A Brute Force Method of Discovering Information Hidden in Datasets

Kristopher Kurt Honetschlager, Computer Science Major, Winona State University-Rochester ‘23

kristopher.honetschlager@go.winona.edu

Winona State University at the Rochester Campus

Computer Science/Statistics/Logic/ Data Science What are they?(Background Information)

- Computer Science can at least partly be defined as "…the study of computers and computing, including their theoretical and algorithmic foundations, hardware and software, and their uses for processing information.".[1]
- Statistics can at least partly be defined as "…the science of collecting, analyzing, presenting, and interpreting data.".[2]
- Logic can at least partly be defined as "The science that investigates the principles governing correct or reliable inference.".[3]
- Data Science can at least partly be defined as "…[a] field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.".[4]

Motivation (Mushroom Dataset Theory/Hypothesis)

- Contrarianism: guide states "…no simple rule for determining the edibility of a mushroom…"
  - unlike "…`leaflets three, let it be' for Poisonous Oak and Ivy."[8]
- The .names file lays out simple rules for determining whether a given mushroom instance is poisonous or not. Such as: "spore-print-color=green 48 cases missed, 99.41% accuracy".[8]
- I figured a brute force method of somehow testing every column and pairing of columns against the poisonous column and retrieving a highest score would be a very good way of obtaining these rules.[8]

[8]Schlimmer, Jeff. "Mushroom Data Set." Edited by Dave Aha, UCI Machine Learning Repository: Mushroom Data Set, University of California, Irvine, 27 Apr. 1987, https://archive.ics.uci.edu/ml/datasets/Mushroom.

Example Mushroom Dataset

no_odor	spore-print-color=green	habitat=leaves	cap-color-white	poisonous-yes
0	0	0	0	0
0	0	0	0	0
0	0	0	0	1
1	0	0	1	0
0	0	0	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	0
0	0	0	0	0
0	0	0	0	0

Number of Columns Selected = 4

Jain, Sumit. “Print All Combinations of Subset of Size K from given Array.” Tutorial Horizon, 31 Aug. 2019, https://algorithms.tutorialhorizon.com/print-all-combinations-of-subset-of-size-k-from-given-array/.

Number of Columns Selected = 1

no_odor
spore-print-color=green
habitat=leaves
cap-color-white

Number of Columns Selected = 2

no_odor
spore-print-color=green
habitat=leaves
cap-color-white

no_odor
spore-print-color=green
habitat=leaves
cap-color-white

Number of Columns Selected = 3

no_odor
spore-print-color=green
habitat=leaves
cap-color-white

no_odor
spore-print-color=green
habitat=leaves
cap-color-white

no_odor
spore-print-color=green
habitat=leaves
cap-color-white

Number of Columns Selected = 4

no_odor
spore-print-color=green
habitat=leaves
cap-color-white

no_odor
spore-print-color=green
habitat=leaves
cap-color-white

no_odor
spore-print-color=green
habitat=leaves
cap-color-white

Extension of Mushroom Dataset Theory/Hypothesis and Terminology

- Why not test every possible column and pairing of columns against every possible label to obtain the best rule obtainable for predicting each label in the dataset.
  - Could be a rule that predicts a label that was previously unknown, the hidden aspect, or finds one contrary to current belief, the contrarian aspect.
- Acts like a hypothesis machine
- Selected columns can be referred to as the X or feature vector in data science[7]
- Row, also referred to as instance[7]
- Column, also referred to as attribute[7]
- Poisonous can be referred to as the label in data science[7]

Example Mushroom Dataset

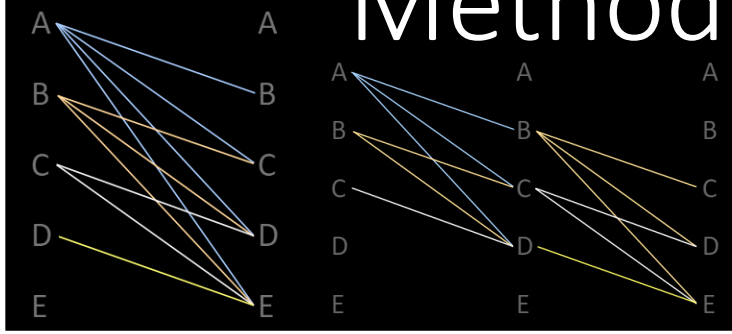
no_odor	spore-print-color=green	habitat=leaves	cap-color-white	poisonous-yes
0	0	0	0	0
0	0	0	0	0
0	0	0	0	1
1	0	0	1	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	1
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

Number of Columns Selected = 1

Photo by sk8erteck on deviantart.com

A	B
1	0
0	1
1	0
0	1

Sort of in-Depth Explanation of Method



ABC, ABD, and A

BCD and BCE

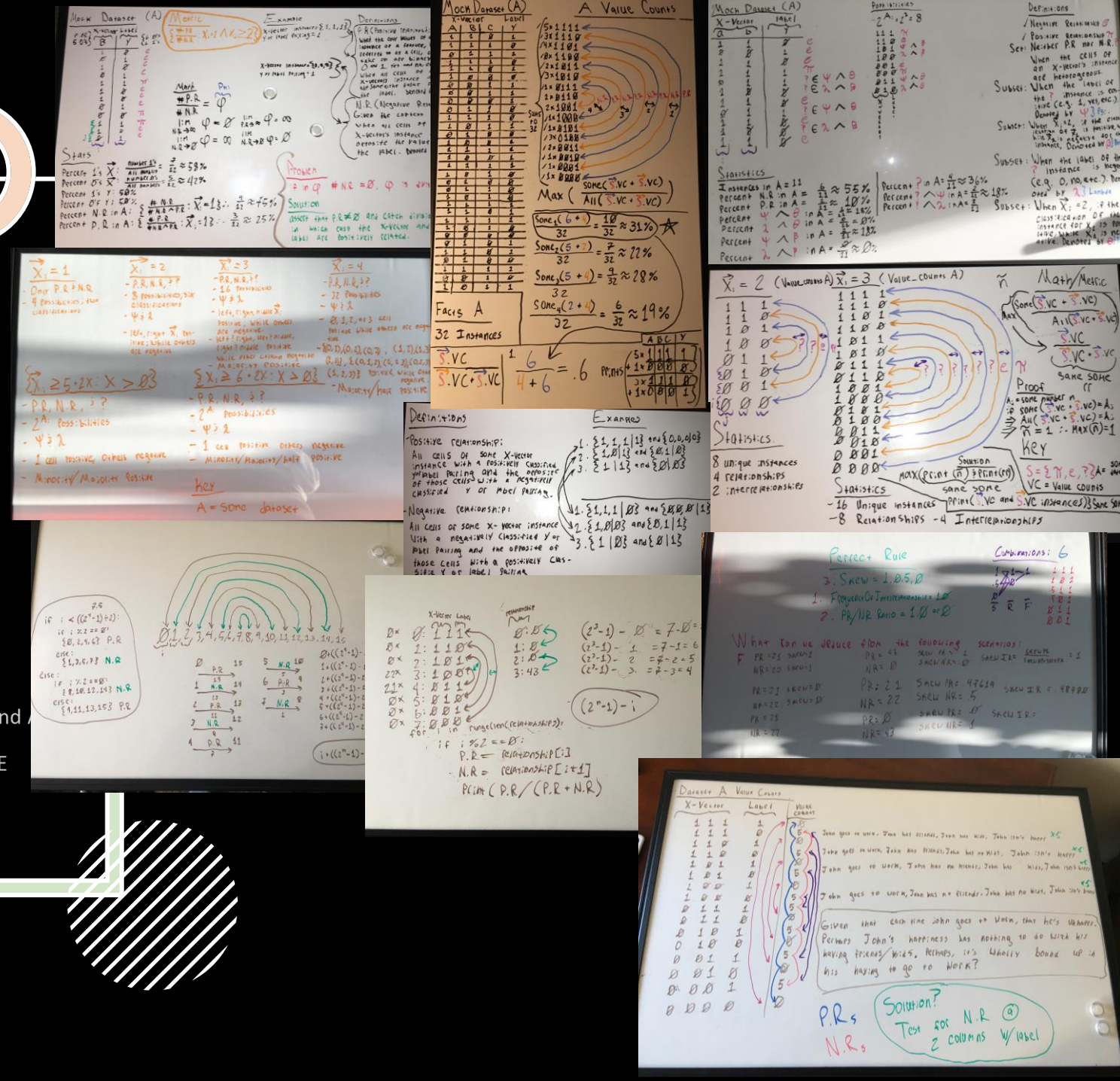
CDE

AB, AC, AD, and AE

BC, BD, and BE

CD and CE

DE



Statistics: Dataset Types

- You can use this method on categorical datasets and binary datasets
  - For categorical datasets you must break categorical columns into binary columns

```
3. cap-color:                    brown=n,buff=b,cinnamon=c,gray=g,green=r,
                                 pink=p,purple=u,red=e,white=w,yellow=y
```

- My method fundamentally operates on binary datasets

poisonous
No
No
Yes
No

poisonous_yes	poisonous_no
0	1
0	1
1	0
0	1

← CATEGORICAL DATASET			BINARY DATASET →		
poisonous	cap-shape	cap-color	poisonous_yes	cap-color_w	cap-shape_b
No	x	n	0	0	0
No	x	y	0	0	1
Yes	b	w	1	1	0
No	x	w	0	0	0

Statistics: Binary Dataset Value Counts

Example Binary Dataset

John Doe is Happy	John Doe had ice cream
1	1
1	1
1	1
1	1
1	1
0	0
0	0
0	0
0	0
0	0

Example Binary Dataset Value Counts

John Doe is Happy	John Doe Had ice cream	Number of Occurences
1	1	5
1	0	0
0	1	0
0	0	5

Value counts represents all possible unique binary combinations given the number of columns in the dataset.

Logic: Conditional Statement Forms

- If p, where p is the hypothesis, then q, where q is the conclusion
  - "If it's sunny and there's no snow on the ground, I will go on a walk outside."
- By traditional logic, "A conditional statement is **not** logically equivalent to its inverse."[7]. Though perhaps instinctually it may seem that way.
  - Therefore, following from the previous conditional statement it cannot be additionally concluded that "If it's not sunny and there is snow on the ground, I won't go on a walk outside."
- But if we have empirical evidence suggesting the first statement and its inverse occur at a respective frequency of greater than 25% in a given dataset. We may be able to imply that the beginning conditional statement, though not logically equivalent to its inverse, is empirically equivalent to its inverse.
- [TERM] The *denial* of the first point is "If p then not q".
  - "If it's sunny and there's no snow on the ground, I won't go on a walk outside."

Sheldon, Frederick T. "Conditional Statement Forms." Review 1.2, Oak Ridge National Laboratory, https://www.csm.ornl.gov/~sheldon/ds/sec1.2.html#Neg,%20Convrs,%20Invrs%20of%20Cond%20Stmnts.

Example: Empirical Observation of a Rule

Example Binary Dataset

Example Binary Dataset Value Counts

John Doe is Happy	John Doe Had ice cream	Number of Occurences
1	1	5
1	0	0
0	1	0
0	0	5
0	0	
0	0	
0	0	

John Doe is happy; did he have ice cream today?

John Doe is not happy; did have ice cream today?

Generation of Disjunctive Rules

- Empirically, rules can be for identifying predictors that a given label's instance is positive.
  - If X is true and Y is true, then Z is true.
- Predictors that the given instance label is negative.
  - If X is false and Y is false, then Z is false.
- Or predictors indicative of the given label's instance being positive or negative based on the inversion of either set of predictors.
  - If X is true and Y is true, then Z is true and if X is false and Y is false, then Z is false.

Statistics: Characteristics of Binary Value Counts

Occurrence frequency of John Doe is happy, and John Doe had Ice Cream = 186/200 = .93 * 100 = 93%

Positive Rule Generation

Occurrence frequency of John Doe isn't happy, and John Doe had Ice Cream = 4/200 = .035 * 100 = 3.5%

Example Binary Dataset Value Counts

John Doe is Happy	John Doe Had ice cream	Number of Occurrences
1	1	186
1	0	2
0	1	7
0	0	5

If one "positive" rule occurs the most frequently in the "positive" rules and in the dataset, greater than 50% respectively; it implies the denials of the rule occur technically less than 50% of the time.

John is happy, did he have ice cream?

Statistics: Characteristics of Binary Value Counts

Occurrence frequency of John Doe is happy, and John Doe didn't have Ice Cream =
182/200 = .91 * 100 = 91%

Negative Rule Generation

Occurrence frequency of John Doe isn't happy, and John Doe didn't have Ice Cream =
4/200 = .02 * 100 = 2%

Example Binary Dataset Value Counts

John Doe is Happy	John Doe Had ice cream	Number of Occurrences
1	1	5
1	0	182
0	1	9
0	0	4

John is happy, did he not have ice cream?

If one "negative" rule occurs the most frequently in the "negative" rules and in the dataset, greater than 50% respectively; it implies the denials of the rule occur technically less than 50% of the time.

Statistics: Characteristics of Binary Value Counts

Total number of rows in dataset = 200

Combinatorial Rule Generation

Occurrence frequency of John Doe is happy and John Doe had ice cream = 95/200 = 0.475 * 100 = 47.5%

Occurrence frequency of John Doe is happy and John Doe didn't have ice cream = 5/200 = 0.025 * 100 = 2.5%

Occurrence frequency of John Doe is not happy and John Doe had ice cream = 95/200 = 0.475 * 100 = 47.5%

Occurrence frequency of John Doe is not happy and John Doe didn't have ice cream = 5/200 = 0.025 * 100 = 2.5%

Example Binary Dataset Value Counts

John Doe is Happy	John Doe Had ice cream	Number of Occurrences
1	1	100
1	0	100
0	1	0
0	0	0

The rule John Doe is happy therefore he had ice cream and John Doe is happy therefore he didn't have ice cream do not deny one another because they are two different combinations of the data.

John Doe is happy therefore he had ice cream and John Doe is happy therefore he didn't have ice cream have a combination of both occurring more than 25% of the time.

Statistics: Characteristics of Binary Value Counts

Total number of rows in dataset = 200

Combinatorial Rule Generation Cont.

Occurrence frequency of John Doe is happy, he's had ice cream and John Doe isn't happy, he hasn't had ice cream = 100/200 = .5 * 100 = 50%

Example Binary Dataset Value Counts

John Doe is Happy	John Doe Had ice cream	Number of Occurrences
1	1	50
1	0	50
0	1	50
0	0	50

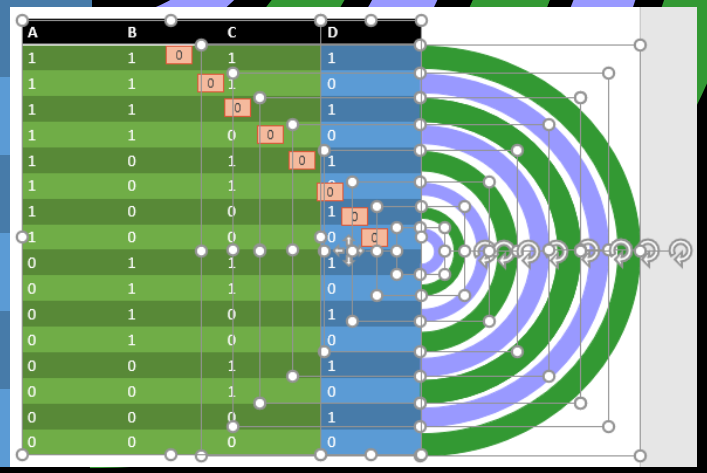
Occurrence frequency of John Doe is happy, he hasn't had ice cream and John Doe isn't happy, John Doe has had ice cream = 100/200 = .5 * 100 = 50%

"If John Doe is happy, he's had ice cream and if John Doe isn't happy, he hasn't had ice cream." is the denial of the rule "If John Doe is happy, he hasn't had ice cream and if John Doe isn't happy, John Doe has had ice cream.".

John Doe is happy; has he had ice cream?

Value Counts Dataset from Hypothetical Dataset

Value Counts	A	B	C	D
100	1	1	1	1
100	1	1	1	0
100	1	1	0	1
100	1	1	0	0
100	1	0	1	1
100	1	0	1	0
100	1	0	0	1
100	1	0	0	0
100	0	1	1	1
100	0	1	1	0
100	0	1	0	1
100	0	1	0	0
100	0	0	1	1
100	0	0	1	0
100	0	0	0	1
100	0	0	0	0



Mathematics/Computer Science: Where's the Math and Computer Science?

Value Counts Dataset from Hypothetical Dataset

Every instance that when it's index modulus 2 is equal to 0 has a positive label

Every instance that when it's index modulus 2 is equal to 1 has a negative label

	Value Counts	A	B	C	D
0	100	1	1	1	1
1	100	1	1	1	0
2	100	1	1	0	1
3	100	1	1	0	0
4	100	1	0	1	1
5	100	1	0	1	0
6	100	1	0	0	1
7	100	1	0	0	0
8	100	0	1	1	1
9	100	0	1	1	0
10	100	0	1	0	1
11	100	0	1	0	0
12	100	0	0	1	1
13	100	0	0	1	0
14	100	0	0	0	1
15	100	0	0	0	0

Total number of columns in value counts dataset (not counting value counts column) = 4 = n

Index of inverse of current instance, when the current instance's index is less than $2^n/2 = (2^n - 1) -$ index current instance

Example using instance at 0: current index $< 2^4/2$, therefore inverse of instance index $= (2^4 - 1) - 0 = 15$

Causality Is Not Proven (Hypothesis [Rule] Machine)

Demonstration

Binary Columns in Dataset (Instances in dataset = 40, Upsampled):

- john-doe_is-happy
- john-doe_had-dinner

If John Doe is happy, John Doe had dinner: {1, 1}

If John Doe is unhappy, John Doe didn’t have dinner: {0, 0}

ULTIMATE GOAL is to find a good pairing with machine learning label

Future Research and Applications

- Phenotyping Algorithms "… are special tools that enable researchers to extract phenotypes from complex, and often messy data that get generated during routine interactions within the healthcare system."[4].
- Induction Logic Programming
- Disjunctive Rule Generation

- UC Irvine for hosting mushrooms dataset
- Collin Engstrom my independent study advisor
- Minnstate Undergraduate Scholars Symposium
- Listeners

Thank you

Works Cited

- [1]Belford, Geneva G. and Tucker, Allen. "computer science". Encyclopedia Britannica, Invalid Date, https://www.britannica.com/science/computer-science. Accessed 20 March 2022.
- [2]Sweeney, Dennis J. , Anderson, David R. and Williams, Thomas A.. "statistics". Encyclopedia Britannica, Invalid Date, https://www.britannica.com/science/statistics. Accessed 20 March 2022.
- [3]"Logic Definition & Meaning." Dictionary.com, Dictionary.com, https://www.dictionary.com/browse/logic.
- [4]"Data Science." *DataRobot*, 20 Jan. 2022, https://www.datarobot.com/wiki/data-science/.
- [5]"Phenotype Library: HDRUK." Phenotype Library | HDRUK, https://phenotypes.healthdatagateway.org/.
- [6]Sk8erteck. "Light Shining through Windows." *Deviant Art*, 18 Mar. 2004, https://www.deviantart.com/sk8erteck/art/Light-Shining-through-Windows-5930403. Accessed 18 Apr. 2022.
- [7]Kohavi, Ron. "Glossary of Terms." Glossary of Terms Journal of Machine Learning, Kluwer Academic Publishers, 1998, http://robotics.stanford.edu/~ronnyk/glossary.html.
- Additional citations on the bottom of some slides