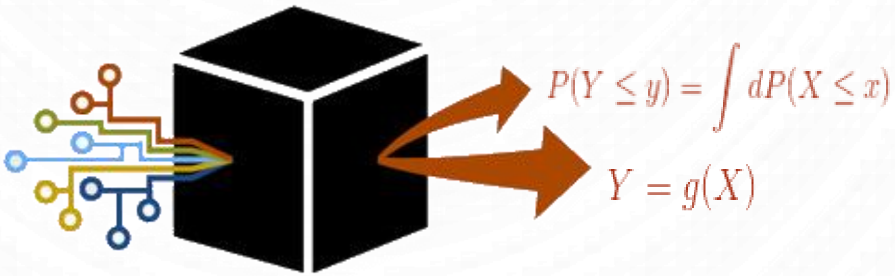
A HYBRID APPROACH TO MODEL RANDOMNESS AND FUZZINESS USING DEMPSTER-SHAFFER METHOD



Taposh Roy & Austin Powell
July 2016

WHO ARE WE ?

Lead Big Data & Data Science
Kaiser Permanente
taposh.d.roy@kp.org

Graduate Intern
Kaiser Permanente
austin.powell@kp.org

OUTLINE

- Motivation, philosophy and some definitions
- Why Dempster Shafer method
- Math behind Dempster Shafer method.
- Our analysis combining predictability with Dempster Shafer.
- Review of our results and interpretation

MOTIVATION

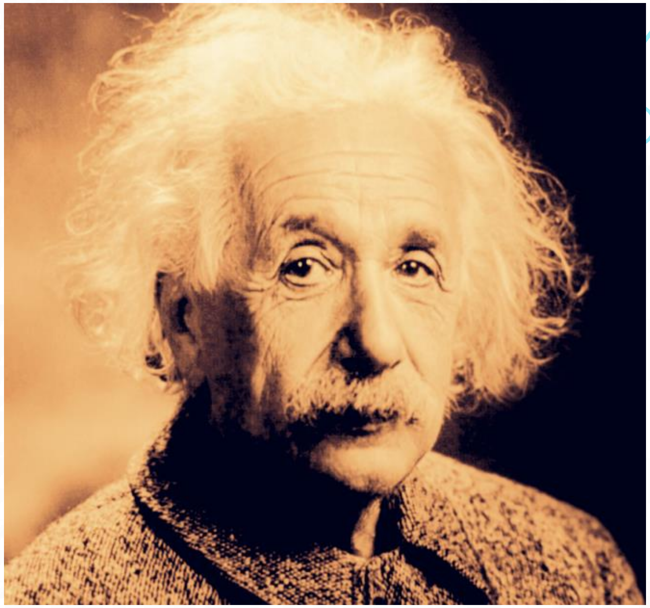
THE WHY?

Majority of predictive analytics today is based on historic data. As data scientists, we hunt and gather historic data, train to generate models using variety of algorithms, test and validate it and finally deploy to score.

This however limits us to situations that have been considered in our training data. Our predictions don't incorporate any external or environmental factors not already trained.

PHILOSOPHY

So far as the laws of mathematics refer to reality, they are not certain. And so far as they are certain, they do not refer to reality.



As complexity rises, precise statements lose meaning and meaningful statements lose precision.

BASIC DEFINITIONS

DEFINITION : RANDOMNESS

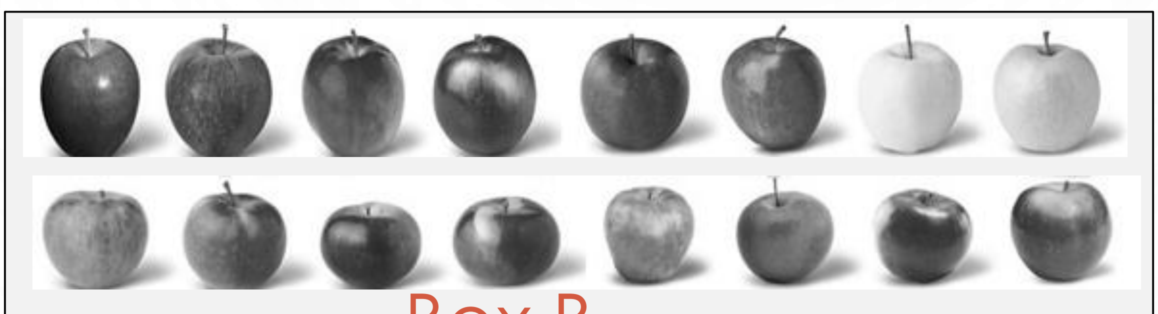
Box A contains 8 small and 8 big apples. You randomly pick an apple from the box. There is uncertainty about the size of the apple that will come out. The size of the apple can be described by a random variable that takes 2 different values.

The probability of selected apple being big or small is : **50 %**

There are 2 distinct possible outcomes and there is no uncertainty after the apple is taken out and observed.

DEFINITION : FUZZINESS

Now, you have a box (Box B) that contains 16 apples with different sizes ranging from very small to very large. You have one of these apples in hand. There is uncertainty in describing the size of the apple but this time the uncertainty is not about the lack of knowledge of the outcome. It is about the lack of the boundary between small and large. You use a fuzzy membership function to define the membership value of the apple in the set ``large`` or ``small``. Everything is known except how to label the objects (apples), how to describe them and draw the boundaries of the sets.



Box B

WHY DEMPSTER SHAFER?

DEMPSTER SHAFFER – WHY?

- Dempster Shafer theory was used in AI and reliability modeling[1].
- Dempster Shafer led the path to Fuzzy logic. It is a base theory for uncertainty modeling.
- Dempster–Shafer theory (a.k.a. theory of belief functions) is a generalization of the Bayesian theory of subjective probability.
- Whereas Bayesian theory requires probabilities for each question of interest, belief functions allow us to base degrees of belief (strength of evidence in favor of some proposition) for one question on probabilities for a related question.

[1]http://www.gnedenko-forum.org/Journal/2007/03-042007/article19_32007.pdf

BRIEF HISTORY:

- 17th Century Roots
- 1968 > Dempster developed means for combining degrees of belief derived from independent items of evidence.
- 1976 > Shafer developed the Dempster-Shafer theory.
- 1988-90 > The theory was criticized by researchers as being confusing probabilities of truth with probabilities of provability[2,3].
- 1998 > Kłopotek and Wierzchoń proposed to interpret the Dempster–Shafer theory in terms of statistics of decision tables
- 2011 > Dr. Hemant Baruah, The Theory of Fuzzy Sets: Beliefs and Realities
- 2012 > Dr. Jøsang proved that Dempster's rule of combination actually is a method for fusing belief constraints.
- 2013 > Dr. John Yen - "Can Evidence Be Combined in the Dempster-Shafer Theory"
- 2014 > Xiaoyan Su,Sankaran Mahadevan,Peida Xu, and Yong Deng - "Handling of Dependence in Dempster–Shafer Theory"

2 http://sartemov.ws.gc.cuny.edu/files/2012/10/Artemov-Beklemishev.-Provability-logic.pdf
3 https://en.wikipedia.org/wiki/Provability_logic

BASIC EXAMPLES

EXAMPLE :

Dempster–Shafer theory has 2 components :

- Obtaining degrees of belief.
- Combining degrees of belief.

	Subjective Probability	
	Reliable	**Unreliable**
Betty	0.9	0.1

Betty

She tells me a limb fell on my car. This statement must be true if she is reliable.
So her testimony alone justifies a 0.9 degree of belief that a limb fell on my car and a Zero (not 0.1) degree of belief that no limb fell on my car.

Belief function is 0.9 and 0 together

Note : Zero degree of belief here does not mean that I am sure no limb fell on my car. It merely means based on Betty's testimony there is no reason to believe that a limb fell on my car.

COMBINING BELIEFS

Sally

	Subjective Probability	
	Reliable	**Unreliable**
Sally	0.9	0.1

I have another friend Sally.
She independent of Betty, <u>tells me a limb fell on my car.</u>
The event that Betty is reliable is independent of the event that Sally is reliable. So we multiply the probabilities of these events.

The probability that both are reliable is 0.9x0.9 =0.81,
The probability that neither is reliable is 0.1x0.1 =0.01, &
The probability that at least one is reliable is 1-0.01 =0.99

Since both said a limb fell on my car, at least one of them being reliable implies that a limb did fell on my car, and hence I may assign this event a degree of belief of 0.99

Reference : http://www.glennshafer.com/assets/downloads/articles/article48.pdf

COMBINING BELIEFS : CONTRADICT

	Subjective Probability	
	Reliable	**Unreliable**
Sally	0.9	0.1

Sally

Now suppose, Sally contradicts Betty. <u>She tells me no limb fell on my car.</u>

In this case both cannot be right and hence both cannot be reliable – one is reliable or neither is reliable.

The prior probabilities :

Only Betty is reliable and Sally is not :

(0.9x0.1) = 0.09

Only Sally is reliable and Betty is not :

(0.9x0.1) =0.09

Neither is reliable is 0.1x0.1 =0.01

COMBINING BELIEFS : CONTRADICT

The posterior probabilities :

Betty

Sally

Only Betty is reliable while Sally is not, provided at least one is unreliable :

P(Betty is reliable, Sally is not)/P(at least one is unreliable)

$0.09/[1-(0.9 \times 0.9)] = 9/19$

Only Sally is reliable while Betty is not, provided at least one is unreliable :

P(Sally is reliable, Betty is not)/P(at least one is unreliable)

$0.09/[1-(0.9 \times 0.9)] = 9/19$

Neither of them are reliable, provided at least one is unreliable :

P(Sally is unreliable and Betty is unreliable)/P(at least one is unreliable)

$[0.1 \times 0.1]/[1-(0.9 \times 0.9)] = 1/19$

Hence we have 9/19 degree of belief that a limb fell on my car (Betty reliable) & 9/19 degree of belief that no limb fell on my car (Sally reliable)

BASIC DEMPSTER SHAFER

We obtain degrees of belief for one question (Did a limb fell on my car?) from probabilities for another question (Is the witness reliable?)[4].

4 http://www.glennshafer.com/assets/downloads/articles/article48.

ALGORITHM

NEXT STEPS

- Quantify fuzziness in binary prediction
- Obtain improved probabilities for binary classification based off a prior classification algorithm using Modified Dempster-Shafer

DATASETS TESTED

- ✓ Higgs Boson
- ✓ Titanic
- ✓ Cdystonia
- ✓ Parkinsons Telemonitoring
- ✓ Abalone
- ✓ Yeast
- ✓ Skin_Nonskin
- ✓ Contraceptive Method Choice
- ✓ Diabetic Renopathy
- ✓ Mammographic Mass
- ✓ Breast Cancer Wisconson (Diagnostic)
- ✓ Breast Cancer Wisconson (Prognastic)
- ✓ Haberman's Survival
- ✓ Statlog (Heart)
- ✓ SPECT Heart
- ✓ SPECTF Heart
- ✓ Hepatitis
- ✓ Sponge
- ✓ Audiology
- ✓ Soybean
- ✓ Ecoli
- ✓ Horse Colic
- ✓ Post-Operative Patient
- ✓ Zoo
- ✓ EEG Eye State
- ✓ Wilt
- ✓ Bach
- ✓ Nursery
- ✓ Phishing Websites
- ✓ Pets
- ✓ Contraceptive Method
- ✓ First Order
- ✓ Spambase
- ✓ Hill Valley
- ✓ Satellite Image
- ✓ Agaricus Lepiota
- ✓ King Rook vs King Pawn
- ✓ Firm Teacher
- ✓ Fertility
- ✓ Dermatology
- ✓ Mice Protein
- ✓ LSVT Voice
- ✓ Thyroid

MODIFIED DS SUMMARY

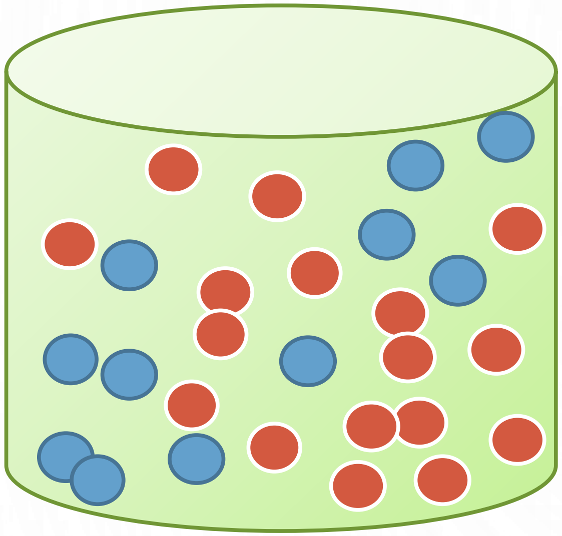
Positives

- Can significantly increase performance metrics and to a greater precision
- Algorithm is not black-box. It can be understood relatively easily
- Prior probabilities can come from algorithms other than logistic regression

Road Forward

- Scale performance with data dimension
- Correct for highly imbalanced classes which may lead to high-class probability when there is none (high/low belief conflict)
- Test different prior algorithms

JAR OF MARBLES



What is the majority color in the jar?

	Person 1	Person 2
Majority is Red	70%	50%
Majority is Blue	20%	30%
Half Blue and Red	10%	20%

A Dempster-Shafer student could then say he/she believes:

Red	Blue	Half-half
81%	17%	2%

DEMPSTER'S RULE OF COMBINATION

Combines multiple pieces of evidence (mass functions).

$$(m_1 \oplus m_2)(A) = \boxed{\frac{1}{1-K}} \sum_{B \cap C = \emptyset} m_1(B) m_2(C) = m_{1,2}(A)$$

Emphasizes the agreement, and ignores the conflict of the mass functions by using the normalization factor 1-K

$$K = \sum_{B \cap C = A \neq \phi} m_1(B) m_2(C), K \neq 1$$

K is the measure of conflict between the two mass sets meaning 1-K is the measure of agreement

DS NOTATION

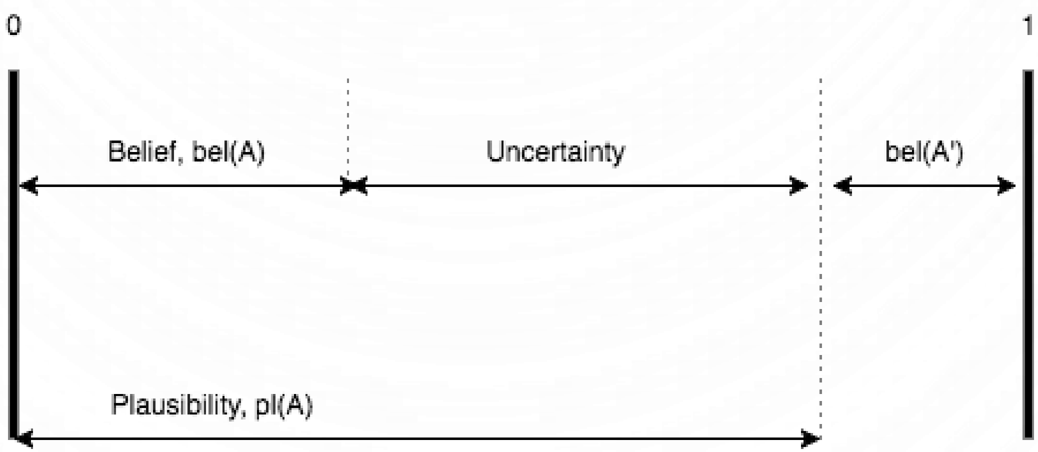
$$bel(A) = \sum_{B \subseteq A} m(B)$$

The **belief** function of a class sums the mass values of all the non-empty subsets of that class. <u>The strength of evidence in favor of some proposition</u>

$$\text{pl}(A) = \sum_{B \cap A \neq \phi} m(B)$$

The **plausibility** function of a class sums the mass values of all sets that intersect that class. <u>Measures what is left over after subtracting evidence against a belief.</u>

$$m(A) \leq bel(A) \leq pl(A)$$



MODIFIED DS: OBTAINING POSTERIOR

Given some evidence and prior probabilities for a class, we can obtain a posterior or updated mass value

E'

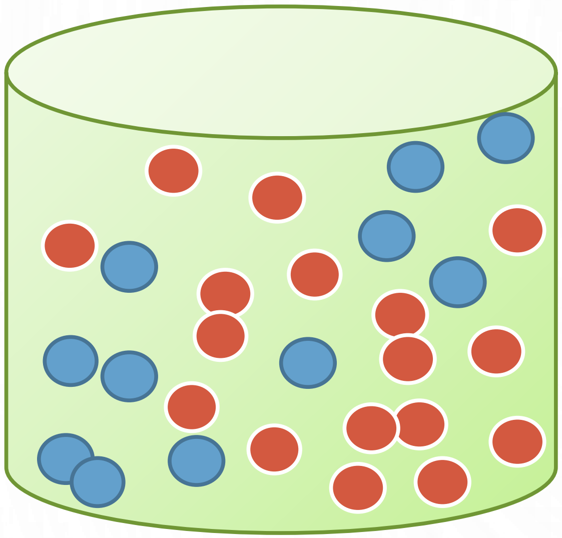
Evidence Space: a set of mutually exclusive outcomes or possible values of an evidence source.

$$C(A|E') = \frac{\frac{m(A|E')}{P(A)}}{\sum_{A\subset\Theta} \frac{m(A|E')}{P(A)}}$$

The **Basic Certainty Value:** a normalized ratio of a hypothesis subset mass to its prior probability*

R. P. S. Mahler D. Fixsen. *The modified dempster-shafer approach to classification*

JAR OF MARBLES



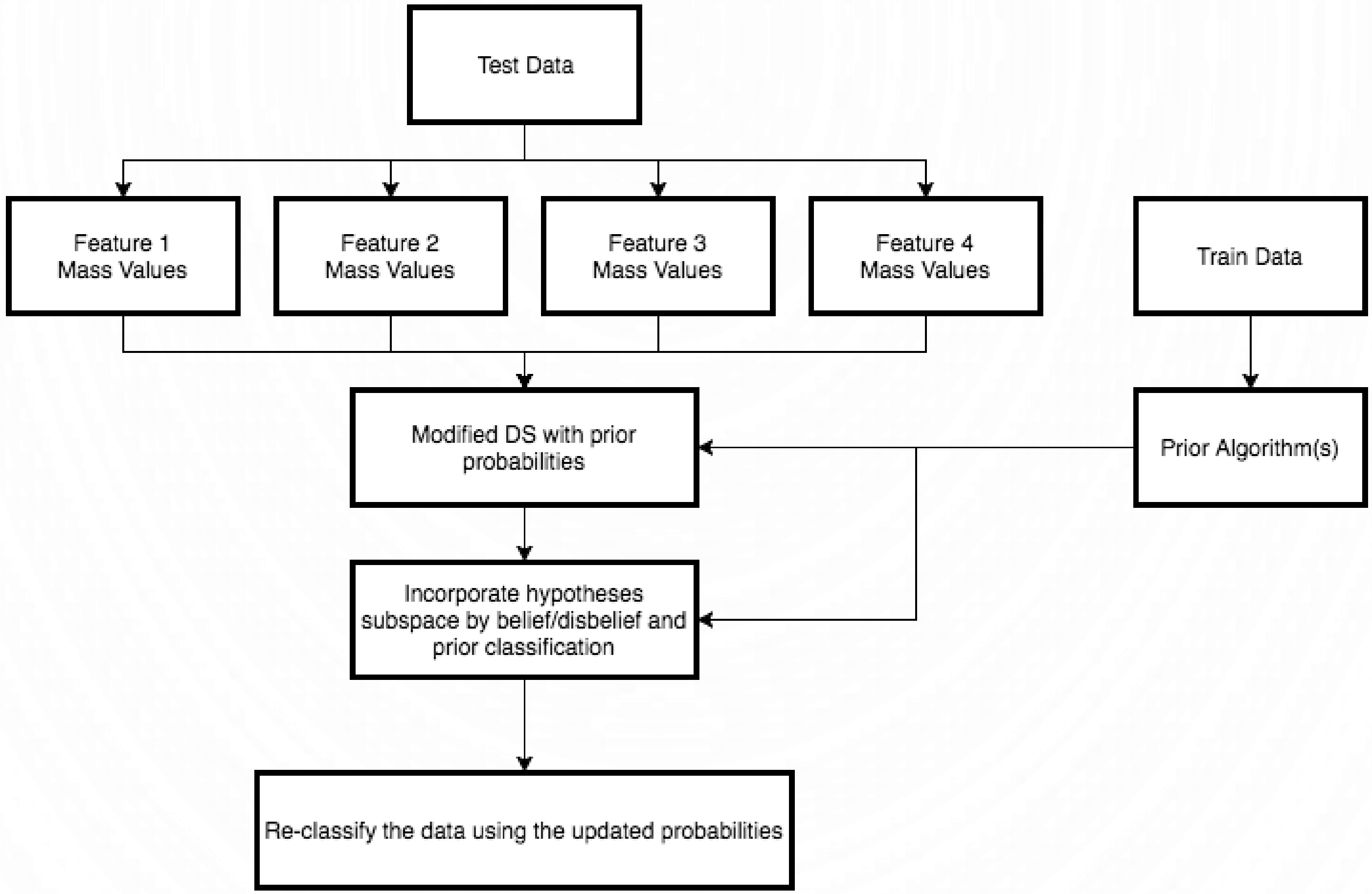
What is the majority color in the jar?

	Person 1	Person 2
Majority is Red	70%	50%
Majority is Blue	20%	30%
Half Blue and Red	10%	20%

Prior Information	
Red	Blue
60%	40%

	A Dempster-Shafer student could then say he/she believes:		
	Red	Blue	Half-half
DS	81%	17%	2%
Modified	84.9%	13.5%	1.6%

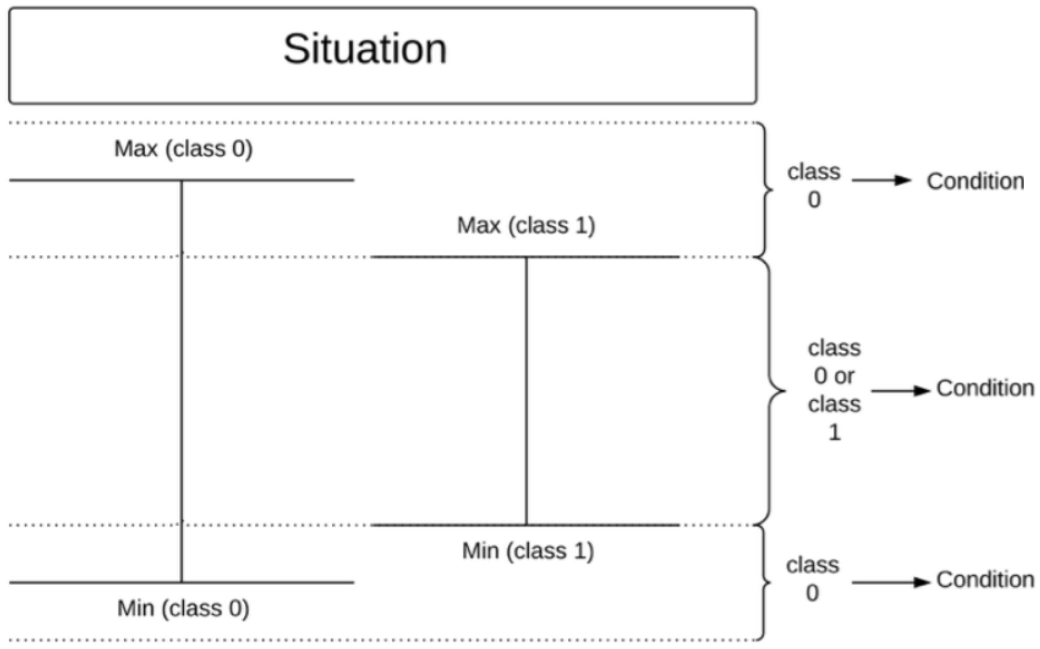
THE ALGORITHM – BUILD A MODEL



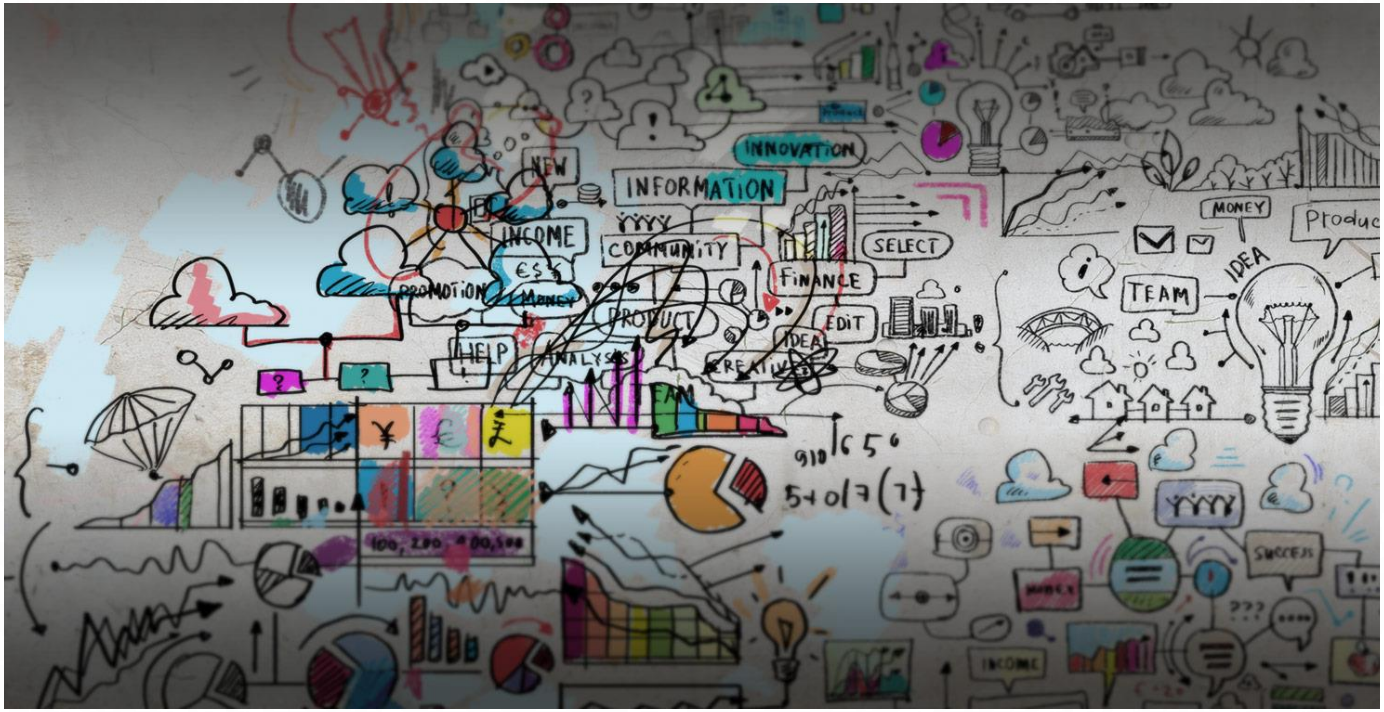
OBTAINING MASS VALUES

<u>For Each Feature</u>

1. Partition the training data by class
2. Calculate the max/min values of each partition
3. Determine which max/min value situation it is by looking at the ranges of each partition



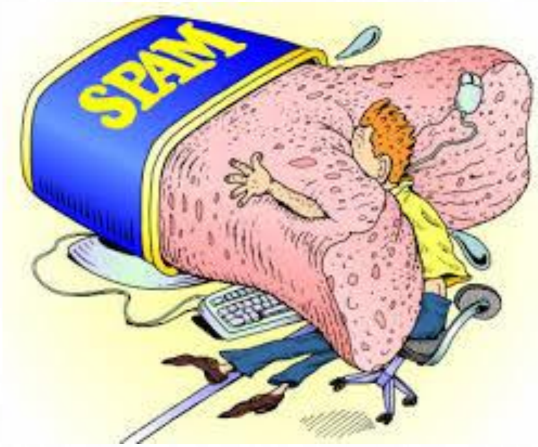
EXAMPLES



So far have tested DS algorithm on **43** datasets:

- UCI Machine Learning Repository
- R package “Datasets”

EMAIL SPAM

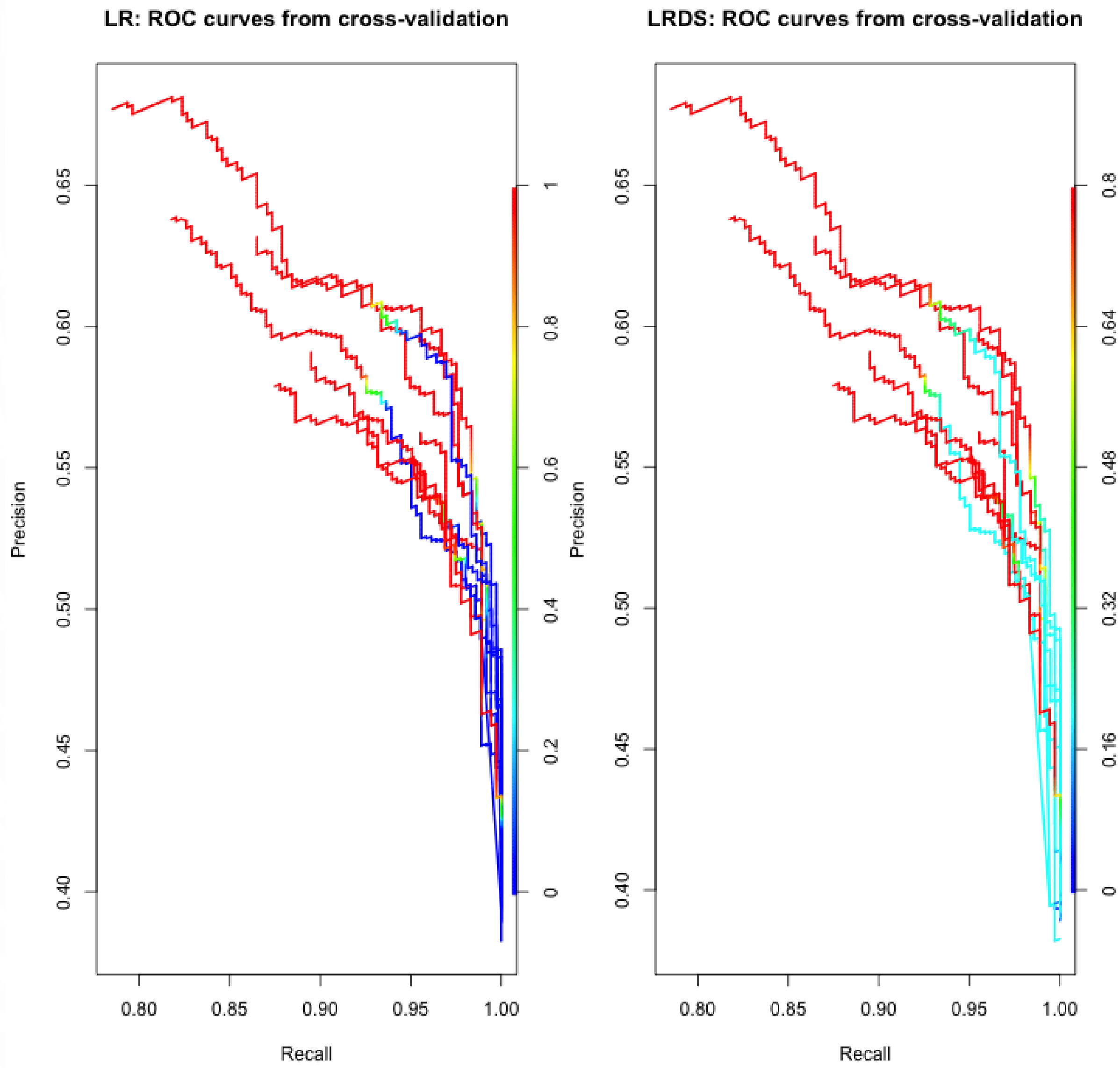


- Email spam is from a diverse group of individuals. Predicting on individuals definition of spam or not.
- 58 features: primarily of different word and character frequencies
- 4601 observations

Median Metrics Based of 10-fold CV		
Metrics	LR	LRDS
Accuracy	0.63587	0.63913
AUC	0.76317	0.76265
F1 Score	0.45753	0.45840
Precision	0.52277	0.52470
True Positive	0.39503	0.40151
False Positive	0.01245	0.01384

https://archive.ics.uci.edu/ml/datasets/Spambase

EMAIL SPAM


LR: ROC curves from cross-validation
LRDS: ROC curves from cross-validation
Precision
Recall
0.40
0.45
0.50
0.55
0.60
0.65
0.80
0.85
0.90
0.95
1.00
0
0.2
0.4
0.6
0.8
1
0.16
0.32
0.48
0.64

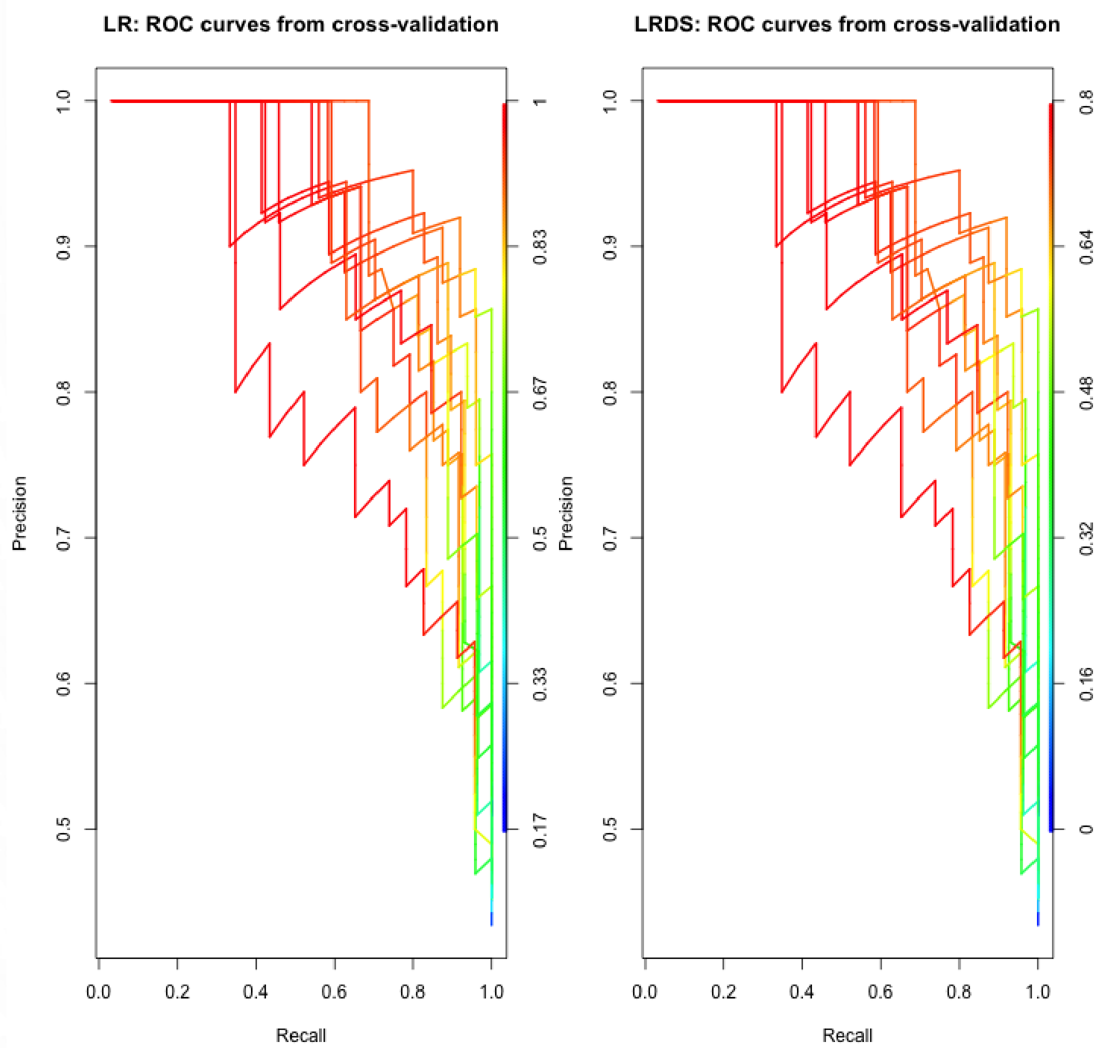
SPECT HEART

- Cardiac **S**ingle **P**roton **E**mission **C**omputed **T**omography images
- Predicting on normal or abnormal heart
- 23 features on partial human diagnoses of abnormality
- 267 observations

Median Metrics Based of 10-fold CV		
Metrics	LR	LRDS
Accuracy	0.55660	0.72641
AUC	0.89846	0.89846
F1 Score	0.52168	0.57441
Precision	0.52618	0.66678
True Positive	0.14074	0.50020
False Positive	0.01562	0.036458

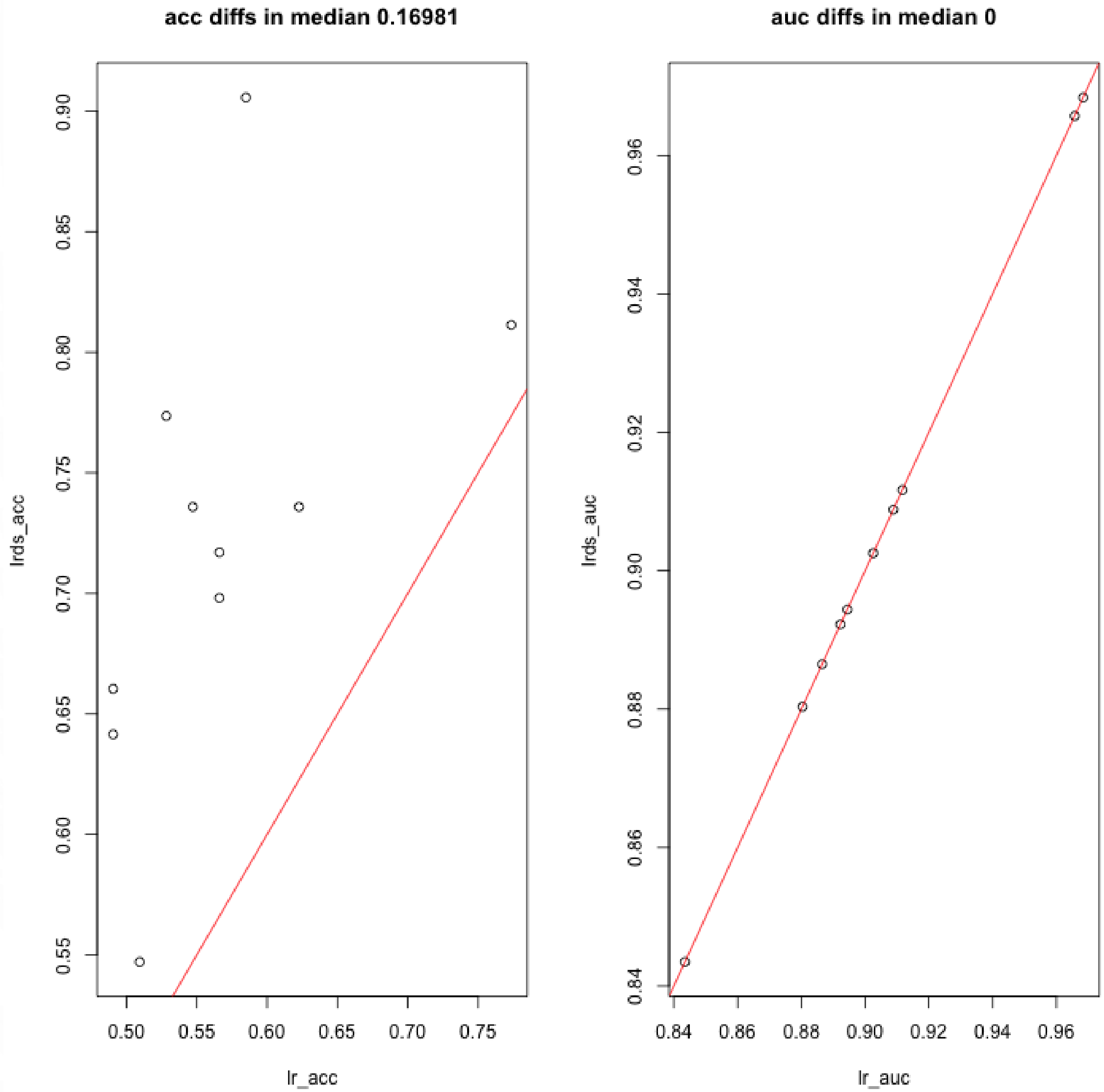
Source: https://archive.ics.uci.edu/ml/datasets/Skin+Segmentation

SPECT HEART



LR: ROC curves from cross-validation
LRDS: ROC curves from cross-validation
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.2
0.4
0.6
0.8
1.0
1
0.83
0.67
0.5
0.33
0.17
0.8
0.64
0.48
0.32
0.16
0

SPECT HEART



acc diffs in median 0.16981
lrds_acc
lr_acc
0.50 0.55 0.60 0.65 0.70 0.75
0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90
auc diffs in median 0
lrds_auc
lr_auc
0.84 0.86 0.88 0.90 0.92 0.94 0.96

REFERENCES

- Application of Dempster Shafer in reliability modeling http://www.gnedenko-forum.org/Journal/2007/03-042007/article19_32007.pdf
- Provability Logic http://sartemov.ws.gc.cuny.edu/files/2012/10/Artemov-Beklemishev.-Provability-logic.pdf
- Wiki Provability Logic https://en.wikipedia.org/wiki/Provability_logic
- Dr. Shafer's Article http://www.glennshafer.com/assets/downloads/articles/article48.pdf
- Home Page of Dr. Lotfi Zadeh https://www2.eecs.berkeley.edu/Faculty/Homepages/zadeh.html
- L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning - I," Information Sciences, vol. 8, no. 3, pp. 199-249, July 1975.
- R. E. Bellman and L. A. Zadeh, "Decision making in a fuzzy environment," Management Science, vol. 17, no. 4, pp. B141-B164, Dec. 1970.
- L. A. Zadeh, "Fuzzy sets," Information and Control, vol. 8, no. 3, pp. 338-353, June 1965.
- Alexander Karlsson. R: Evidence Combination in R. R Foundation for Statistical Computing, MIT, 2012.
- R. P. S. Mahler D. Fixsen. The modified dempster-shafer approach to classification. IEEE Std. 1516-2000, pages 96 – 104, Jan 1997.
- Jiwen Guan, Jasmina Pavlin, and Victor R. Lesser. Combining Evidence in the Extended Dempster-Shafer Theory, pages 163–178. Springer London, London, 1990
- M. Lichman. UCI machine learning repository, 2013
- Simon Parsons and London Wca Px. Some qualitative approaches to applying the dempster-shafer theory, 1994.

QUESTIONS?

APPENDIX

DS NOTATION

$\Theta = \{a, b\}$

A hypothesis for the possibilities which is a subset of all possibilities

$2^\Theta = \{\phi, a, b, \Theta\}$

The power set is the set of all possible subsets of Θ, including Θ itself and the empty set Φ

$m: 2^\Theta \rightarrow [0,1]$

Mass function: Each element of the power set has a mass value assigned to it

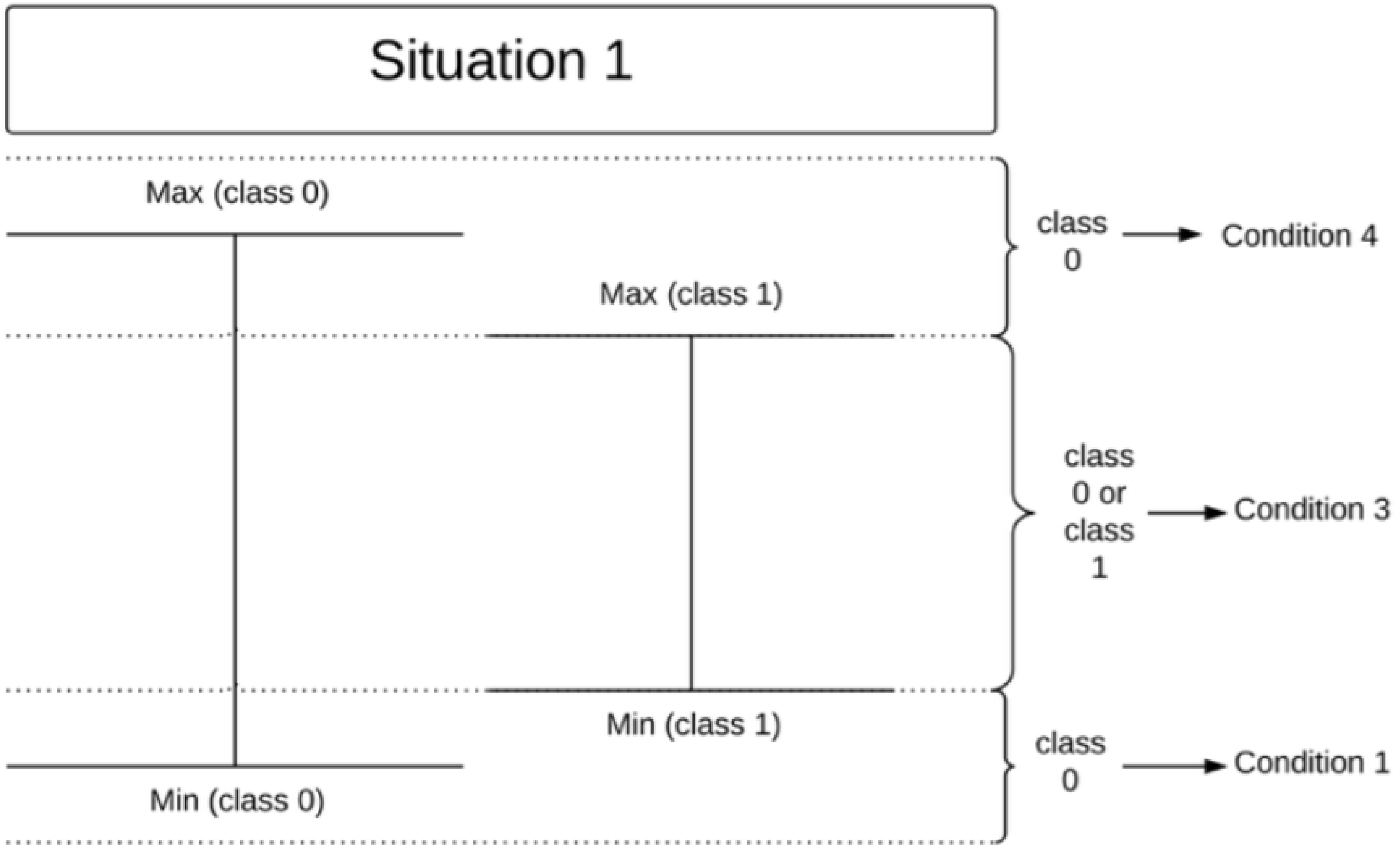
$m(\phi) = 0$

For the purposes of our algorithm, the mass of the empty set is zero

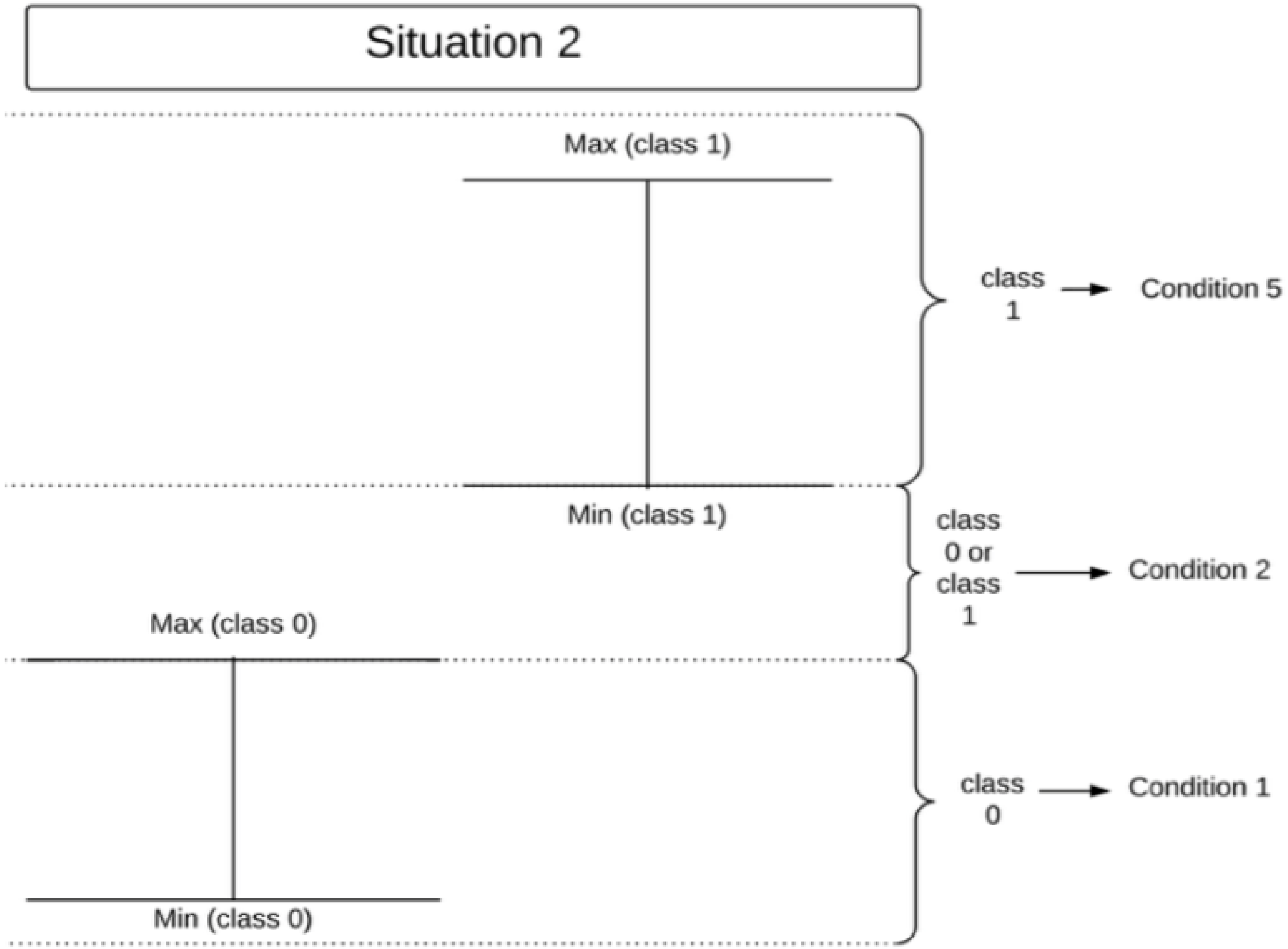
$\sum_{A \subseteq \Theta} m(A) = 1$

Sum of masses (for each row) is 1 (currently)

CLASSIFICATION PROCESS



CLASSIFICATION PROCESS



CLASSIFICATION PROCESS