DMML, 21 Jan 2020

Constructing decision trees

Building smallest tree is NP-Complete

Greedy heuristic — maximize improvement in purity

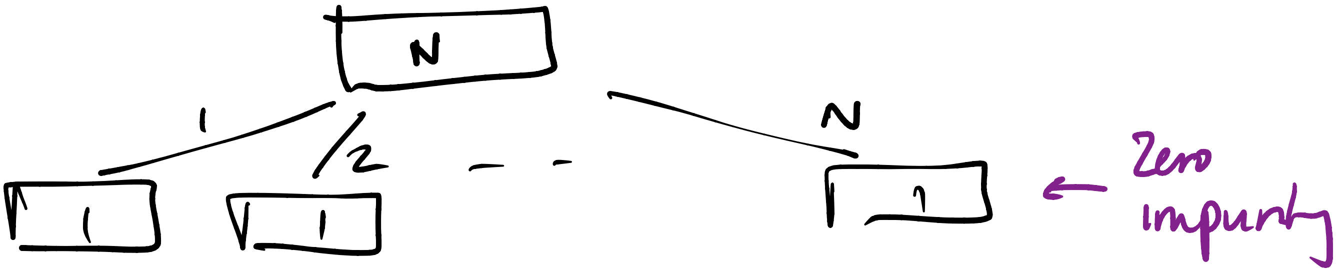
Measuring impurity

- Misclassification rate
- Entropy — "Information gain"
- Gini Index

Problem with using information gain directly:

Attribute = Aadhaar no / Passport No / Serial No / ...

Suppose we pick Aadhaar as next question



Highest possible information gain, but totally useless

Penalize attributes with too many possible values

- Compute entropy / Gini index of attribute itself !

Aadhaar — N values, each once in table

$p_i = \frac{1}{N}$ for each value

$$- \sum_{i=1}^{N} p_i \log p_i = - \sum_{i=1}^{N} \frac{1}{N} \log \frac{1}{N}$$

$$= -\log \frac{1}{N} = \log N$$

information-gain (A_i) – improvement in purity if we choose A_i

entropy (A_i)

$$\text{information-gain-ratio}(A_i) = \frac{\text{information-gain}(A_i)}{\text{entropy}(A_i)}$$

Two well known implementations of decision trees

- CART – Classification & Regression Tree

  └ Gini Index   Breiman et al

- C4.5 – Quinlan – Entropy

Continuous attributes ?   e.g "Salary"

"Cost"

May know lower & upper bound for A_i

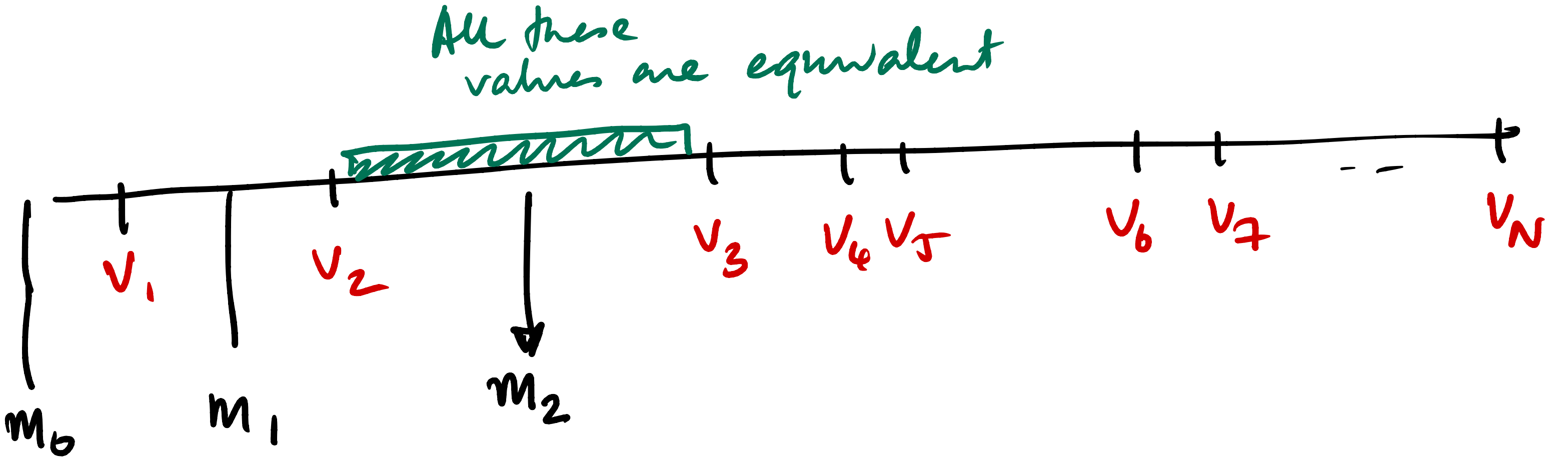
- Granularity

Typical question should be

How do we choose v? → $\left[\begin{array}{l} A_i \leq v\,? \\ A_i \geq v\,? \\ A_i = v\,? \end{array}\right.$

All we know about A_i is what we see in table

$\leq N$ values across N rows $\quad v_1 < v_2 < \dots < v_N$



$A_i < m_j$?

Try each m_j, evaluate purity gain, keep the best

Should we choose midpoints for m_j, or actual v_j'?

Better to use actual v_j's. — for interpretability of model not all values are possible

CART

Regression – produce a number as an answer

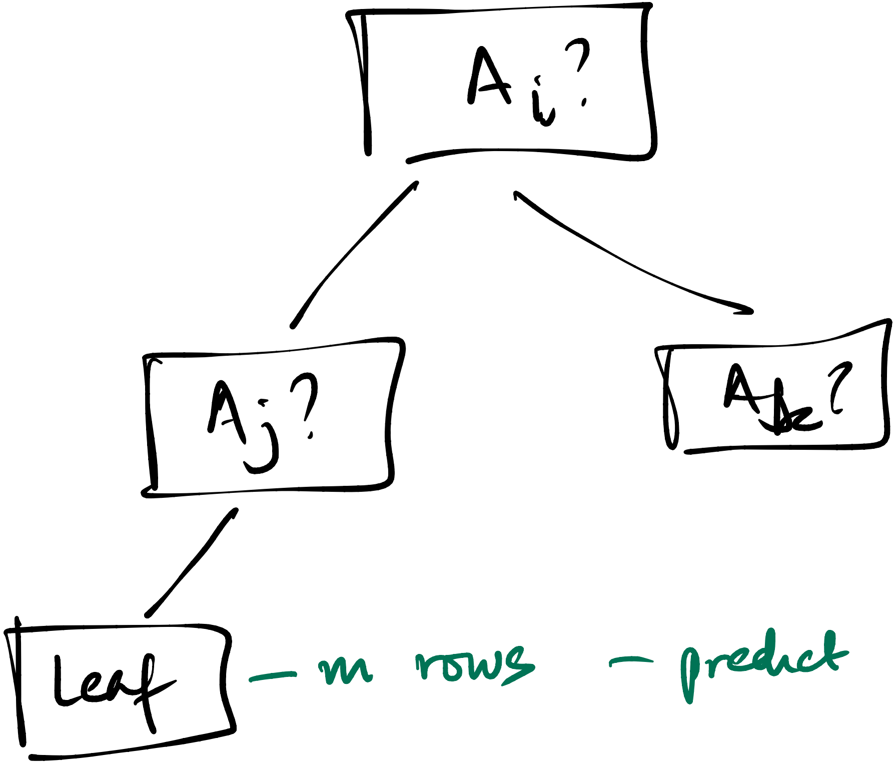
└ Typically, fit a function $f(x_1,..,x_n)$ to data

Training data has a numeric value as target

Age	Education	...		Salary

Regression Tree

Like a classification tree



— m rows — predict mean salary

Variance is a good error metric

CART – Classification And Regression Tree

└ Only asks binary questions

How good is the classifier?

- What is the correct measure?
- On what data can we compute the measure?

  └ Need data for which correct ans is known

Only labelled data we have is training data

Withhold some training data for testing

Typically 70% to build model, 30% to test

(Be careful to select 30% "randomly"

- Build a model on training set
- Evaluate on test set

Sometimes - training data is too sparse to "waste"

Cross Validation

- Choose different subsets as test data
- Repeatedly build models

10-fold crossvalidation

- 90% train, 10% test

  └ covers whole data across 10 iterations

If results are good, go back and build model on full data