DMML, 29 Jan 2019

Decision Trees

- Mathematically define impurity, to select attributes to explore on a path

Guard against "high entropy" attributes

- Information gain ratio

What about continuous (numerical) attributes?

Age ⟶ Young / Middle / Old

Numeric — Categorical

How do we create these boundaries?

Suppose column is Age, in no. of years

Find a cutoff

Ask Age ≤ Cutoff ?

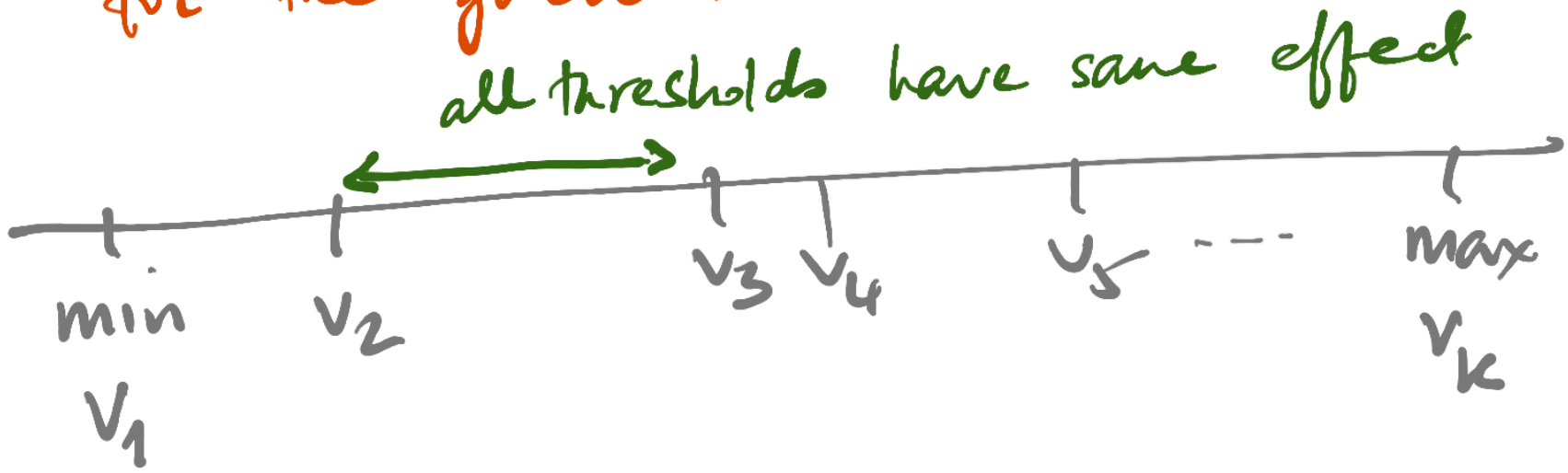
Y ／ ＼ N

Choosing the cutoff

Training data is a finite table

Table has fixed values $v_1, v_2, \ldots, v_k$

for the given attribute



$k-1$ "logical" thresholds corresponding to $k-1$ intervals $v_i - v_{i+1}$

Which interval to pick?

- Pick each possible threshold, compare its usefulness (information gain)

Choice of threshold $\overline{v_i \quad v_{i+1}}$

$\leq \frac{v_i + v_{i+1}}{2}$

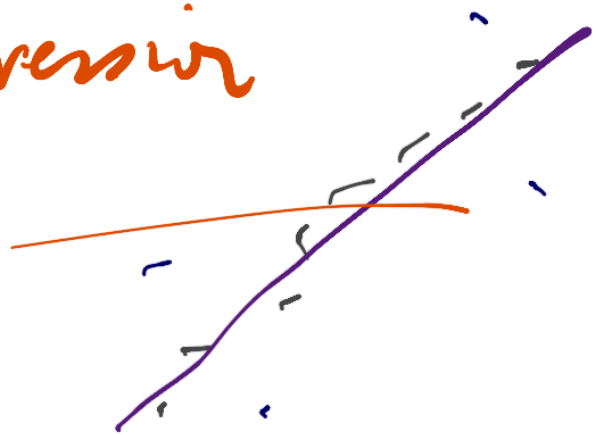
$\leq v_{i+1}$ ← May be more "interpretable"

Construction of decision tree depends heavily on training data

- Small perturbation in input can drastically change the tree

Variance

In contrast: linear regression



How to evaluate a model?

Split input data as training data & test data (maybe 75% – 25%)

Build model on training data

Validate on test data

Careful to choose "randomly"

Cross Validation

Take out 10% as test data by turns

Build 10 models, each with 90% of data

After cross validation, what model to use?

- Build a fresh model on 100% data
- Use voting across multiple models

What is a good measure of correctness

Accuracy - Fraction of correct answers

Problem - Most classification problems are highly asymmetric

E.g. Suppose 1% of card transactions are fraud

Report "Not fraud"

- 99% accurate

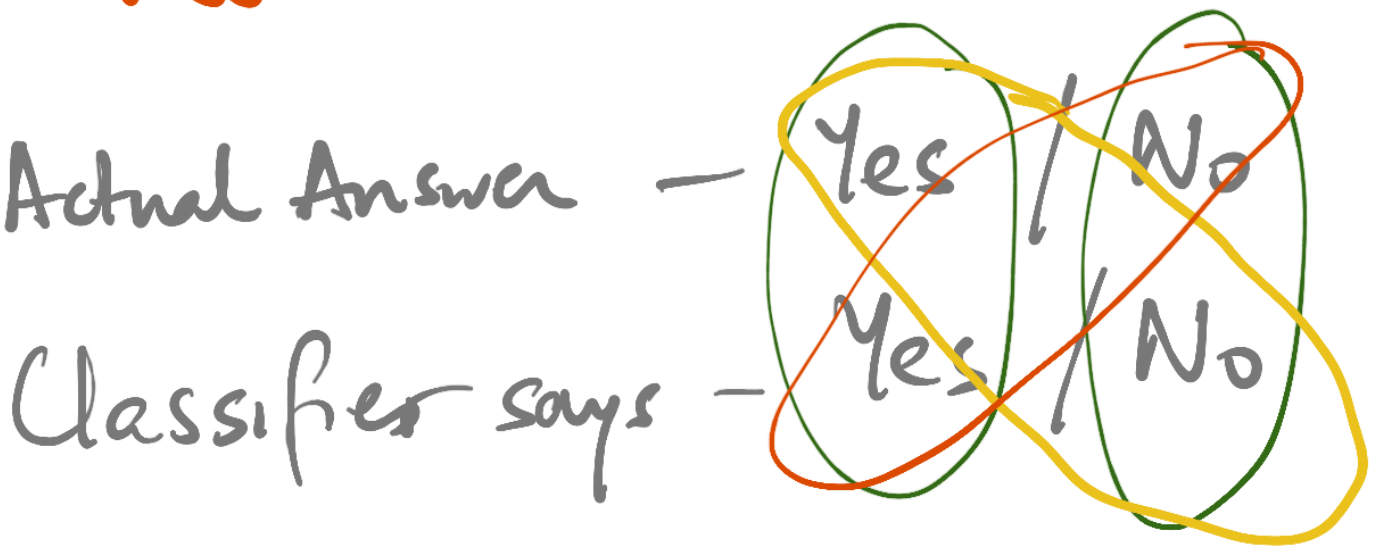
Want classifier to identify minority case

Need a better measure

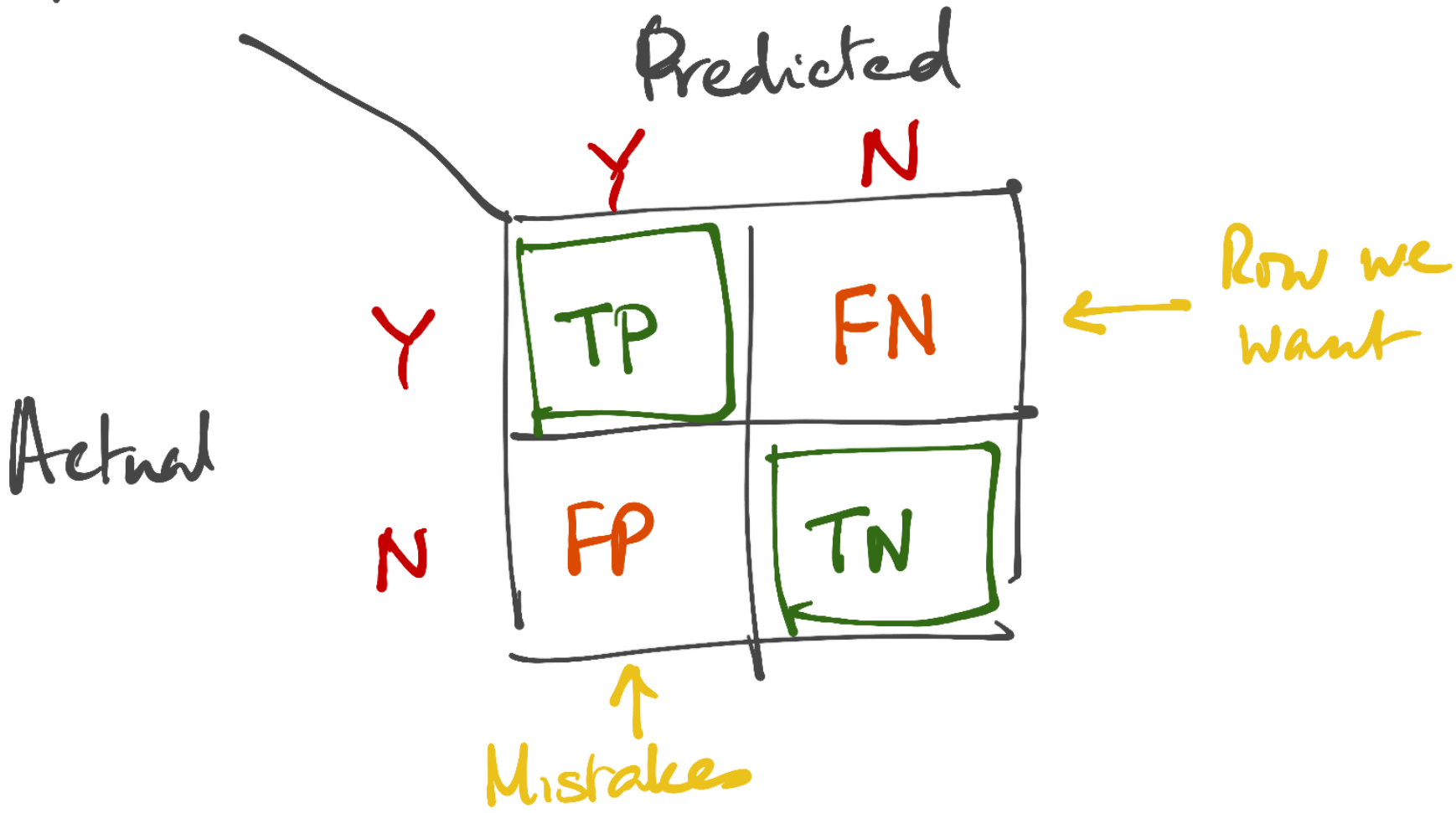
Assume answers are Yes/No

Yes is the minority case (Fraud, Spam, ..)

A good fraction of true Yes cases are flagged by classifier



Confusion Matrix

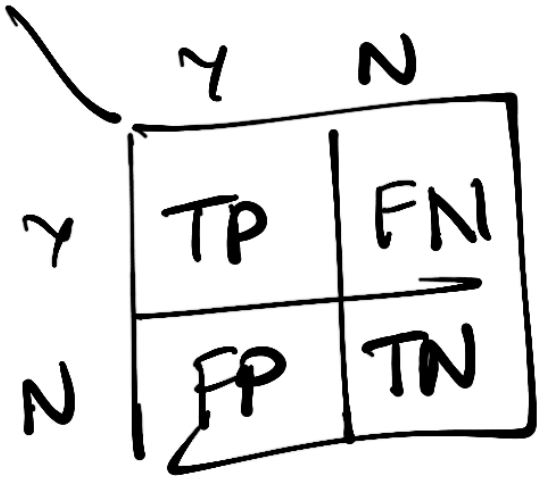


True Positive

True Negative

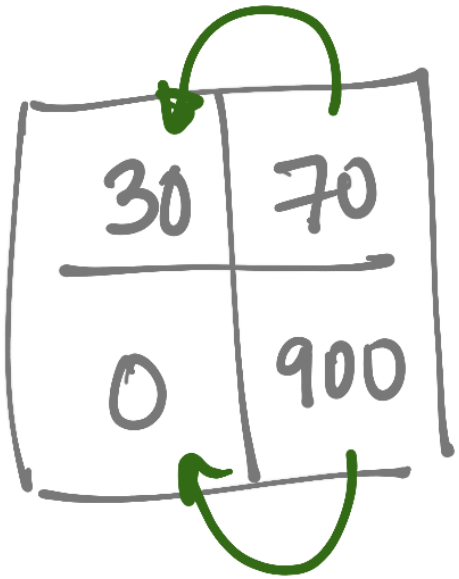
False +ve

False -ve



Recall $\frac{TP}{TP+FN}$ Pred Y / Act Y

Precision $\frac{TP}{TP+FP}$ Act Y / Pred Y



Recall = 0.3

Precision = 1

Recall ↑

Precision ↓

Example

Screening Test vs Interview

High Recall

↓

HINI

High Precision

↓

Pancreatic Cancer

Medical Diagnosis

Make a choice!

Overfitting*

Building a model too close to training data

* Model behaves well on training data but generalizes poorly

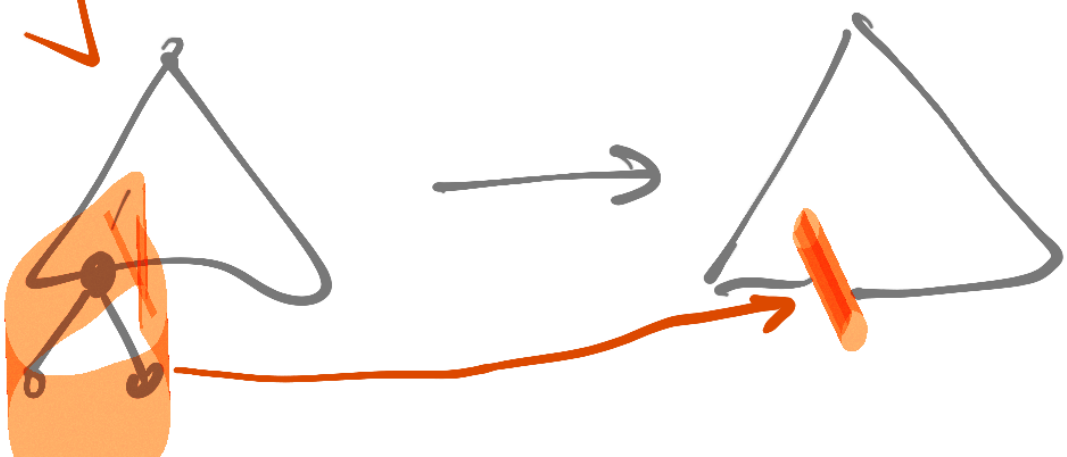
Asking too many questions

Pruning a tree

Forward – stop exploring if info gain drops below some threshold

– How to get this threshold right?

Backward – build full tree & start deleting leaves



Quinlan's strategy

Sampling theory

Extrapolate from 71 heads/100 coin tosses

to $P(\text{heads}) = 0.71 \pm \epsilon$

Original decision tree:

```
physician fee freeze = n:
|   adoption of the budget resolution = y: democrat (151)
|   adoption of the budget resolution = u: democrat (1)
|   adoption of the budget resolution = n:
|   |   education spending = n: democrat (6)
|   |   education spending = y: democrat (9)
|   |   education spending = u: republican (1)
physician fee freeze = y:
|   synfuels corporation cutback = n: republican (97/3)
|   synfuels corporation cutback = u: republican (4)
|   synfuels corporation cutback = y:
|   |   duty free exports = y: democrat (2)
|   |   duty free exports = u: republican (1)
|   |   duty free exports = n:
|   |   |   education spending = n: democrat (5/2)
|   |   |   education spending = y: republican (13/2)
|   |   |   education spending = u: democrat (1)
physician fee freeze = u:
|   water project cost sharing = n: democrat (0)
|   water project cost sharing = y: democrat (4)
|   water project cost sharing = u:
|   |   mx missile = n: republican (0)
|   |   mx missile = y: democrat (3/1)
|   |   mx missile = u: republican (2)
```

97 R
3 D

2R

After pruning:

```
physician fee freeze = n: democrat (168/2.6)
physician fee freeze = y: republican (123/13.9)
physician fee freeze = u:
|   mx missile = n: democrat (3/1.1)
|   mx missile = y: democrat (4/2.2)
|   mx missile = u: republican (2/1)
```

Now, this description does violence to statistical notions of sampling and confidence limits, so the reasoning should be taken with a large grain of salt. Like many heuristics with questionable underpinnings, however, the estimates that it produces seem frequently to yield acceptable results.