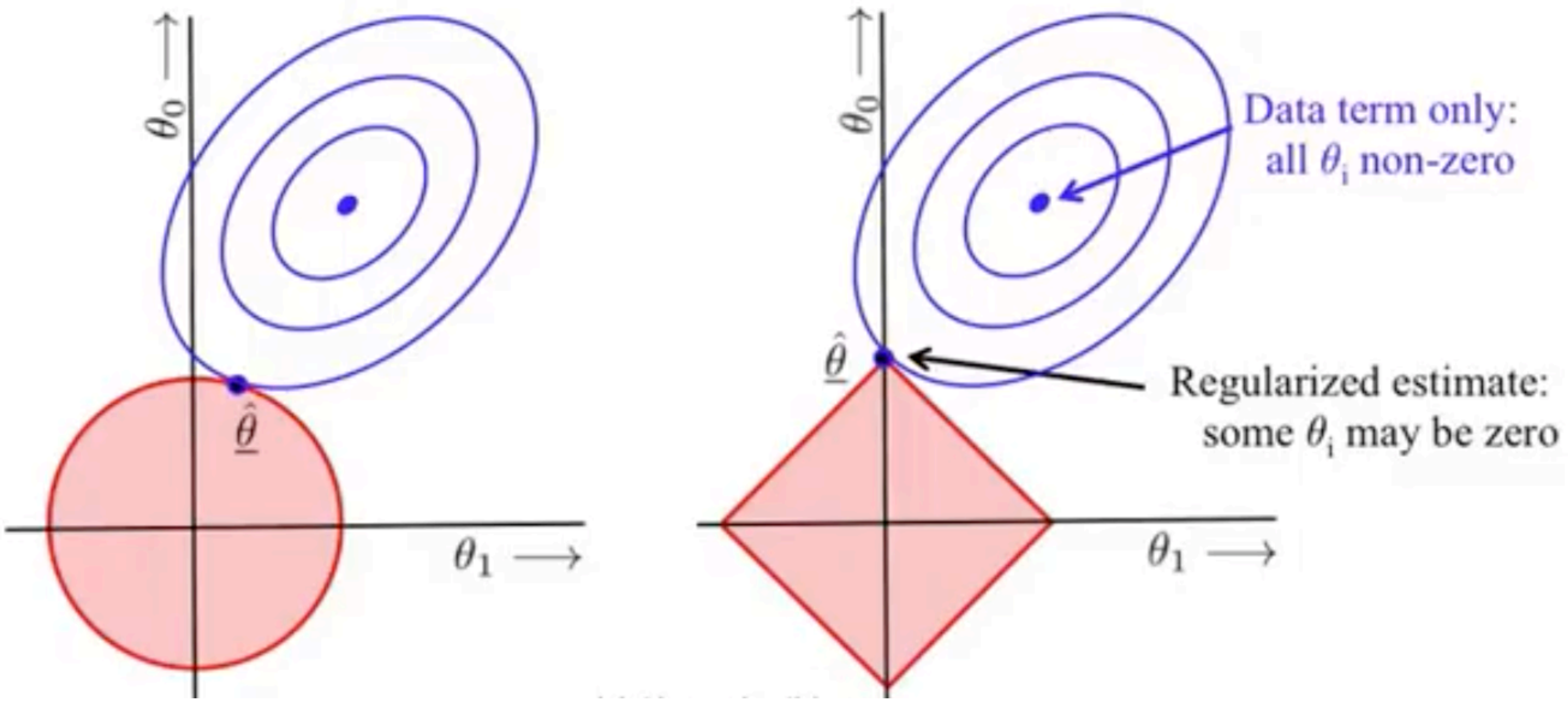
L_p regularizer: $\left(\sum_i |\theta_i|^p \right)^{\frac{1}{p}}$

Isosurfaces: $||\theta||_p = \text{constant}$



REGULARIZATION

- Estimate balances data term & regularization term
- Lasso tends to generate sparser solutions than a quadratic regularizer.



L1 PENALTY

- L1 weight regularization penalizes weight values by adding the sum of their absolute values to the error term

```
double sumAbsVals = 0.0; // L1 penalty
for (int i = 0; i < weights.Length; ++i)
  sumAbsVals += Math.Abs(weights[i]);
```

L2 PENALTY

- L2 weight regularization penalizes weight values by adding the sum of their squared values to the error term

```
double sumSquaredVals = 0.0; // L2 penalty
for (int i = 0; i < weights.Length; ++i)
  sumSquaredVals += (weights[i] * weights[i]);
```

RANDOMIZED LOGISTIC REGRESSION WITH L1 REGULARIZATION

- Randomized Logistic Regression works by subsampling the training data and fitting a L1-penalized Logistic Regression model where the penalty of a random subset of coefficients has been scaled. By performing this double randomization several times, the method assigns high scores to features that are repeatedly selected across randomizations. This is known as stability selection. In short, features selected more often are considered good features.