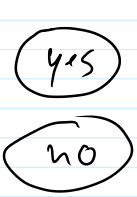
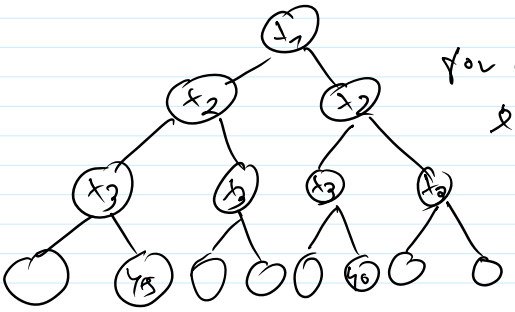
Example: $P(x,y)$ with properties as follows
- $P(Y=yes) = P(Y=no) = 0.5$
- three binary features $x_1 \dots x_3$
- there is a deterministic relationship between x and y

DT learner $A_k(S)$ via ERM: Pick k-node tree that minimizes $err_S(h)$

$k=1$ → $k = 2^4 - 1$



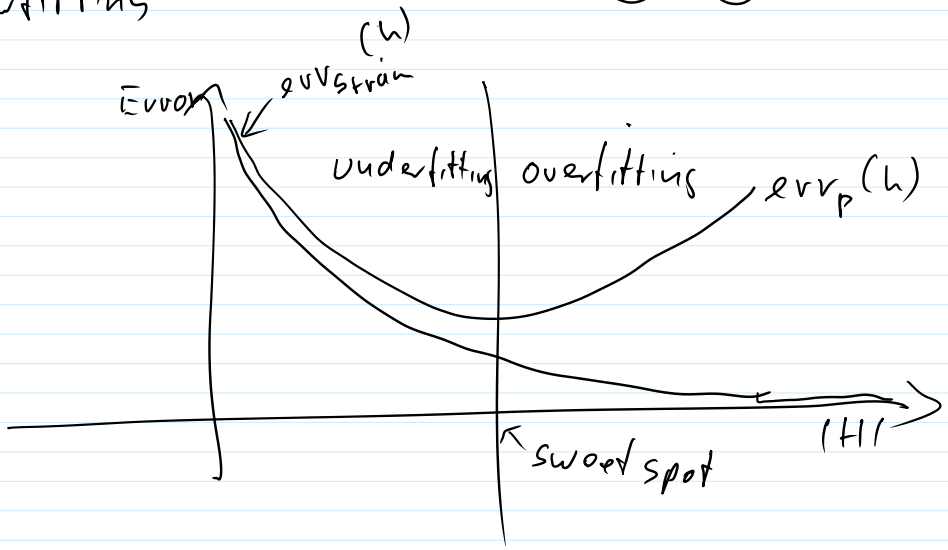


for many h $err_S(h) = 0$

overfitting

$\forall h: Err_S(h) \approx 0.5$

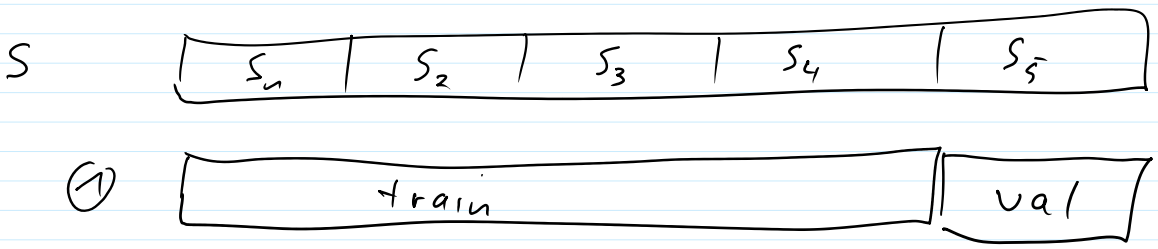
underfitting

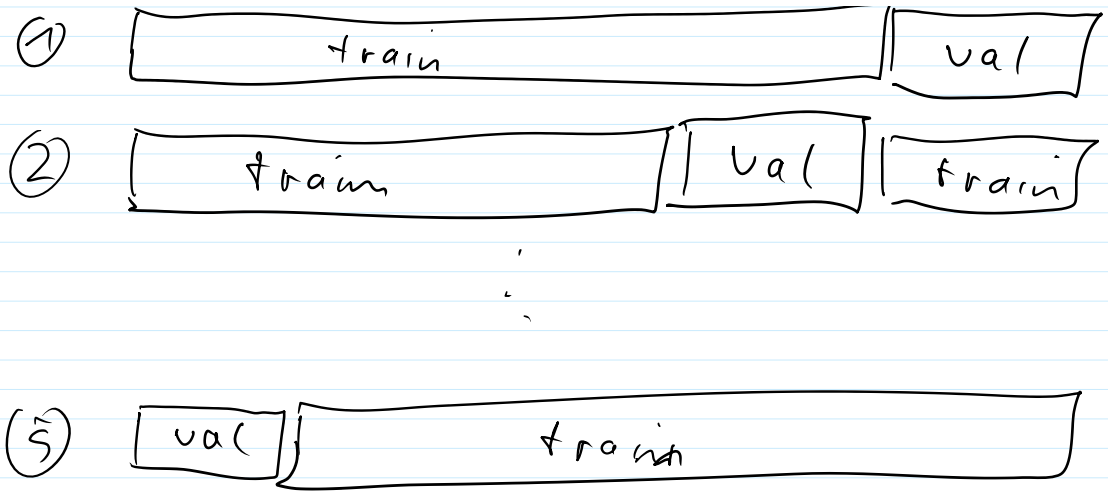


Model selection → secondary learning problem

$H = \{h_{full}, h_{early}, h_{pruning}, h_{rule}\}$

→ pick minimizing $err_{S_{val}}(h)$ → $\hat{h}$





Scenario: Is text classification rule accurate enough for shipping?
($err_p(h) < 0.1$)

Null hypothesis: $err_p(h) \geq 0.1$

$n = 600$

$x = 59$ → What is the probability of seeing 59 test errors under the null hypothesis's

$p = 0.1$

$Binom(X \leq 59 \mid p=0.1, n=600) = 0.426$

→ cannot reject

Null hypothesis: $err_p(h) \geq 0.2$

$Binom(X \leq 59 \mid p=0.2, n=600) = 10^{-12}$

→ reject null hypothesis

95% confidence interval: $err_p(h) \in [0.043, 0.154]$

Scenario: Is early stopping DT more/less accurate than full DT?

Sign test:

d_1 = number of examples h_1 makes error but not h_2

d_2 = " " " h_2 " " " h_1

$n = 600$

h_1 makes 59 errors

h_1 makes 57 errors

h_2 makes 66 errors

$d_1 = 1$, $d_2 = 8$

Null hypothesis: $err_p(h_1) = err_p(h_2)$

$\rightarrow D_1$ is $Binom(D_1 \mid p=0.5, n=d_1+d_2)$

$Binom(D_1 \leq 1 \mid p=0.5, n=9) = \cancel{0.002}$ 0.02 $\rightarrow$ reject null hypothesis