model overfitting

underfitting = model is too simple, both training and test errors are large

overfitting = model is too complex, training error is smal, but test error is large

- increasing the size of training data reduces the differences between training and testing errors

reasons for overfitting = not enoug training data or high model complexity

multiple comparison procedure =

Day 1	Up
Day 2	Down
Day 3	Down
Day 4	Up
Day 5	Down
Day 6	Down
Day 7	Up
Day 8	Up
Day 9	Up
Day 10	Down

$$P(\text{number of correct guesses} \geq 8) = \frac{\binom{10}{8}+\binom{10}{9}+\binom{10}{10}}{2^{10}} = 0.0547$$

$$P(\text{among 50 analyst, at least one makes at least 8 correct guesses}) = 1 - (\underbrace{1-0.0547}_{\text{all 50 not guess more than 8}})^{50} = 0.9399$$

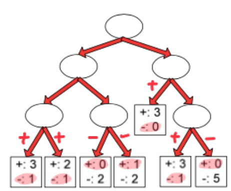
generalization error (model) = training error (model, training data) + $\alpha \times$ complexity (model)

complexity of decision trees → # of leaf nodes

pessimistic error estimate =

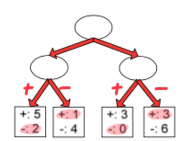
$$err_{gen}(T) = err(T) + \Omega \times \frac{k}{N_{train}}$$

- err(T): error rate on all training records
- Ω: trade-off hyper-parameter (similar to α)
  - Relative cost of adding a leaf node
- k: number of leaf nodes
- N_{train}: total number of training records



$e(T) = \frac{4}{24}$

$e_{gen}(T) = \frac{4}{24} + 1 \cdot \frac{7}{24} = \frac{11}{24}$



$e(T) = \frac{6}{24}$ $\quad e_{gen}(T) = \frac{6}{24} + 1 \cdot \frac{4}{24} = \frac{10}{24}$

→ optimistic errors

optimistic error estimate = $err_{gen}(T) = err(T)$ using only training error = resubstitution estimate

minimum description length (MDL) = $\underbrace{\text{Cost (Model, Data)}}_{\text{\# of bits needed for encoding}} = \underbrace{\text{Cost (Data | Model)}}_{\text{encoding of misclassification err}} + \alpha \underbrace{\text{Cost (Model)}}_{\text{encoding of nodes + split condition}}$