Carnegie Mellon

Center for Automated Learning and Discovery

and Robotics Institute

and School of Computer Science

Accelerating Machine Learning Algorithms with Cached Sufficient Statistics

Andrew Moore

Mary Soon Lee

*Thanks to Brigham Anderson, Scott Davies, Anthony Dilello, Nir Friedman. Funded by NSF.

www.cs.cmu.edu/~AUTON

Here’ s the top-right corner of a dataset

age	workc lass	marital-status	sex	...
31-60	Self-employed	Divorced	Male	...
0-30	Unknown	Never-married	Female	...
0-30	Unknown	Never-married	Female	...
31-60	Private	Married-spouse-present	Male	...
31-60	Private	Married-spouse-present	Male	...
0-30	Unknown	Never-married	Female	...
31-60	Private	Married-spouse-present	Male	...
31-60	Private	Divorced	Female	...
0-30	Unknown	Never-married	Female	...
31-60	Unknown	Married-spouse-present	Female	...
0-30	Unknown	Never-married	Male	...
0-30	Self-employed	Married-spouse-present	Male	...
31-60	Private	Married-spouse-present	Male	...
60-90	Unknown	Married-spouse-present	Female	...
0-30	Unknown	Never-married	Male	...
:	:	:	:	:

(143,000 records. 13 attributes).

From 1990 census data [Kohavi 1995]

The Census Dataset

```
ds_edit> details
age                    Attribute   0 (Symbolic)
     v0:30- (66163) v1:30+ (55388) v2:60+ (20970)

workclass                    Attribute   1 (Symbolic)
     Federal-government (2054) Local-government (5622)
     Never-worked (298)
     Private (51451) Self-employed-incorporated (2290)
     Self-employed-not-incorporated (6173) State-government (2995)
     Unknown (71523) Without-pay (115)

marital-status                    Attribute   2 (Symbolic)
     Divorced (9102) Married-A-F-spouse-present (460)
     Married-civilian-spouse-present (59992)
     Married-spouse-absent (1005)
     Never-married (62287) Separated (2403) Widowed (7272)

occupation                    Attribute   3 (Symbolic)
     Adm-support-including-clerical (10748) Armed-Forces (26)
     Executive-admin-and-managerial (8764)
     Farming-forestry-and-fishing (2325)
     Handlers-equip-cleaners-etc (3006)
     Machine-operators-assmblrs-&-inspctrs (4583) Other-service (8802)
     Precision-production-craft-&-repair (7619)
     Private-household-services (572) Professional-specialty (9556)
     Protective-services (1181) Sales (8405)
     Technicians-and-related-support (2174)
     Transportation-and-material-moving (2939) Unknown (71821)

sex                    Attribute   4 (Symbolic)
     Female (74176) Male (68345)

capital-gain                    Attribute   5 (Symbolic)
     v0:40000- (142276) v1:40000+ (4) v2:80000+ (241)

capital-loss                    Attribute   6 (Symbolic)
     v0:40000- (142462) v1:40000+ (46) v2:80000+ (13)

area                    Attribute   7 (Symbolic)
     Abroad (613) Midwest (5129) Northeast (3179) South (6771)
     Unknown (120950) West (5879)

parents                    Attribute   8 (Symbolic)
     Both-parents-present (28294) Father-only-present (1323)
     Mother-only-present (9279) Neither-parent-present (1159)
     Unknown (102466)
```

The Census Dataset ctd

```
citizenship                  Attribute   9 (Symbolic)
     Foreign-born--Not-a-citizen-of-U-S (8157)
     Foreign-born--U-S-citizen-by-naturalization (3431)
     Native--Born-abroad-of-American-Parent(s) (978)
     Native--Born-in-Puerto-Rico-or-U-S-Outlying (1066)
     Native--Born-in-the-United-States (128889)

hours-per-week                  Attribute  10 (Symbolic)
     v0:20- (76684) v1:20+ (9450) v2:40+ (56387)

class                  Attribute  11 (Symbolic)
     --50000. (134359) 50000+. (8162)
```

The Birth Dataset

s3emcsec	vol	s1anem	s1desexp	s1digest	s1diab	s1renal	...
0	0-1200	0	0	0	0	0	...
0	0-1200	0	0	0	0	0	...
0	0-1200	0	0	0	0	0	...
0	0-1200	0	0	0	0	0	...
0	0-1200	0	0	0	0	0	...
0	0-1200	0	0	0	0	0	...
0	0-1200	0	0	0	0	0	...
1	0-1200	0	0	0	1	0	...
0	0-1200	0	0	0	0	0	...
1	0-1200	0	0	0	0	0	...
0	0-1200	0	0	0	0	0	...
0	0-1200	0	0	0	0	0	...
1	0-1200	0	0	0	0	0	...
0	0-1200	0	0	0	0	0	...
0	0-1200	1	0	0	0	0	...
0	0-1200	0	0	0	0	0	...
:	:	:	:	:	:	:	:

97 Attributes

9672 Records

Birth Attributes

s3emcsec , vol , s1hbpath , s1anem , s1htn , s1ahd , s1desexp , s1digest , s1diab , s1renal , s1seiz , s1hac he , s1resp , s1depr , s1suic , s1coke , s1etoh , s1smoke , s1plate , s1petop , s1vagdel , s1pcsect , s1dlvprb , s1inccvx , s1pprtcn , s1rhinc , s1aboinc , s1ec l , s1htnec l , s1genabn , s1astprg , s1ac hd , s1gbd , s1dka , s1guinf , s1vomit , s1thbpt , s1bleed1 , s1bleed2 , s1bleed3 , s1prmlbr , s1iugr , s1abrupt , s1placab , s1olpoly , s1cpd , s1cancer , cragl18 , cragg35 , crblac k , crasian , crrothr , s2amnio , s2octcst , s2nstfat , s2htn1 , s2anem1 , s2vanem1 , s2glycu1 , s2protu1 , s2std , s2bpr2 , s3malprs , s3anem , s3vanem , s3fetdis , mlbrth , vanem , s1genscr , s1extvrs , s3amninf , s3intmtr , s3extmtr , s3bthmtr , s3fetsc l , s3attob , s3attfp , s3attmw , s3attrn , s3attcvl , s3atotmd , s3lfrcps , s3mffrcp , s3vcmfcp , s3vagdel , s3vbac , s3breec h , s3prcsec , s3rpcsec , s3flvbac , s3elcsec , s3preec l , s3ec l , s3prllbr , s3ohydr , mrvu , brvu

Counting

$C(Sex{=}Female, Area{=}Midwest, HoursPerWeek{=}40{+}) = 999$

There are 999 women in the Midwest who work more than 40 hours per week.

$C(Occupation{=}Farming\text{-}etc, HoursPerWeek{=}{<}20) = 403$

There are 403 people with agricultural occupations who work less than 20 hours per week.

Notation:

$C(a_{i(1)} = v_1, a_{i(2)} = v_2, \ldots a_{i(n)} = v_n)$ denotes the count of number of records matching all the equalities, where $\{a_{i(1)}, a_{i(2)}, \ldots a_{i(n)}\}$ is some set of attributes.

Contingency T ables

SEX	AREA	HOURS-PER-WEEK	COUNT
Female	Abroad	v0:20-	237
Female	Abroad	v1:20+	27
Female	Abroad	v2:40+	34
Female	Midwest	v0:20-	1329
Female	Midwest	v1:20+	261
Female	Midwest	v2:40+	999
Female	Northeast	v0:20-	891
Female	Northeast	v1:20+	174
Female	Northeast	v2:40+	623
Female	South	v0:20-	1996
Female	South	v1:20+	343
Female	South	v2:40+	1180
Female	Unknown	v0:20-	37326
Female	Unknown	v1:20+	3955
Female	Unknown	v2:40+	21863
Female	West	v0:20-	1697
Female	West	v1:20+	301
Female	West	v2:40+	940
Male	Abroad	v0:20-	188
Male	Abroad	v1:20+	32
Male	Abroad	v2:40+	95
Male	Midwest	v0:20-	1077
Male	Midwest	v1:20+	249
Male	Midwest	v2:40+	1214
Male	Northeast	v0:20-	625
Male	Northeast	v1:20+	149
Male	Northeast	v2:40+	717
Male	South	v0:20-	1385
Male	South	v1:20+	330
Male	South	v2:40+	1537
Male	Unknown	v0:20-	28645
Male	Unknown	v1:20+	3324
Male	Unknown	v2:40+	25837
Male	West	v0:20-	1288
Male	West	v1:20+	305
Male	West	v2:40+	1348

Contingency Tables

Every possible set of attributes $\{ a_{i(1)}, a_{i(2)}, \ldots a_{i(n)} \}$ has an associated contingency table denoted by **ct** $(a_{i(1)}, a_{i(2)}, \ldots a_{i(n)})$

It countains counts for all queries that involve the given attribute set.

Claim

For Machine Learning involving discrete (symbolic) attributes, counting or building contingency tables is very frequently the inner loop of the algorithm.

- Bayesian Networks
- Causal Networks
- Rule Learning
- Feature Selecting
- GMDH
- Frequent Sets
- Decision Tree Learning
- Decision List Learning
- TFIDF
- Bayes Classifiers
- Logistic Regression

(But Not: Nearest Neighbor, Projection Pursuit, Neural Nets, Support Vector Machines)

Building a contingency table

Start with a contingency table full of zeroes.

For each k in $\{1, 2, \ldots R\}$

- Let v_1 = Value of $a_{i(1)}$ in kth record
- Let v_2 = Value of $a_{i(2)}$ in kth record

  :
- Let v_n = Value of $a_{i(n)}$ in kth record

- Increment the $a_{i(1)} = v_1, a_{i(2)} = v_2, \ldots a_{i(n)} = v_n$ row of the contingency table by 1.

- Cost (binary attributes) is $O(2^n + n\,R)$
- This can be very expensive with a large number of records.
- Wouldn't it be nice to make it faster?

Idea 1: Pre-compute all contingency tables

After the initial pre-computation phase ,
fetc hing a contingency table would be O(L).

Two problems

- A lot of memory . The number of numbers you need to store is the number of possible counting queries:

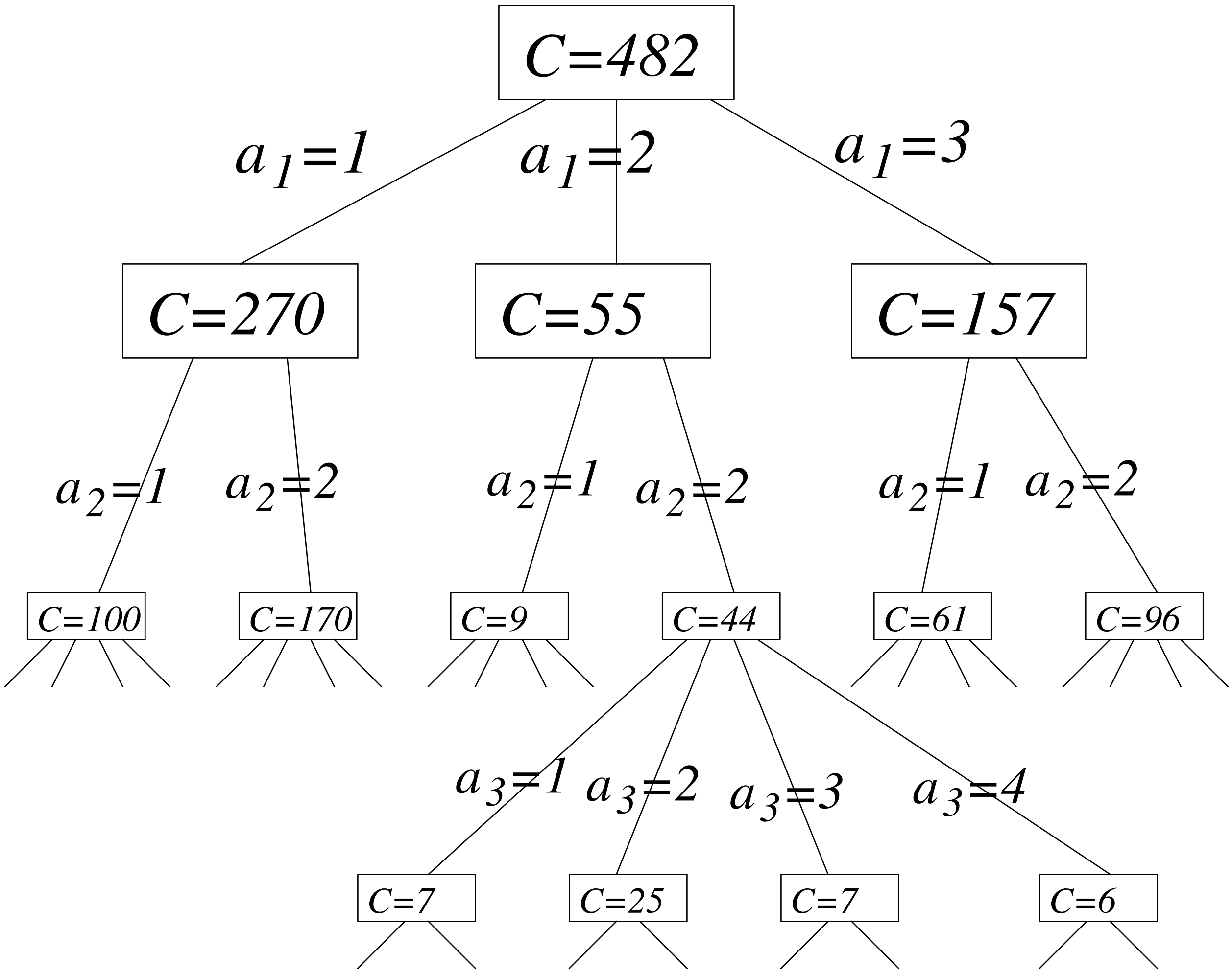
  $$\prod_{j=1}^{M} (n_j + 1)$$

  (The jth attribute ma y take on one of n_j values or it ma y be omitted...equivalent to an a_j = * "dont-care" value)

- A lot of time to pre-build

  $\prod_{j=1}^{M} (n_j + 1)$ at the very very least!

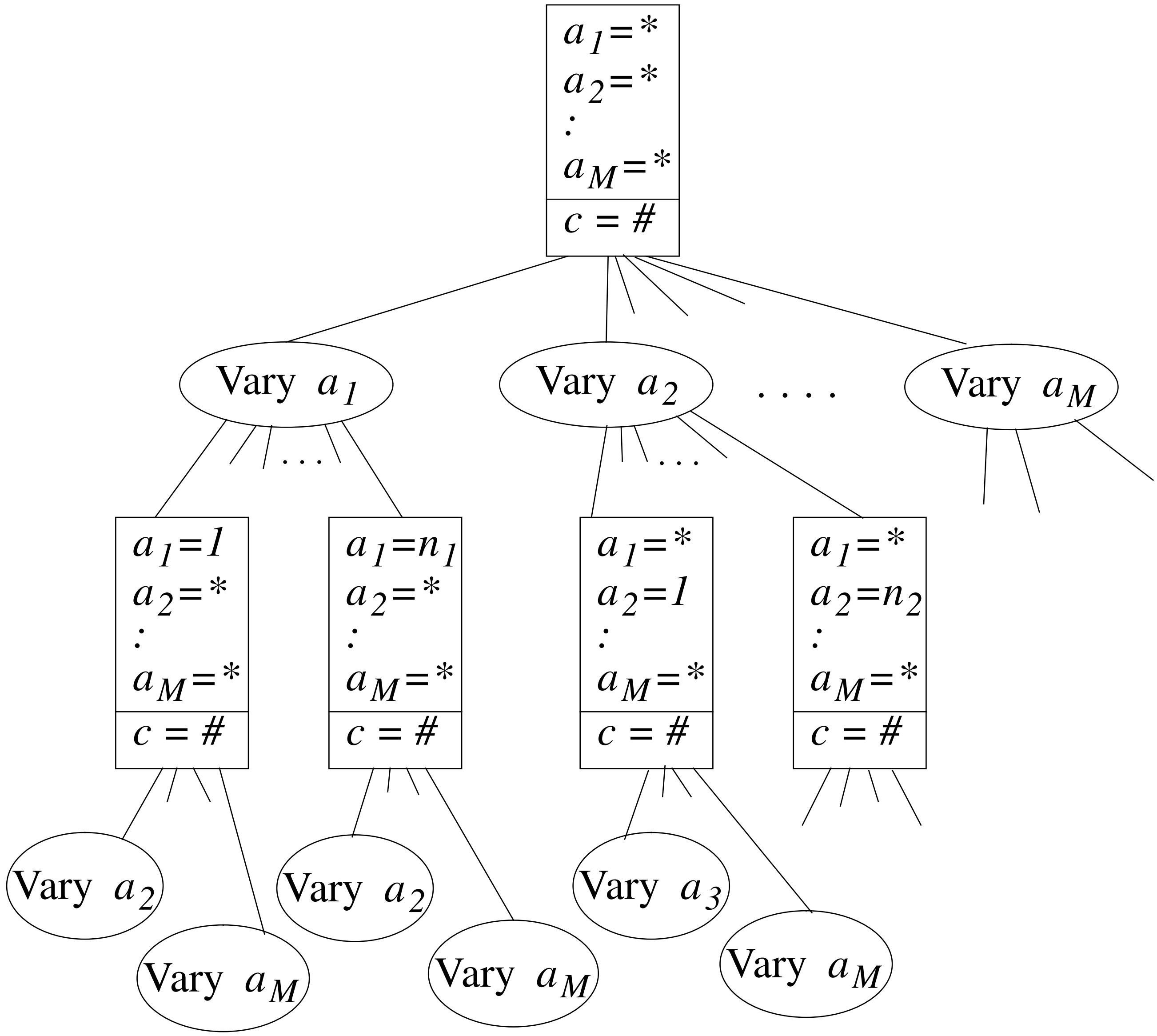
Idea 2: Use kd-trees



Good News: Cheap to compute C(a1=2,a3=4) or ct(a1,a3).

Bad News: Stupidly expensive to compute C(a1=2,a17=4) or ct(a3,a15)

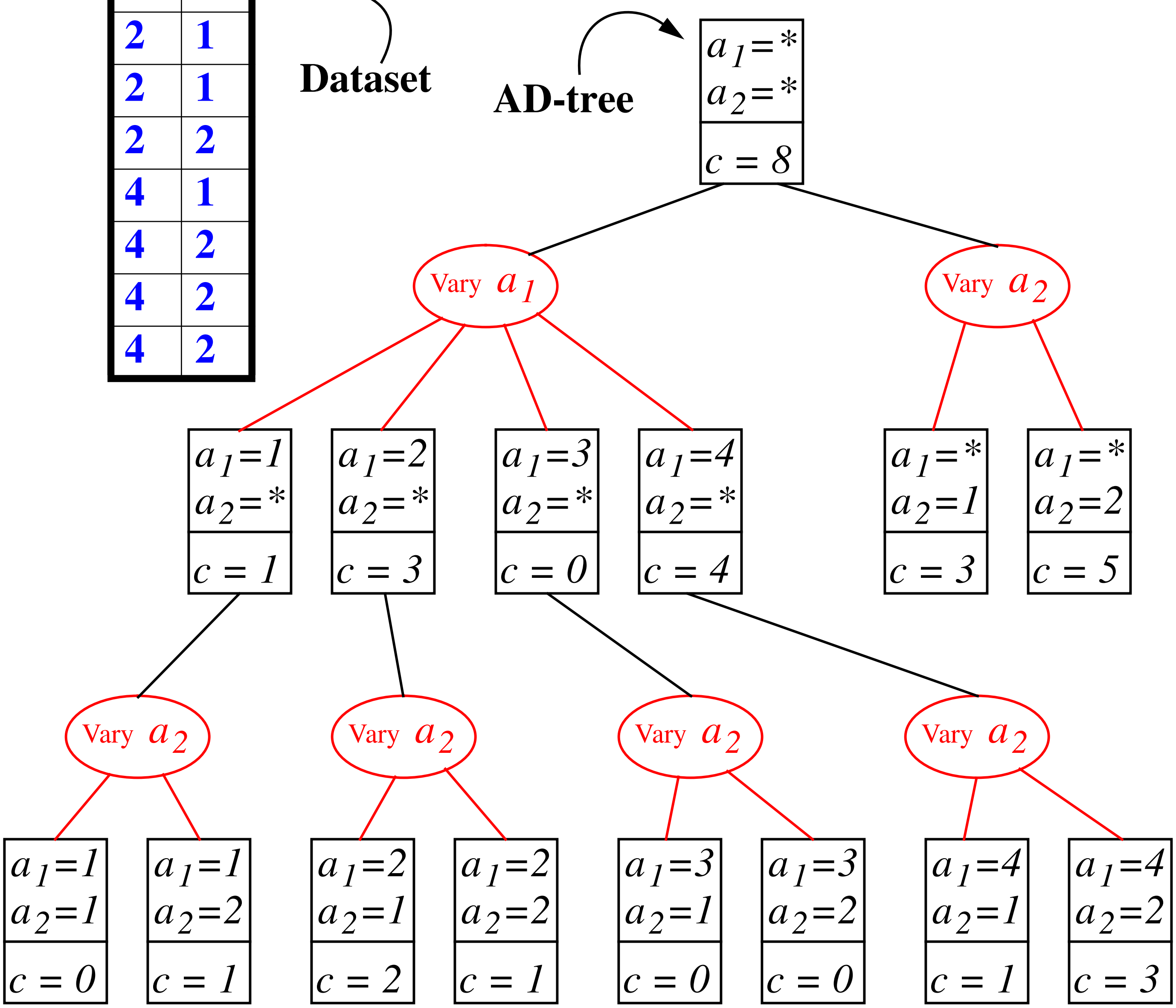
AD-tree Version 1.0



AD-T ree v1.0 Example

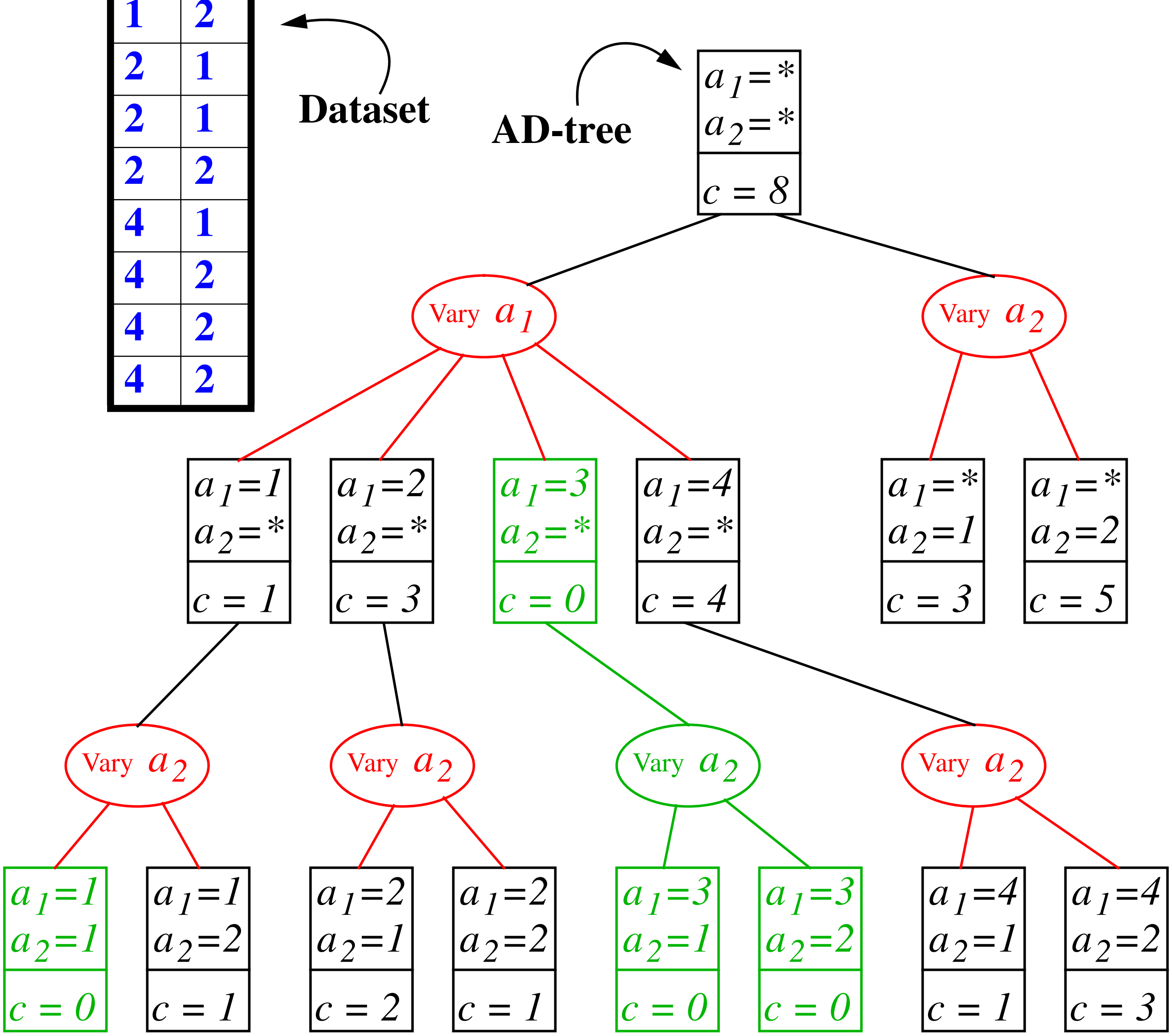
a_1	a_2
1	2
2	1
2	1
2	2
4	1
4	2
4	2
4	2

Dataset



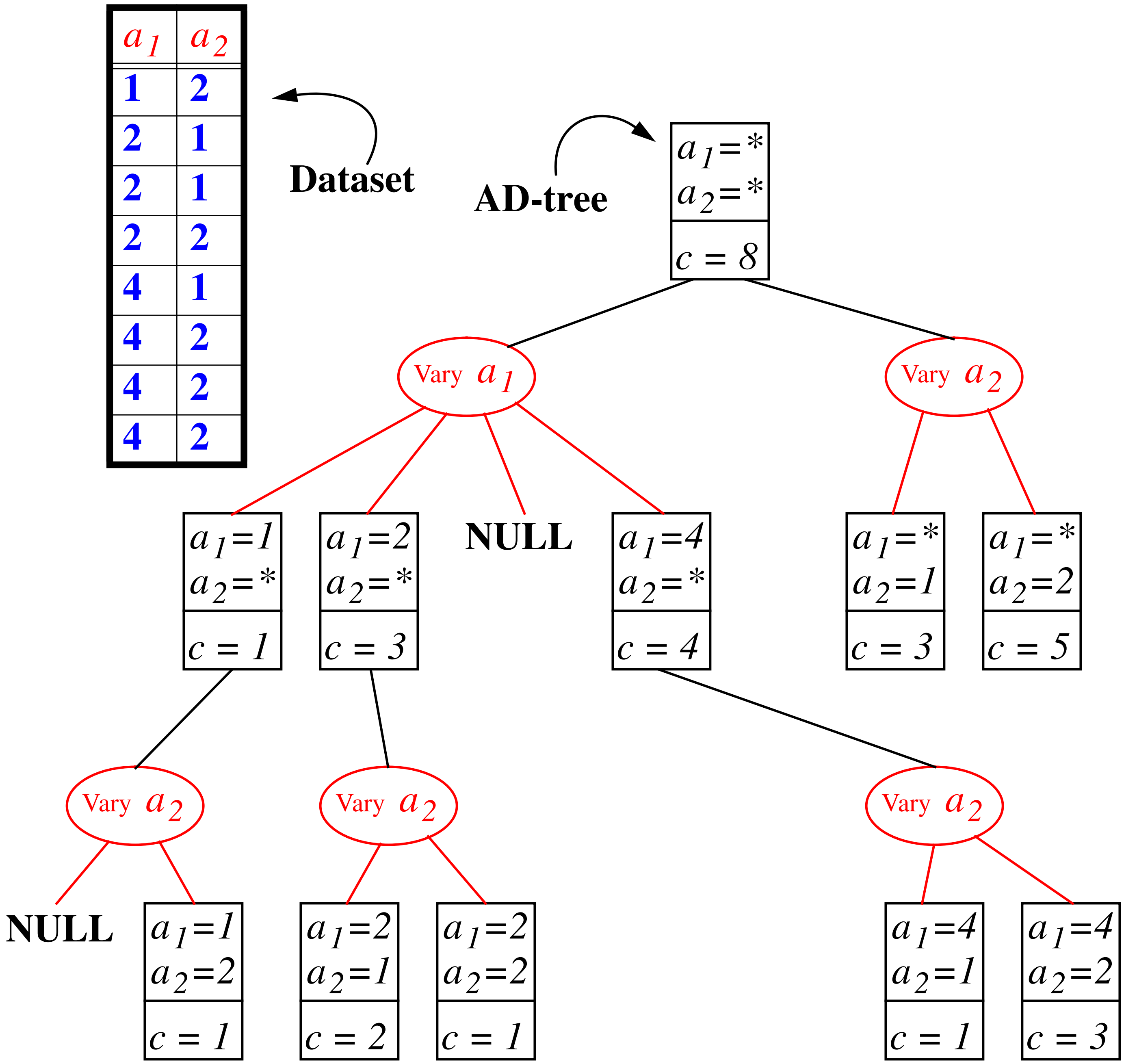
AD-T ree v1.1

a_1	a_2
1	2
2	1
2	1
2	2
4	1
4	2
4	2
4	2



A blindingly obvious improvement is to sa ve memory by storing NULL pointers for counts of zero .

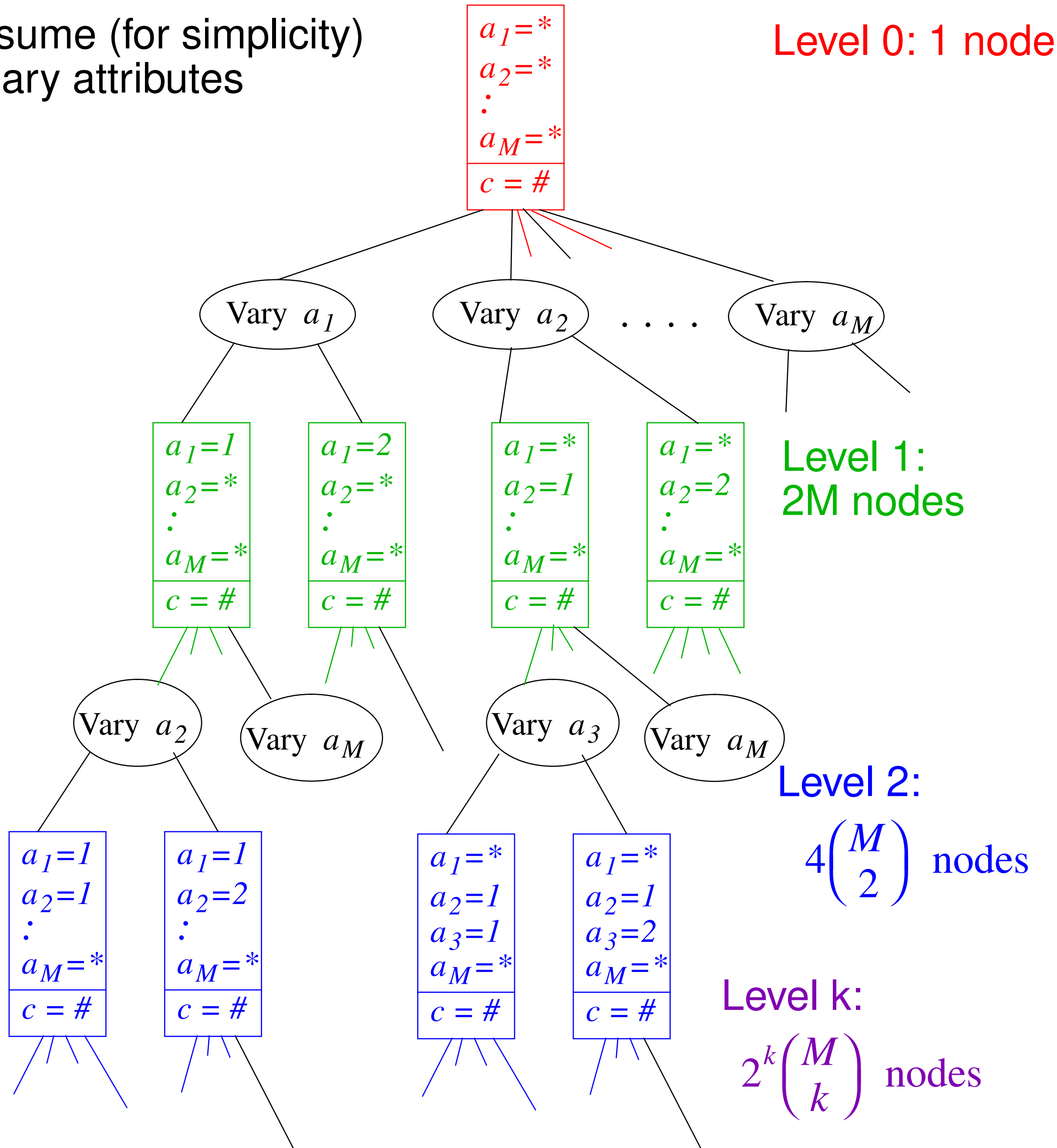
AD-Tree v1.1



A blindingly obvious improvement is to save memory by storing NULL pointers for counts of zero .

Ad-Tree v1.1 Size

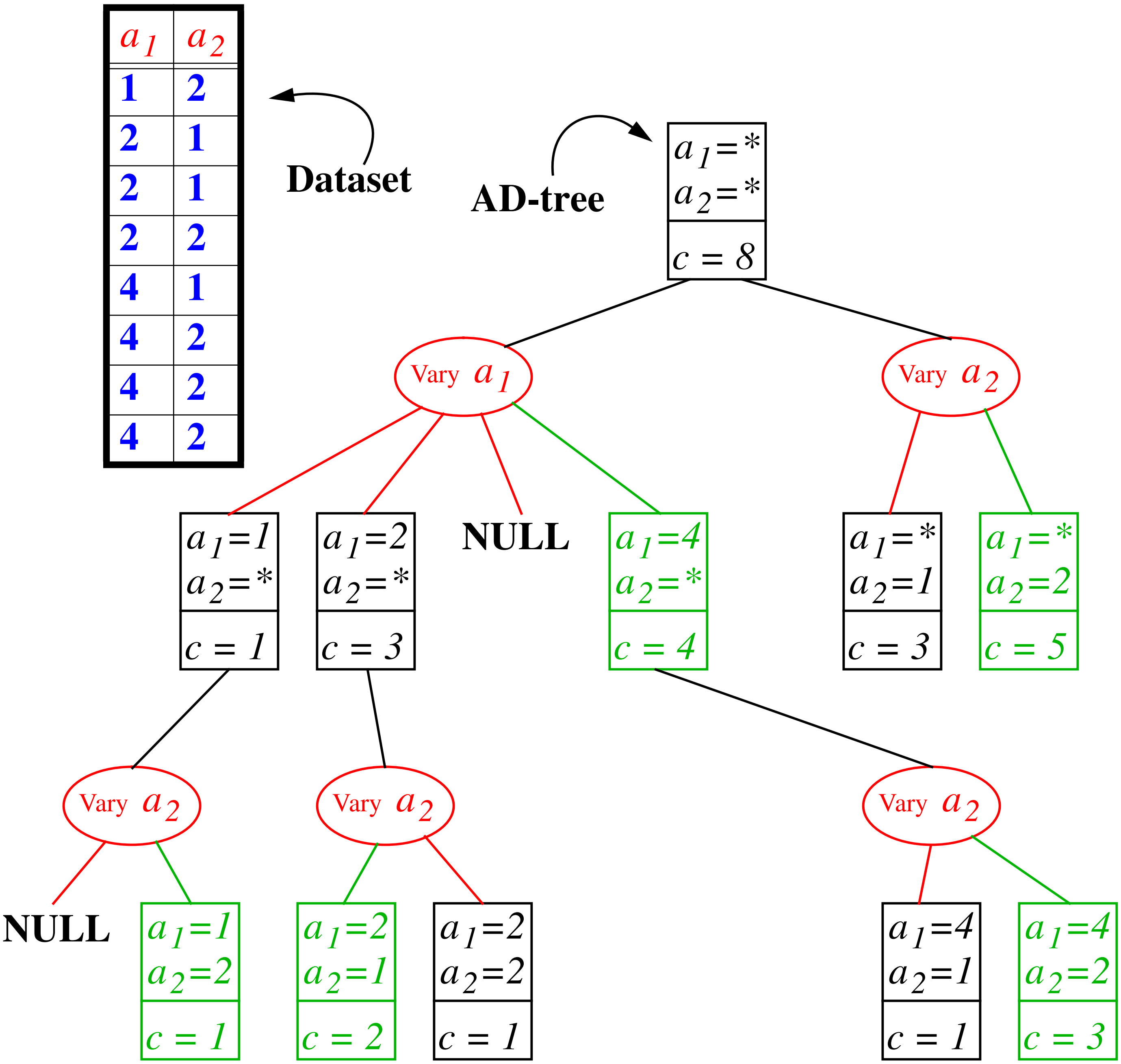
Assume (for simplicity) binary attributes



Number nodes is 3^M in the worst case. But if $R < 2^M$ then (in best case) number of nodes is about

$$\sum_{k=0}^{B} 2^k \binom{M}{k} \quad \text{where} \quad B = \log_2 R$$

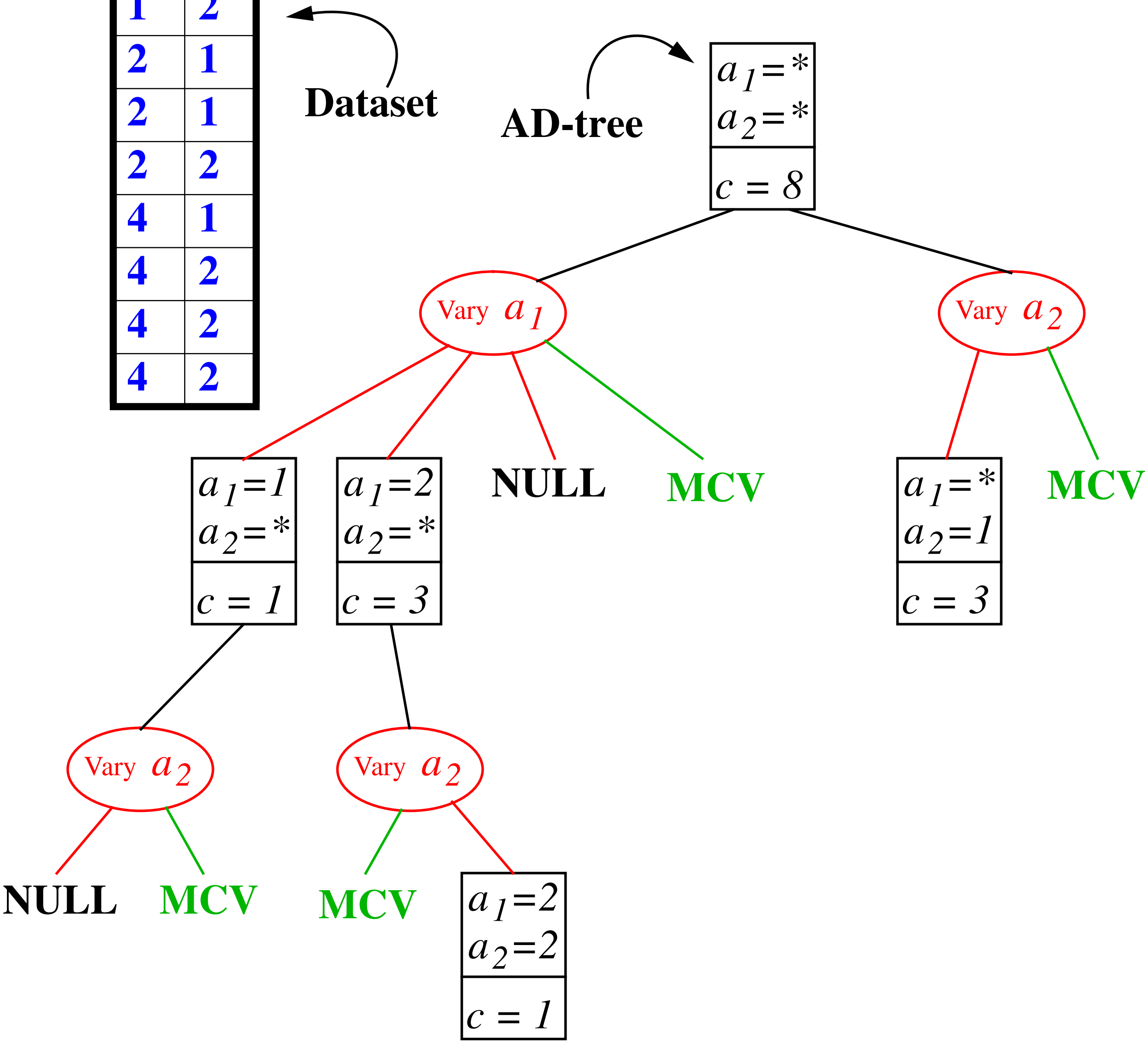
AD-Tree Version 2.0



For every Vary node in the tree, consider the Most Common Value.

AD-Tree Version 2.0

a_1	a_2
1	2
2	1
2	1
2	2
4	1
4	2
4	2
4	2



Replace any subtree that is topped by a MCV node with a single token: **MCV**

Wonderfully, counterintuitively, you've lost nothing!!!!