Prediction and Overfitting

CS4780/5780 – Machine Learning
Fall 2012

Thorsten Joachims
Cornell University

Reading: Mitchell Sections 3.6 – 3.7

Learning as Prediction

Definition: *A particular* **instance of a learning problem** *is described by a probability distribution* $P(X, Y)$.

Definition: *A* **sample** $S = ((\vec{x}_1, y_1), ..., (\vec{x}_n, y_n))$ *is independently identically distributed (i.i.d.) according to* $P(X, Y)$ *if*

$$P(S = ((\vec{x}_1, y_1), ..., (\vec{x}_n, y_n))) = \prod_{i=1}^{n} P(X = \vec{x}_i, Y = y_i)$$

Sample Error and Generalization Error

Definition: *The* **error on sample** S $Err_S(h)$ *of a hypothesis* h *is* $Err_S(h) = \frac{1}{n}\sum_{i=1}^{n} \Delta(h(\vec{x}_i), y_i)$.

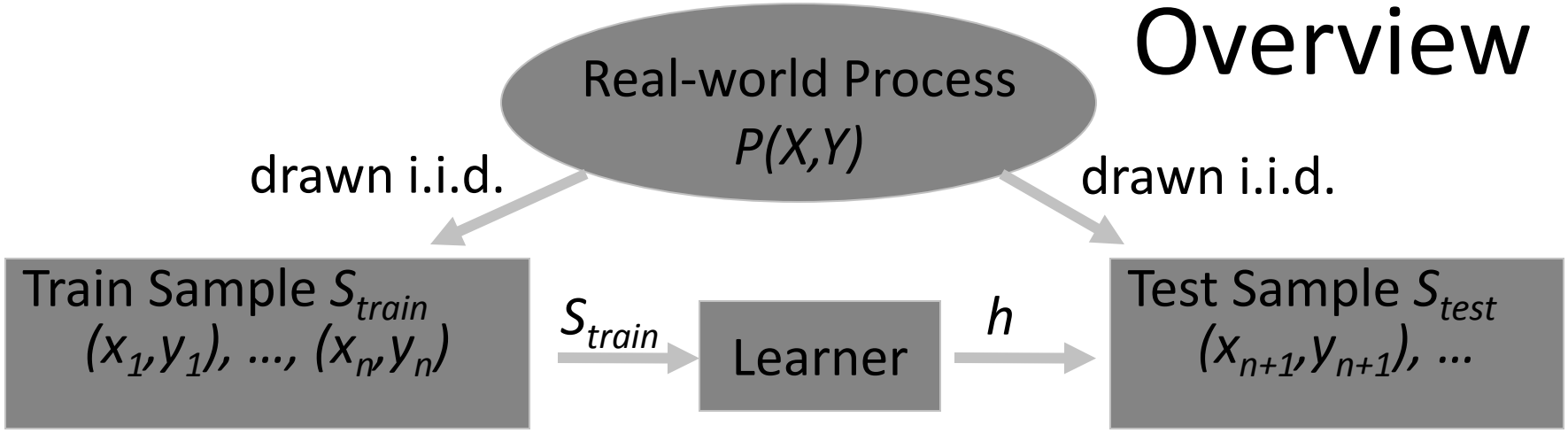
Definition: $\Delta(a, b)$ *is the 0/1-loss function*

$$\Delta(a,b) = \begin{cases} 0 & if(a == b) \\ 1 & else \end{cases}$$

Definition: *The* **prediction/generalization/true error** $Err_P(h)$ *of a hypothesis* h *for a learning task* $P(X, Y)$ *is*

$$Err_P(h) = \sum_{\vec{x} \in X, y \in Y} \Delta(h(\vec{x}), y) P(X = \vec{x}, Y = y).$$

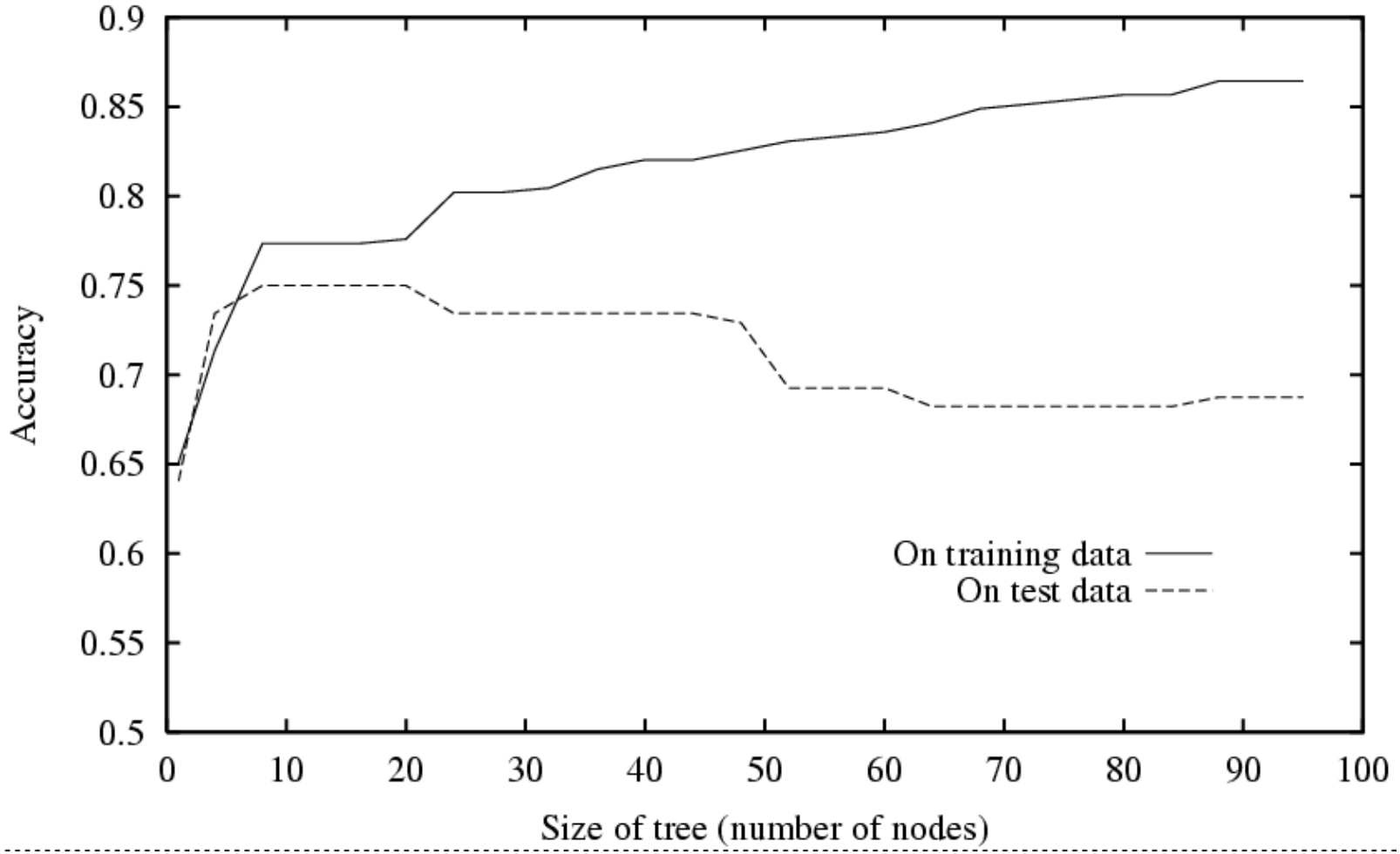
Learning as Prediction Overview



- Goal: Find h with small prediction error $Err_P(h)$ over $P(X,Y)$.
- Strategy: Find (any?) h with small error $Err_{S_{train}}(h)$ on training sample S_{train}.

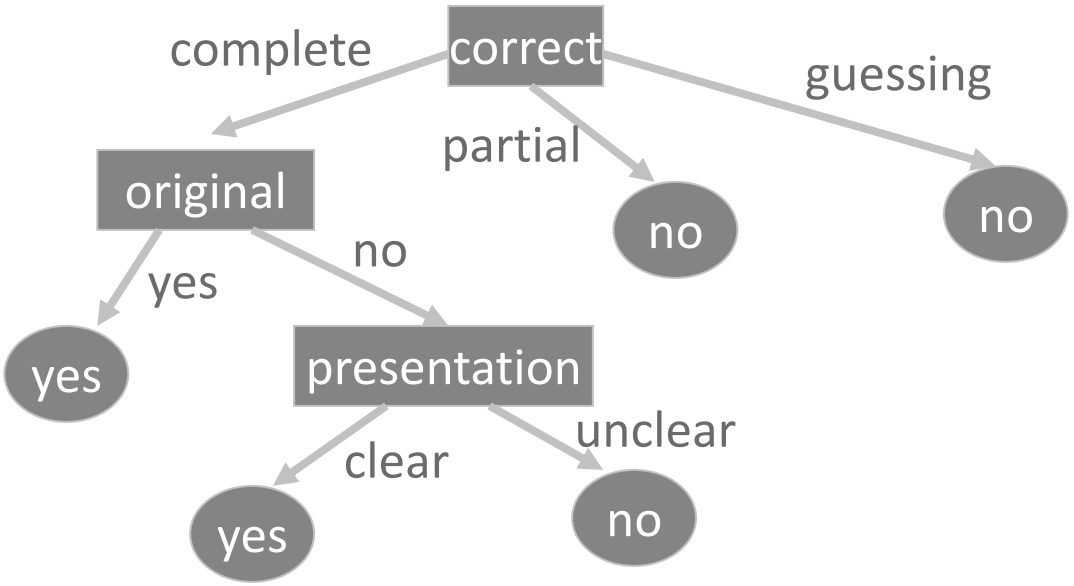
- Training Error: Error $Err_{S_{train}}(h)$ on training sample.
- Test Error: Error $Err_{S_{test}}(h)$ on test sample is an estimate of $Err_P(h)$.

Overfitting



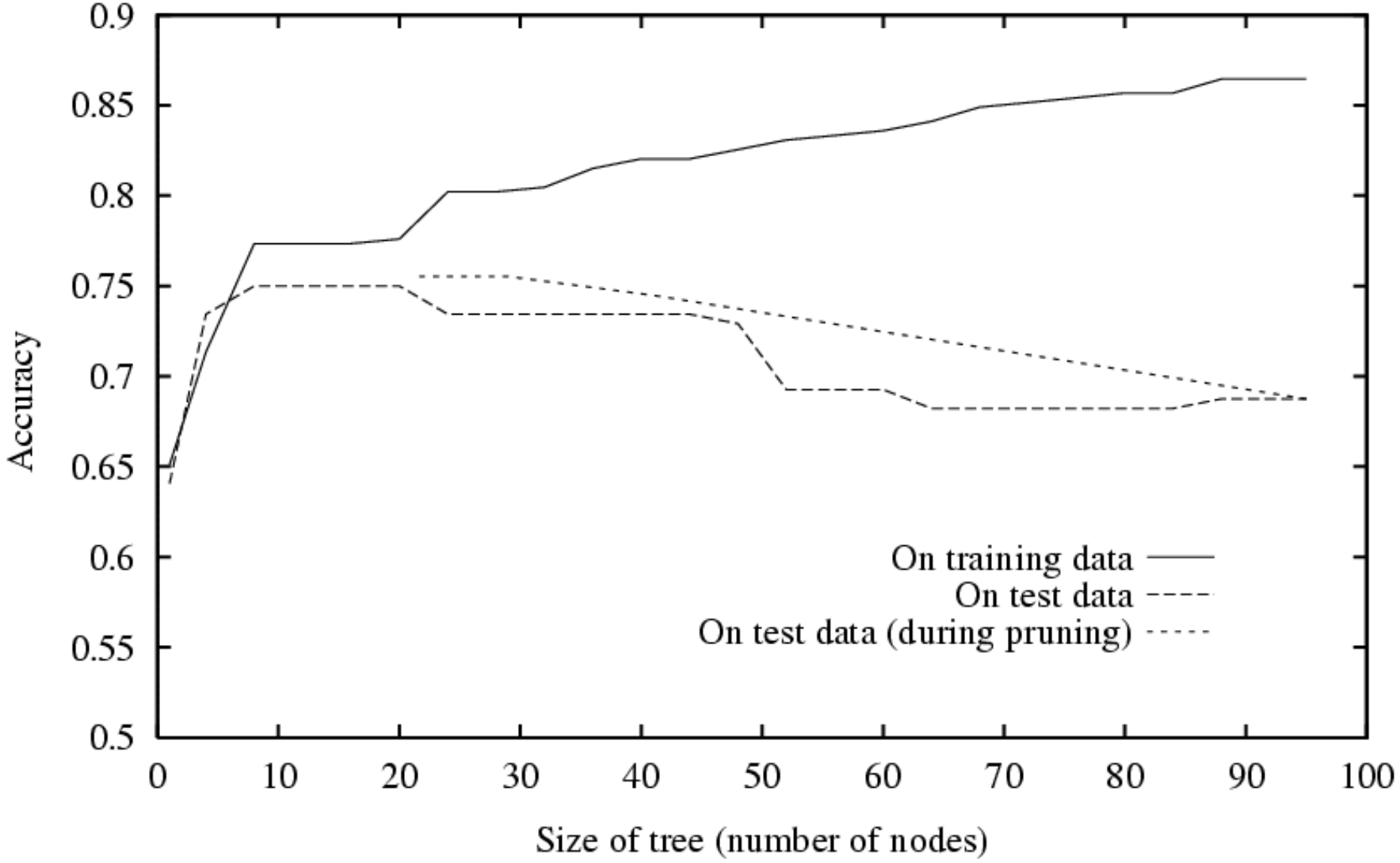
- Note: Accuracy = 1.0-Error

Decision Tree Example: revisited



	A^+
$\vec{x}_1 = (\mathrm{c}, \mathrm{y}, \mathrm{c})$	$y_1 = +1$
$\vec{x}_2 = (\mathrm{c}, \mathrm{n}, \mathrm{u})$	$y_2 = -1$
$\vec{x}_3 = (\mathrm{c}, \mathrm{y}, \mathrm{u})$	$y_3 = +1$
$\vec{x}_4 = (\mathrm{c}, \mathrm{n}, \mathrm{c})$	$y_4 = +1$
$\vec{x}_5 = (\mathrm{p}, \mathrm{y}, \mathrm{c})$	$y_5 = -1$
$\vec{x}_6 = (\mathrm{g}, \mathrm{y}, \mathrm{c})$	$y_6 = -1$
$\vec{x}_7 = (\mathrm{c}, \mathrm{y}, \mathrm{c})$	$y_7 = +1$
$\vec{x}_8 = (\mathrm{c}, \mathrm{y}, \mathrm{u})$	$y_8 = +1$
$\vec{x}_9 = (\mathrm{p}, \mathrm{y}, \mathrm{c})$	$y_9 = -1$
$\vec{x}_{10} = (\mathrm{c}, \mathrm{y}, \mathrm{c})$	$y_{10} = +1$

Reduced-Error Pruning



Text Classification Example Results

- Unpruned Tree:
  - Size: 437 nodes Training Error: 0.0% Test Error: 11.0%
- Early Stopping Tree:
  - Size: 299 nodes Training Error: 2.6% Test Error: 9.8%
- Post-Pruned Tree:
  - Size: 167 nodes Training Error: 4.0% Test Error: 10.8%
- Rule Post-Pruning:
  - Size: 164 tests Training Error: 3.1% Test Error: 10.3%
  - Examples of rules
    - IF vs = 1 THEN - [99.4%]
    - IF vs = 0 & export = 0 & takeover = 1 THEN + [93.6%]