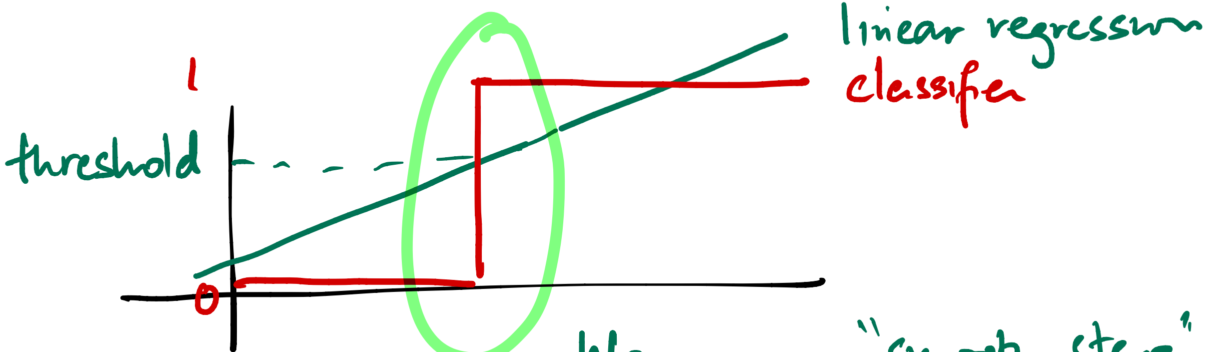
DMML, 28 Jan 2020

Decision Trees → CART → Regression

Regression - predicting a value

Can use a threshold on predicted value to classify

≥ 50% mock exam prediction ⇒ will pass boards

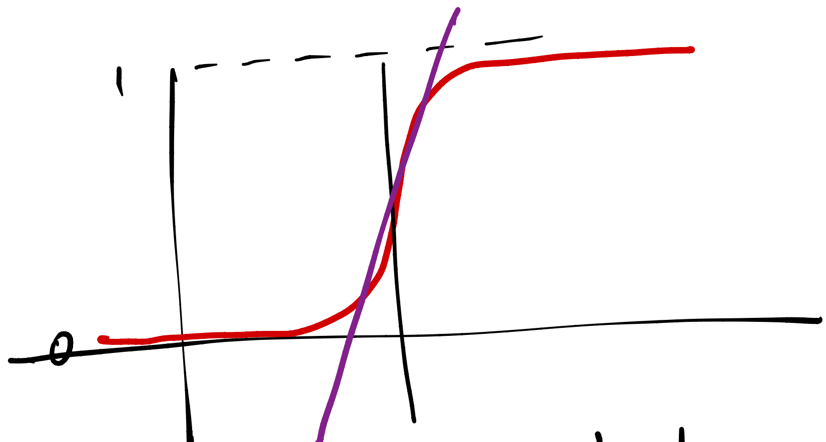


Want a "smooth step"

Sigmoid function

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$z = mx + b$ is our linear regression output



determined by b

determined by m

$z \to \infty$

$$\frac{1}{1+e^{-z}} \to \frac{1}{1} = 1$$

$z \to -\infty$

$$\frac{1}{1+e^{\infty}} \to 0$$

Logistic Regression

Minor point.

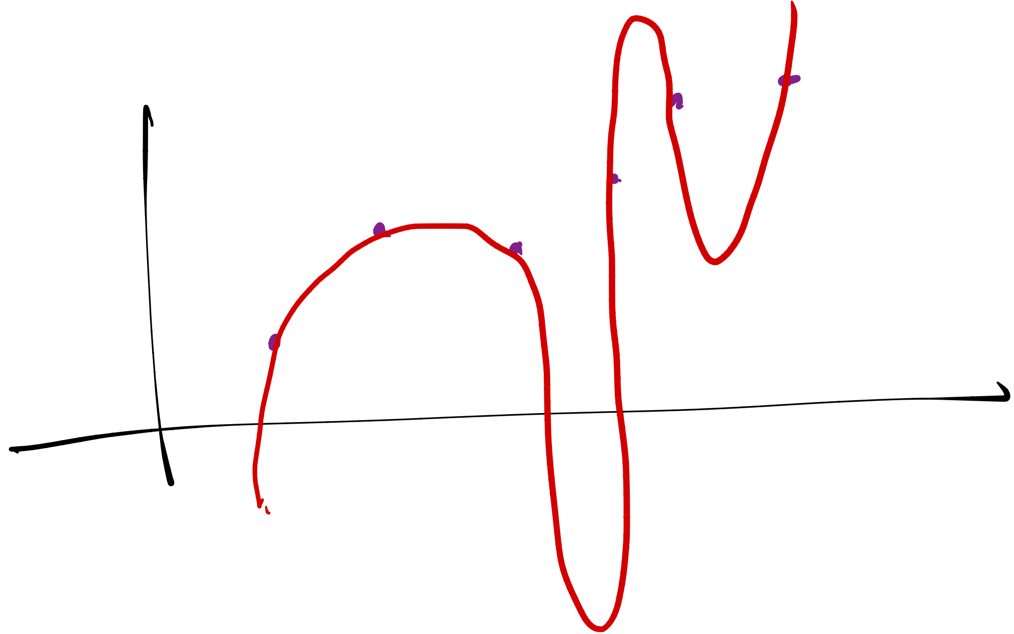
Typically $x_1, x_2, \ldots, x_n$

$$y = w_1x_1 + w_2x_2 + \cdots + w_nx_n + w_0$$

Many implementations will extend $x_1 \ldots x_n$ with $x_0 = 1$

$$w_0x_0 + w_1x_1 \ldots + w_nx_n = \vec{w} \cdot \vec{x}$$

Higher dimension regression - Can always exactly fit n pt with $n-1$ degree polynomial



Overfitting

Model fits training data "too well" – performs poorly on unseen data

There is another model that performs worse on this training data, but better in general

Decision trees also can overfit

Too many questions

Shallow tree may in fact "generalize" better

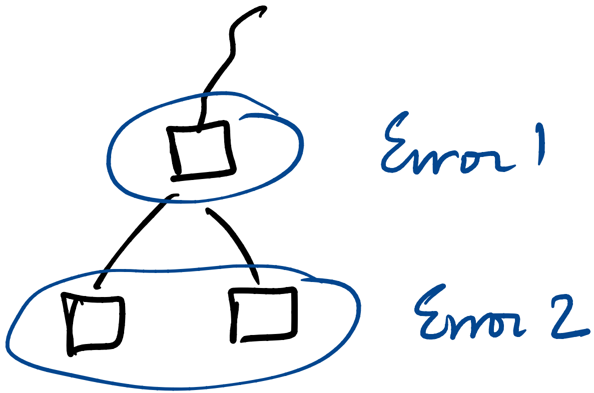
Prune the tree

Pre-pruning – limit the depth in advance or based on threshold of improvement

Post-pruning – build the tree in full & remove leaves bottom up

Quinlan (C4.5)

Estimate error before after pruning



Error function cannot be misclassification rate

Error2 < Error1 always

Quinlan - use sampling theory

Want to estimate some voter preference

Assume random sampling works

How big a sample do we need?

Depends on expected outcome

10,000 voters. Either 9999 prefer A

or 9999 prefer B

Sample size 3 suffices

As ratio goes toward 50%, sample size increases

Allow an error of δ

Check if candidate A will get $\geq p \pm \delta$ percentage of votes

Can compute sample size to guarantee this

Quinlan's interpretation.

$$\boxed{\begin{matrix} 9\ Y \\ 3\ N \end{matrix}}$$

Interpret as sample size $= 12$

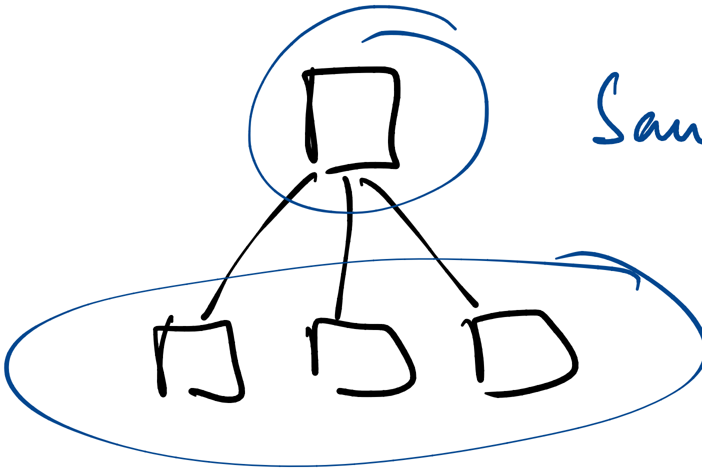
$$p = \frac{9}{12}$$

compute δ

Suppose I see 7 Heads in 10 coin tosses

70 Heads in 100 coin tosses

700 Heads in 1000 coin tosses



Sampling error E_1

Weighted sampling error E_2

$E_2 > E_1$ is possible – prune

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)

Using sampling error as a prediction of error rate

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)

QUINLAN : 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.

Nevertheless, overfitting is a very standard problem

More general strategy — penalize model complexity

Regression : $\underbrace{\text{Usual cost}}_{\text{MSE}} + \underbrace{\delta \cdot f(\text{coefficient size})}_{\text{Model cost}}$

Regularization

└ Simplifying model by adding structural penalty to overall cost