Decision trees and their use in NLP

Jan Hajič

Additional Lecture to NPFL067

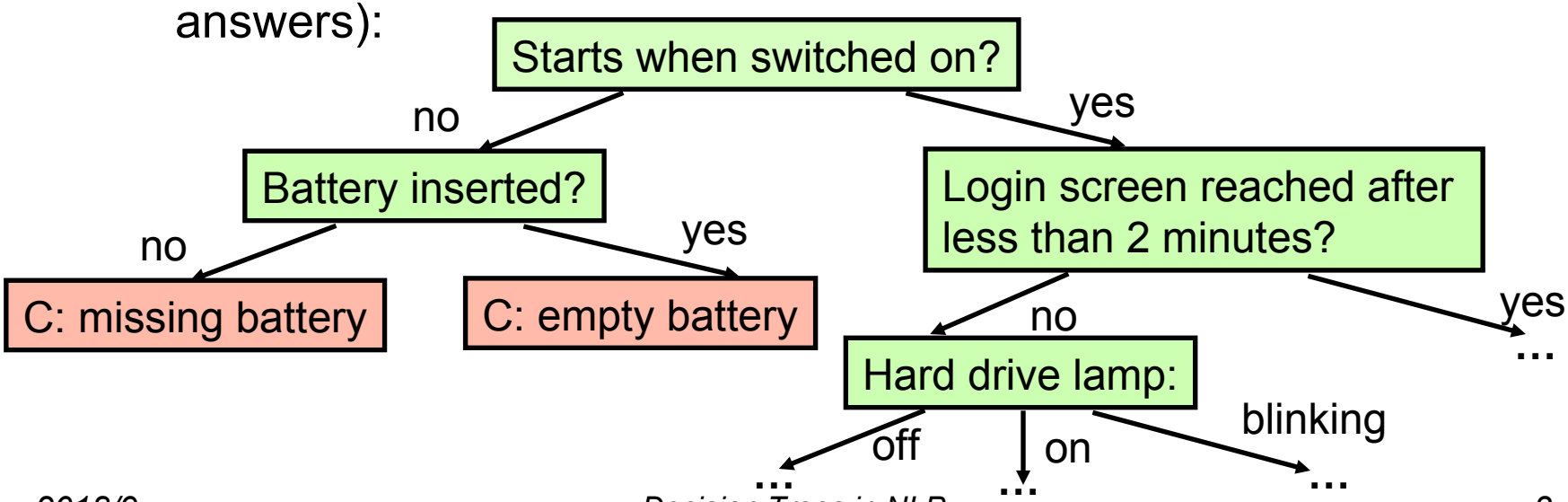
Fall 2018/19

Decision Trees

- Goal: Categorical or numerical predictions
  - i.e., classification (prevalent in NLP) / regression
- Use in NLP (examples)
  - Standard classification
    - POS/morphological tagging C_{POS}: $W \rightarrow T$
    - Named entity recognition NE: $W \rightarrow \{0,1\}^{|W|}$
  - Modeling conditional distributions
    - Language modeling LM: $<w_1,\ldots,w_i> \rightarrow P_{i+1}(.)$
    - In general D: $H \rightarrow P(.)$
      - H is “history” (context)
      - P(.) is a set probabilistic distributions on the variable of interest (e.g. “next word”)

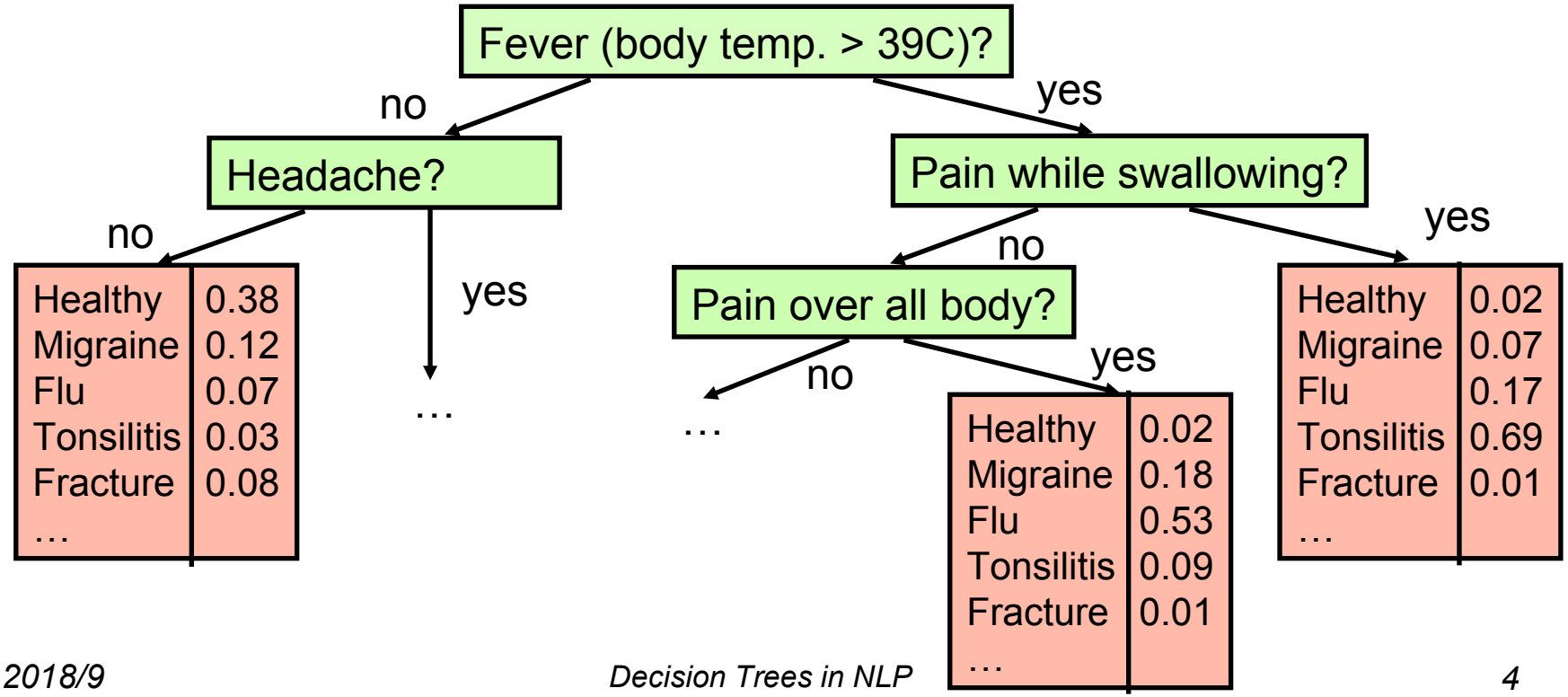
Decision Trees – the Idea

- Queries organized in a rooted tree
  - Queries evaluated against (input) data
    - Precisely, against context of the item of interest to be classified
  - Value returned → edge to be followed down the tree
  - (Global) answer (class, distribution) found at leaf
  - Example classifier (customer helpdesk, data: customer's answers):



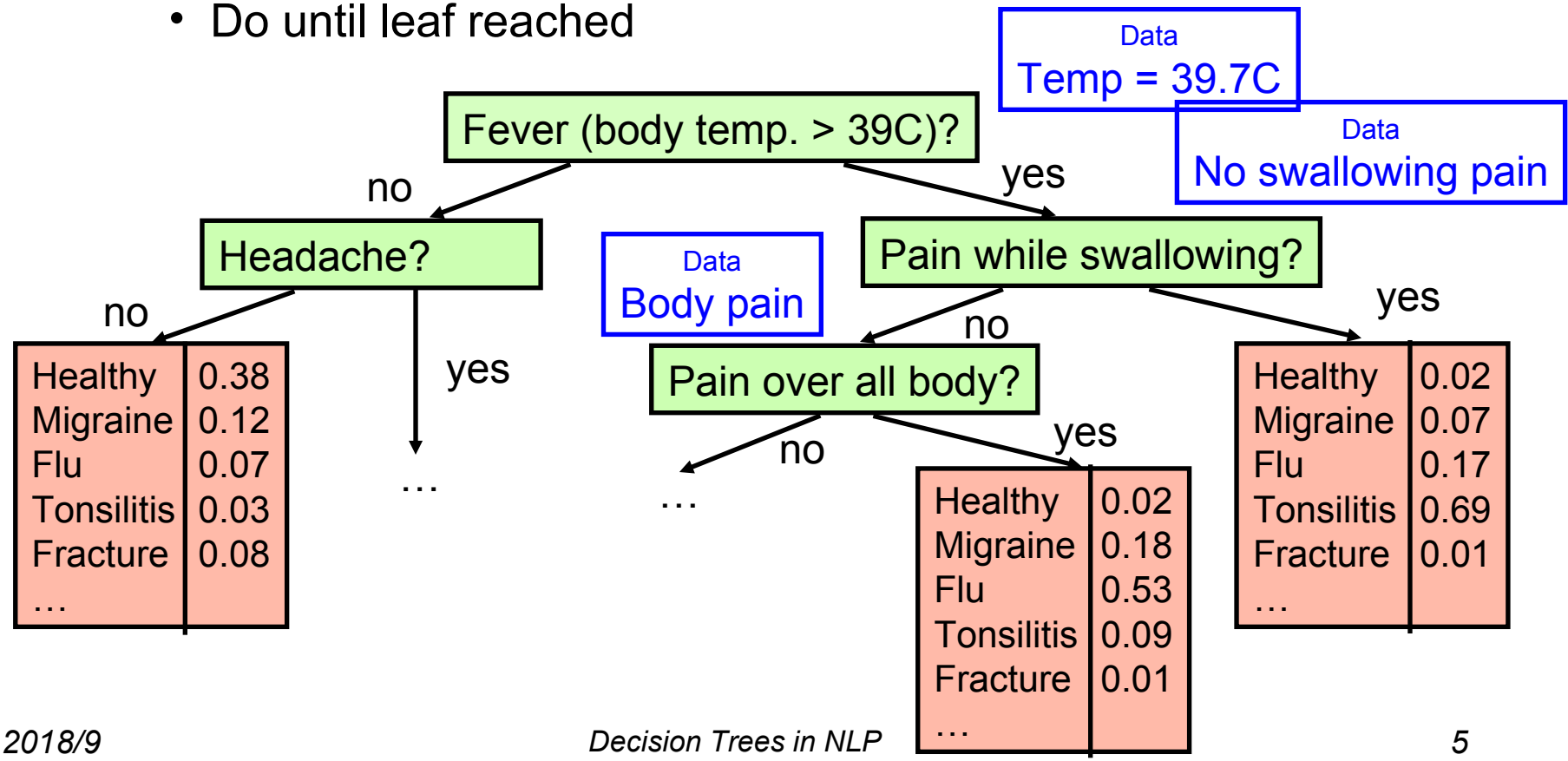
Goal: A Distribution

- Generalization (from a categorical answer)
  - Leaves contain a prob. distribution on 'classes'



Using Decision Trees

- Collect data at each query, evaluate
  - Follow edge, evaluate next query
  - Do until leaf reached



Constructing (Training) Decision Trees

- Set-of-Queries Acquisition
  - Manual (given)
  - (Semi-)automatic ([man.] templates → instances)
- Tree building
  - Machine learning (supervised)
  - Objective function
    - Probability of data (maximize) given the (trained) tree (~ model)
    - MERT (Minimum error rate training)
      - (If it is) hard to define probability of data (esp. categorical classifiers)
  - NP complete problem
    - Heuristics needed (e.g., greedy search)
    - Approximations
  - Technique: Top-down, Node-splitting

Queries

- Binary (Yes/No)
  - two edges outgoing from a query node
    - Is the previous word “to”?
    - Is there the word “car” within the same sentence?
    - Is the rel. unigram frequency of the prev. word > 0.05?
    - Is the entropy of the trigram distribution $P(.|w_{i-2},w_{i-1})$ below 0.1 (at the position *‘i’* in the data)?
- General (discrete)
  - N (>2) edges outgoing from a query node
    - Number of children (0, 1, 2, >2) → 4 edges down
    - POS of previous word (N,V,A,…,) → 10 (11, 12, …) edges
    - Discretization intervals: (0..0.05, ..0.10, ..1.00) → 20 edges

Queries: Acquisition

- Defined ahead of time
  - Diagnosis (problem, medical), Financials, Profiling, …
  - NLP: POS tagging, Word Sense Disambiguation
- Template-based
  - Analogy:
    - Brill's TBEDL
    - Feature templates in MaxEnt models, perceptrons, …
  - Used for most tasks in NLP
    - Language modeling, other models w/conditional distributions
  - Two steps:
    - Template definition (manual)
    - Query Instance Generation (template "expansion") – data-based
    - [Selection (cf. feature selection in MaxEnt): part of tree building]

Tree Building: Data

- (Large) vector of pairs (y,x): D = (y_i,x_i), i = 1..|D|
  - y_i – value of interest (the one being predicted)
  - x_i – context ('history')
- Examples, modeling p(y|x):
  - Language modeling
    - n-gram LM: y_i – current word, x_i – previous (n-1) words
  - POS tagging
    - y_i – POS tag, x_i – words in sentence & tags to the left
  - Word Sense Disambiguation
    - y_i – sense of the word at position 'i' (from a fixed set),
    - x_i – words +- 50 positions away from position 'i'
  - (Non-NLP:) Disease diagnostics
    - y_i – disease, x_i – vector of symptoms (numerical, categorical)

Tree Building: the Objective Function (Θ)

- Form
  - Function to maximize/minimize: Θ: T x D $\rightarrow \Re$
    - T – the decision tree being built, D – the training data
- Distribution-based
  - As usual: (Max.) Probability of data ~ (Min.) Entropy

$$\text{argmax}_T\ \Theta(T,D) = P_T(D) \sim \text{argmin}_T\ \Theta(T,D) = -\sum_{y,x} p_T(y,x)\ \log(p_T(y|x))$$

$$= -1/|D| \sum_{i=1..|D|} \log(p_T(y_i|x_i))$$

- Error-based (~Minimum Error Rate Training, MERT)

$$\text{argmin}_T\ \Theta(T,D) = ER = 1/|D| \sum_{i=1..|D|} \delta(\text{Classify}_T(x_i),y_i)$$

Classify_T: X $\rightarrow$ Y chooses y $\in$ Y given context x $\in$ X using T

Tree Building: the Algorithm (~ID3)

- Using training data and the objective function:

$$T_{final} = \text{argmin}_T \, \Theta(T,D)$$

  - Too (exponentially) many possible trees (vs. no. of queries)
- Greedy search (Q: pool of possible queries)
  1. Start with 'empty' T (single leaf node), set $\beta_0 = \Theta(T,D)$
  2. Iterate (iteration index $k = 1..k_{max}$); set $\beta^{min}_k = \beta_{k-1}$
     - For all leafs l_i in T
     - For all $q_j \in Q$:
       - Split l_i into a query node q_j, and n_j (≥ 2) leafs → call it $T_{i,j}$
       - If $\Theta(T_{i,j},D) < \beta^{min}_k$: set $\beta^{min}_k = \Theta(T_{i,j},D)$ and remember i,j

         i.e. β^{min}_k holds the minimal value of $\Theta(T_{i,j},D)$ so far
  3. Set $\beta_k = \beta^{min}_k$ and $T = T_{i,j}$; repeat (2) until termination

Tree Building: Termination

- Terminating condition(s):
  - For any new query splitting any leaf node of T:
    - No improvement based on objective function $\Theta(T,D)$
      - i.e., $\beta^{min}_k = \beta_{k-1}$ at the end of the iteration, for example:
        - » Entropy does not go down
        - » Error rate does not go down
    - Small improvement only ($\beta_{k-1} - \beta^{min}_k < \tau$)
      - To avoid overtraining
  - Tree too big (or too deep)
    - To avoid overtraining, long running time; space constraints etc.
      - Set k_{max} to desired maximum size of tree, or watch T's depth

Example: POS tagging

- Data:

Positions:	1	2	3	4	5	6	7	8
X:	John	can	bring	the	can	to	the	table
Y:	NN	MOD	VBF	DET	NN	PRE	DET	NN

- Objective function: error rate

$$\Theta(T,D) = 1/|D| \sum_{i=1..|D|} \delta(C_T(x_i),y_i)$$

C_T is the "Classify" function: Words (in/with context) → Tags

- Query pool:

$q_1..q_6$: current word (w_i = John, …, table)

$q_7..q_{11}$: previous tag (t_{i-1} = NN, …, PRE)

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=bring? q_3

w_i=the? q_4

w_i=to? q_5

w_i=table? q_6

t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

t_{i-1}=PRE? q_{11}

C: NN l_1

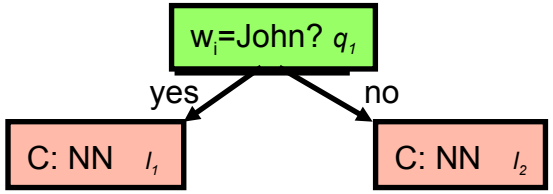
Iteration "0"

$\beta_0 = \Theta(T,D) = 5/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



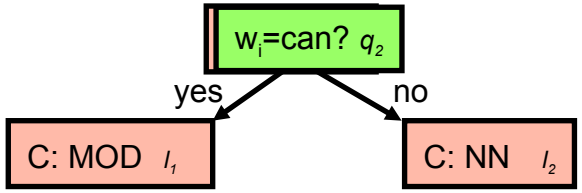
$T_{1,1}$

Iteration 1 | Leaf 1 | Query 1 | i,j: ---- | $\Theta(T_{1,1},D) = 5/8$ | $\beta^{min}_1 = 5/8$ | $\beta_0 = 5/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



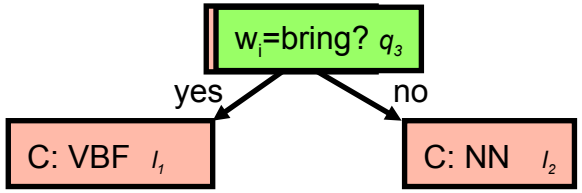
$T_{1,2}$

Iteration 1 | Leaf 1 | Query 2 | i,j: ---- | $\Theta(T_{1,2},D) = 5/8$ | $\beta^{min}_1 = 5/8$ | $\beta_0 = 5/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



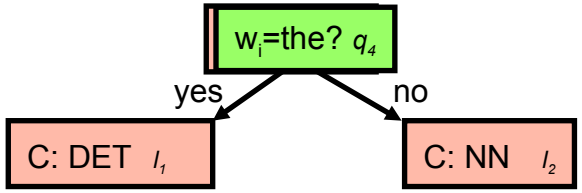
$T_{1,3}$

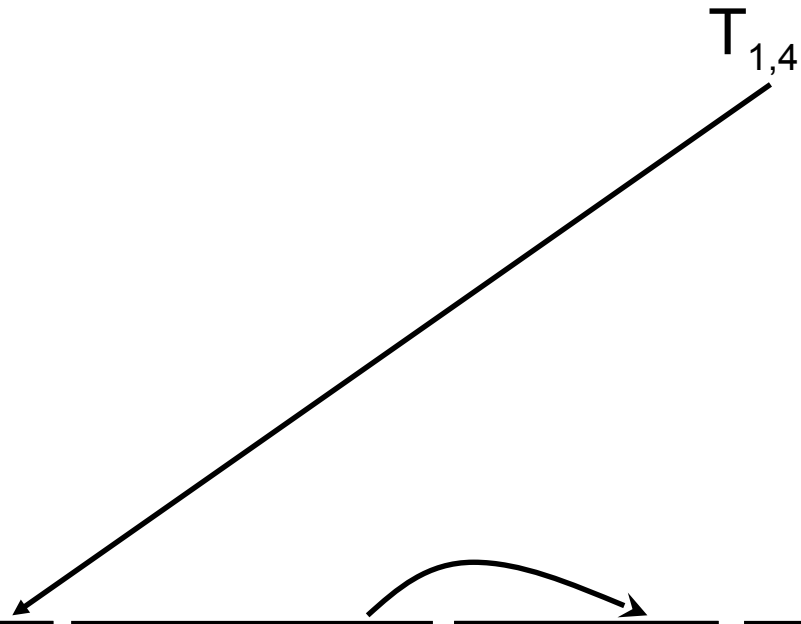
Iteration 1 | Leaf 1 | Query 3 | i,j: 1,3 | $\Theta(T_{1,3},D) = 1/2$ | $\beta^{min}_1 = 1/2$ | $\beta_0 = 5/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



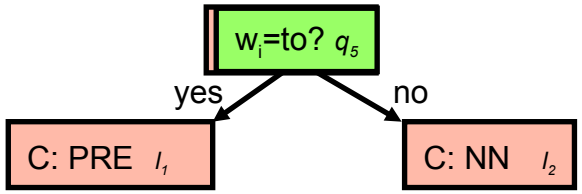


Iteration 1 | Leaf 1 | Query 4 | i,j: 1,4 | $\Theta(T_{1,4},D) = 3/8$ | $\beta^{min}_1 = 3/8$ | $\beta_0 = 5/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



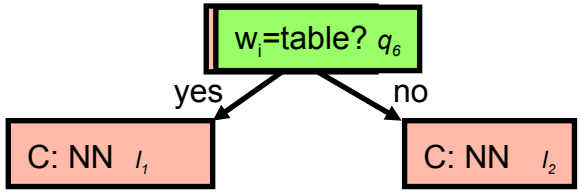
$T_{1,5}$

Iteration 1 | Leaf 1 | Query 5 | i,j: 1,4 | $\Theta(T_{1,5},D) = 1/2$ | $\beta^{min}_1 = 3/8$ | $\beta_0 = 5/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



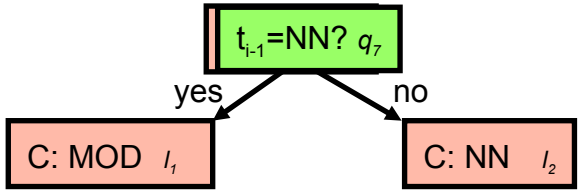
$T_{1,6}$

Iteration 1 | Leaf 1 | Query 6 | i,j: 1,4 | $\Theta(T_{1,6},D) = 5/8$ | $\beta^{min}_1 = 3/8$ | $\beta_0 = 5/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



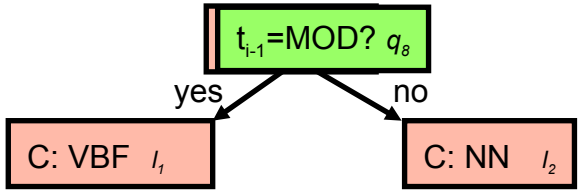
$T_{1,7}$

Iteration 1 | Leaf 1 | Query 7 | i,j: 1,4 | $\Theta(T_{1,7},D) = 1/2$ | $\beta^{min}_1 = 3/8$ | $\beta_0 = 5/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



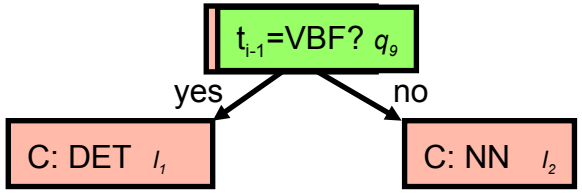
$T_{1,8}$

Iteration 1 | Leaf 1 | Query 8 | i,j: 1,4 | $\Theta(T_{1,8},D) = 1/2$ | $\beta^{min}_1 = 3/8$ | $\beta_0 = 5/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



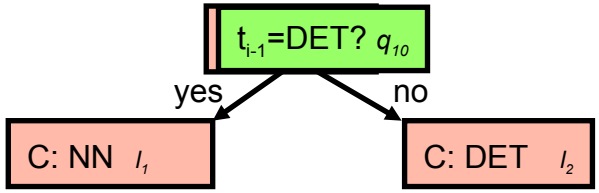
$T_{1,9}$

Iteration 1 | Leaf 1 | Query 9 | i,j: 1,4 | $\Theta(T_{1,9},D) = 1/2$ | $\beta^{min}_1 = 3/8$ | $\beta_0 = 5/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



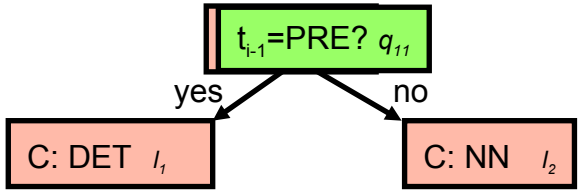
$T_{1,10}$

Iteration 1 | Leaf 1 | Query 10 | i,j: 1,4 | $\Theta(T_{1,10},D) = 1/2$ | $\beta^{min}_1 = 3/8$ | $\beta_0 = 5/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



$T_{1,11}$

End of iteration 1

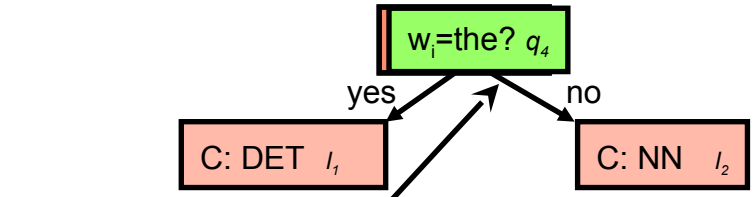
Iteration 1 | Leaf 1 | Query 11 | i,j: 1,4 | $\Theta(T_{1,11},D) = 1/2$ | $\beta^{min}_1 = 3/8$ | $\beta_0 = 5/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}

Iteration 2



Query q_4

Leaf l_1

Start of iteration 2

i,j: 1,4

Decreasing Θ… go on to next iteration

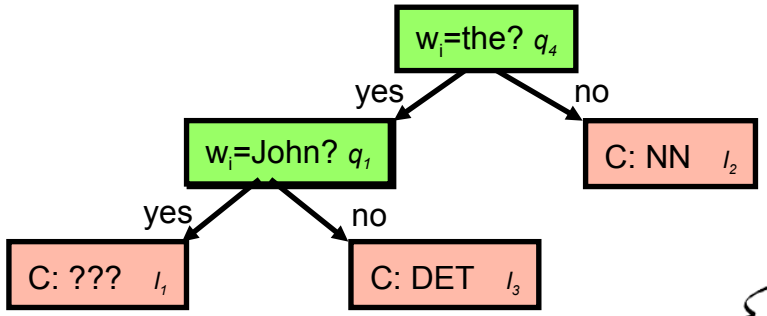
β^{min}_1 = 3/8

β_1 = 3/8

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



$T_{1,1}$

Can we ignore the queries with other words?

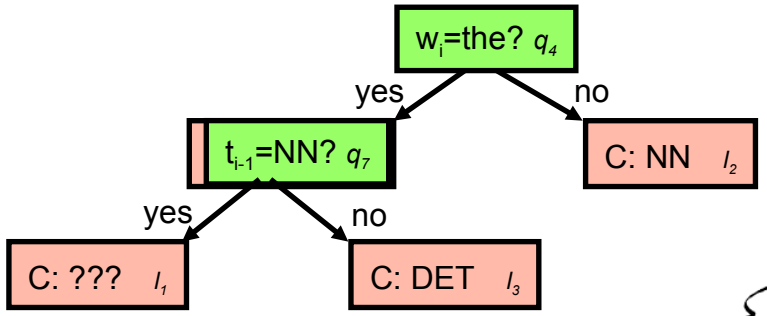
… yes (same w_i)

Iteration 2 | Leaf 1 | Query 1 | i,j: ---- | $\Theta(T_{1,1},D) = 3/8$ | $\beta^{min}_2 = 3/8$ | $\beta_1 = 3/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



$T_{1,7}$

Can we ignore the queries with tags at *i-1*?

… no (but see later for more powerful heuristics)

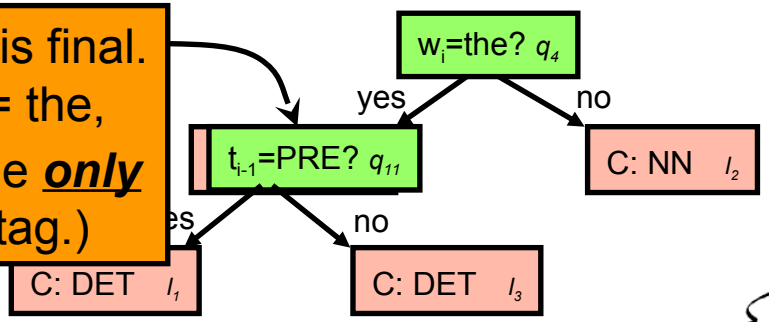
Iteration 2 | Leaf 1 | Query 7 | i,j: ---- | $\Theta(T_{1,7},D) = 3/8$ | $\beta^{min}_2 = 3/8$ | $\beta_1 = 3/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}

This leaf is final. (If word$_i$ = the, DET is the ***only*** possible tag.)



$T_{1,11}$

OK, so what's the heuristics?

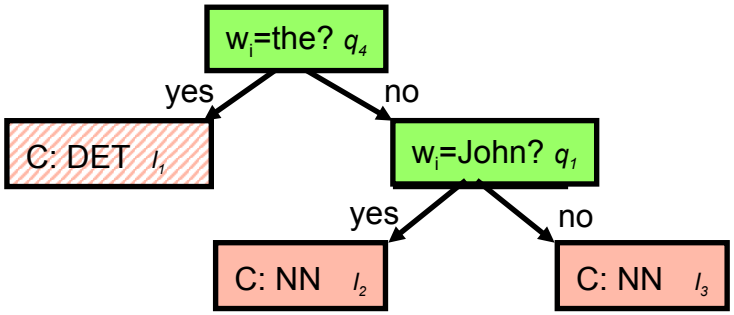
Leafs with <u>absolute classification certainty</u> need not be split

Iteration 2 | Leaf 1 | Query 11 | i,j: ---- | $\Theta(T_{1,11},D) = 3/8$ | $\beta^{min}_2 = 3/8$ | $\beta_1 = 3/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



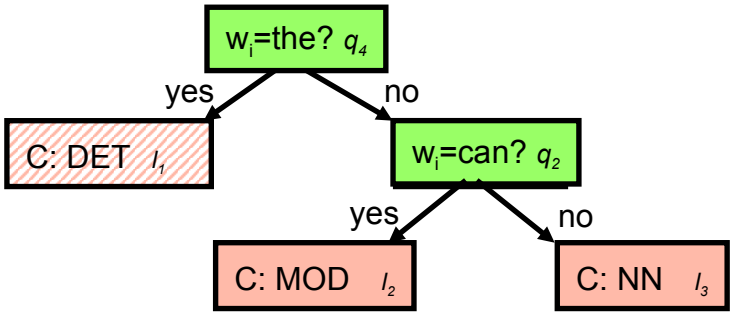
$T_{2,1}$

Iteration 2 | Leaf 2 | Query 1 | i,j: ---- | $\Theta(T_{2,1},D) = 3/8$ | $\beta^{min}_2 = 3/8$ | $\beta_1 = 3/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



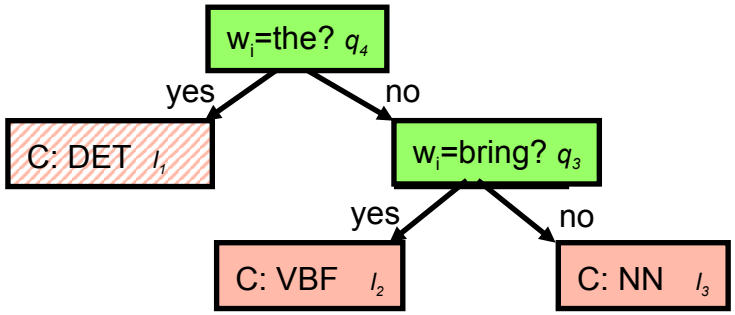
$T_{2,2}$

Iteration 2 | Leaf 2 | Query 2 | i,j: ---- | $\Theta(T_{2,2},D) = 3/8$ | $\beta^{min}_2 = 3/8$ | $\beta_1 = 3/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



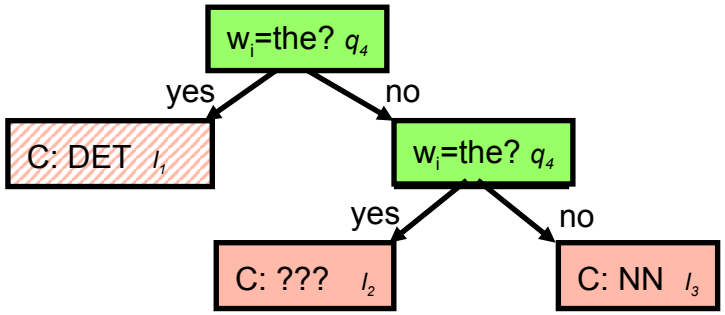
$T_{2,3}$

Iteration 2 | Leaf 2 | Query 3 | i,j: 2,3 | $\Theta(T_{2,3},D) = 1/4$ | $\beta^{min}_2 = 1/4$ | $\beta_1 = 3/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=the? q_4
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



$T_{2,4}$

Should I look at the same question again?

Queries once used may be excluded

…unlike in TBEDL!

Iteration 2 | Leaf 2 | Query 4 | i,j: 2,3 | $\Theta(T_{2,4},D) = 3/8$ | $\beta^{min}_2 = 1/4$ | $\beta_1 = 3/8$

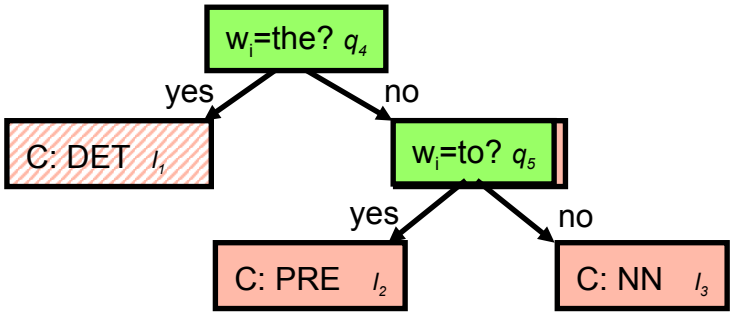
Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=bring? q_3



$T_{2,5}$

w_i=to? q_5

w_i=table? q_6

t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

t_{i-1}=PRE? q_{11}

Iteration 2 | Leaf 2 | Query 5 | i,j: 2,3 | $\Theta(T_{2,5},D) = 1/4$ | $\beta^{min}_2 = 1/4$ | $\beta_1 = 3/8$

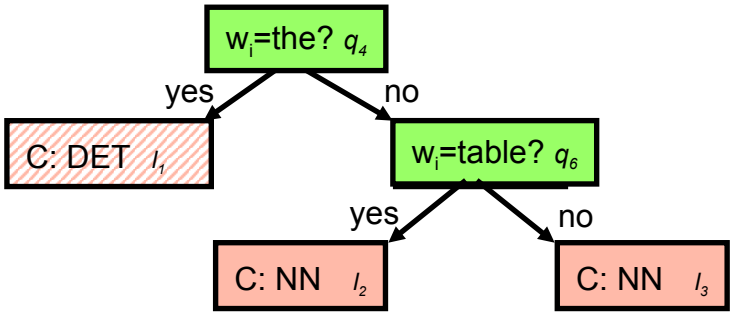
Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=bring? q_3



$T_{2,6}$

w_i=to? q_5

w_i=table? q_6

t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

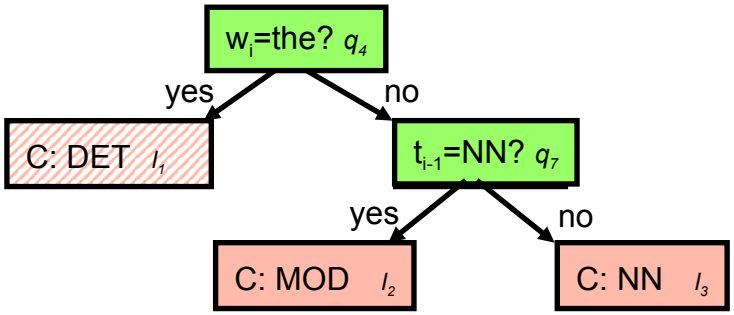
t_{i-1}=PRE? q_{11}

Iteration 2 | Leaf 2 | Query 6 | i,j: 2,3 | $\Theta(T_{2,6},D) = 3/8$ | $\beta^{min}_2 = 1/4$ | $\beta_1 = 3/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



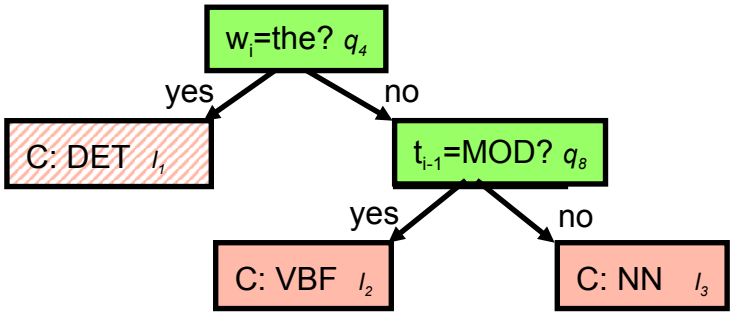
$T_{2,7}$

Iteration 2 | Leaf 2 | Query 7 | i,j: 2,3 | $\Theta(T_{2,7},D) = 1/4$ | $\beta^{min}_2 = 1/4$ | $\beta_1 = 3/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



$T_{2,8}$

Iteration 2 | Leaf 2 | Query 8 | i,j: 2,3 | $\Theta(T_{2,8},D) = 1/4$ | $\beta^{min}_2 = 1/4$ | $\beta_1 = 3/8$

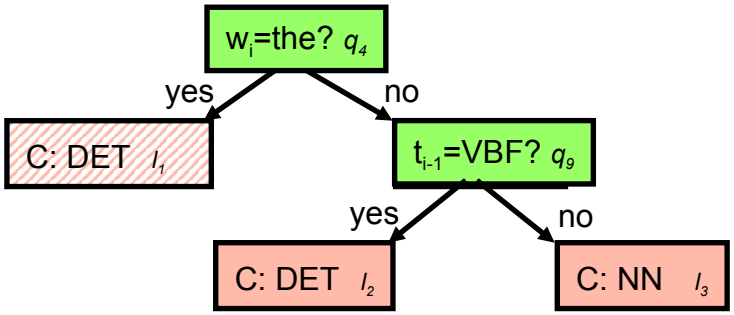
Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=bring? q_3



$T_{2,9}$

w_i=to? q_5

w_i=table? q_6

t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

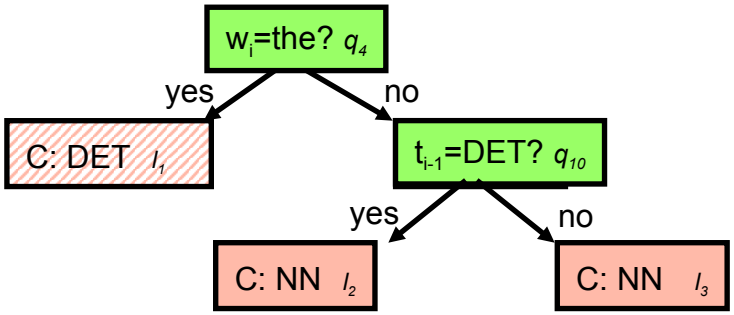
t_{i-1}=PRE? q_{11}

Iteration 2 | Leaf 2 | Query 9 | i,j: 2,3 | $\Theta(T_{2,9},D) = 3/8$ | $\beta^{min}_2 = 1/4$ | $\beta_1 = 3/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



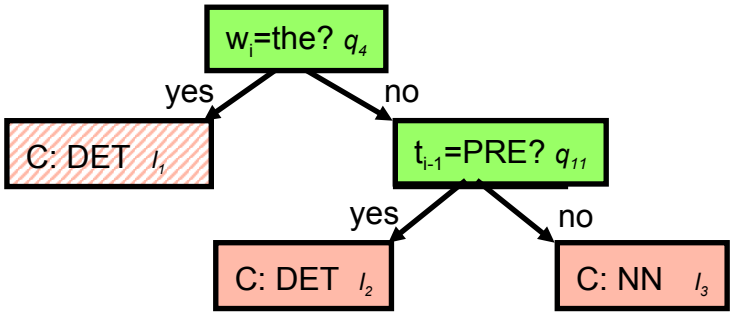
$T_{2,10}$

Iteration 2 | Leaf 2 | Query 10 | i,j: 2,3 | $\Theta(T_{2,10},D) = 3/8$ | $\beta^{min}_2 = 1/4$ | $\beta_1 = 3/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

- w_i=John? q_1
- w_i=can? q_2
- w_i=bring? q_3
- w_i=to? q_5
- w_i=table? q_6
- t_{i-1}=NN? q_7
- t_{i-1}=MOD? q_8
- t_{i-1}=VBF? q_9
- t_{i-1}=DET? q_{10}
- t_{i-1}=PRE? q_{11}



$T_{2,11}$

Iteration 2 | Leaf 2 | Query 11 | i,j: 2,3 | $\Theta(T_{2,11},D) = 3/8$ | $\beta^{min}_2 = 1/4$ | $\beta_1 = 3/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=bring? q_3

w_i=to? q_5

w_i=table? q_6

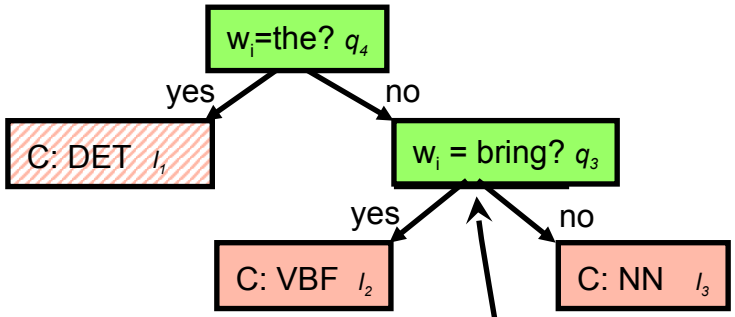
t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

t_{i-1}=PRE? q_{11}



Query q_3

End of iteration 2

Leaf l_2

Iteration 2 | Leaf 2 | Query 11 | i,j: 2,3

β^{min}_2 = 1/4 | β_2 = 1/4

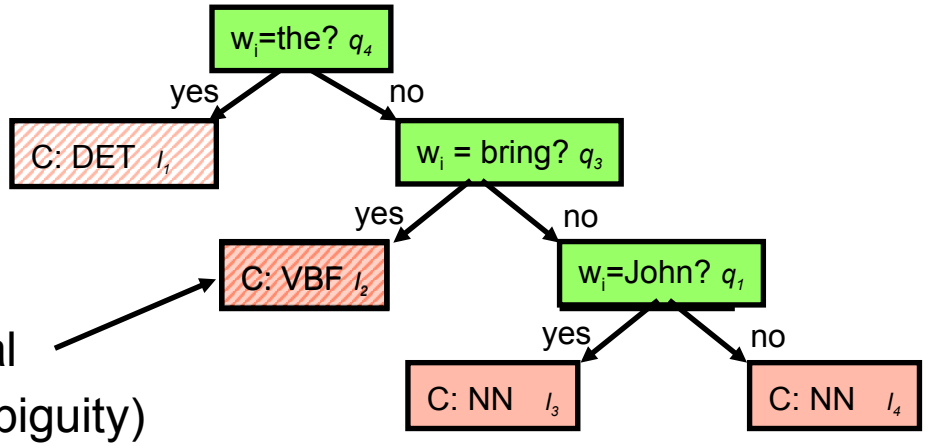
Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1
w_i=can? q_2
w_i=bring? q_3

Delete query used

w_i=to? q_5
w_i=table? q_6
t_{i-1}=NN? q_7
t_{i-1}=MOD? q_8
t_{i-1}=VBF? q_9
t_{i-1}=DET? q_{10}
t_{i-1}=PRE? q_{11}



$T_{3,1}$

l_2 is final (no ambiguity)

Start of iteration 3

Iteration 3	Leaf 3	Query 1	i,j: ----	$\Theta(T_{3,1},D) = 1/4$	$\beta^{min}_3 = 1/4$	$\beta_2 = 1/4$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=to? q_5

w_i=table? q_6

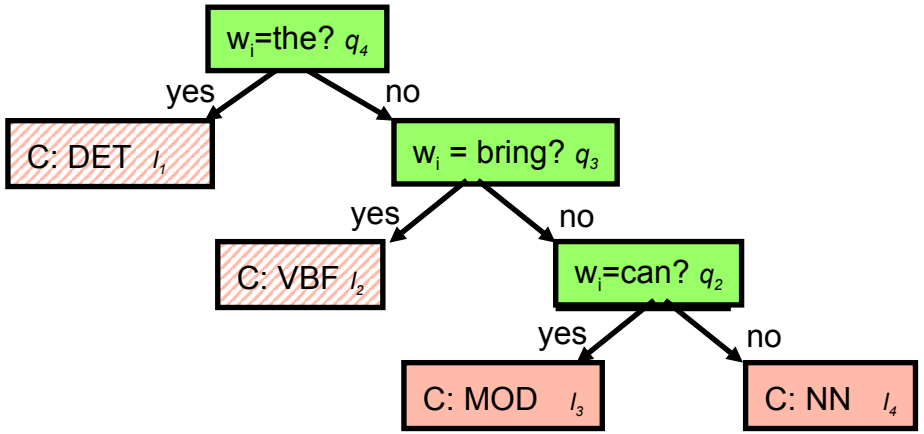
t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

t_{i-1}=PRE? q_{11}



$T_{3,2}$

Iteration 3 | Leaf 3 | Query 2 | i,j: ---- | $\Theta(T_{3,2},D) = 1/4$ | $\beta^{min}_3 = 1/4$ | $\beta_2 = 1/4$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=to? q_5

w_i=table? q_6

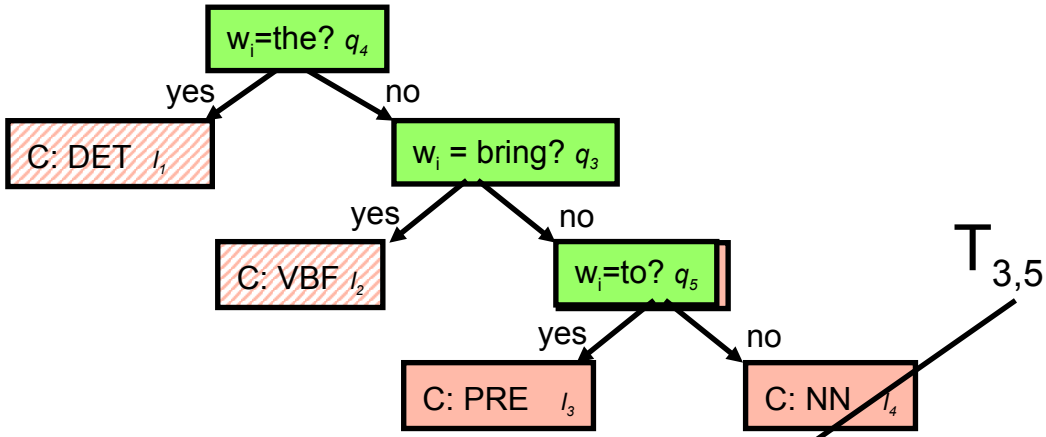
t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

t_{i-1}=PRE? q_{11}



$T_{3,5}$

Iteration 3 | Leaf 3 | Query 5 | i,j: 3,5 | $\Theta(T_{3,5},D) = 1/8$ | $\beta^{min}_3 = 1/8$ | $\beta_2 = 1/4$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=to? q_5

w_i=table? q_6

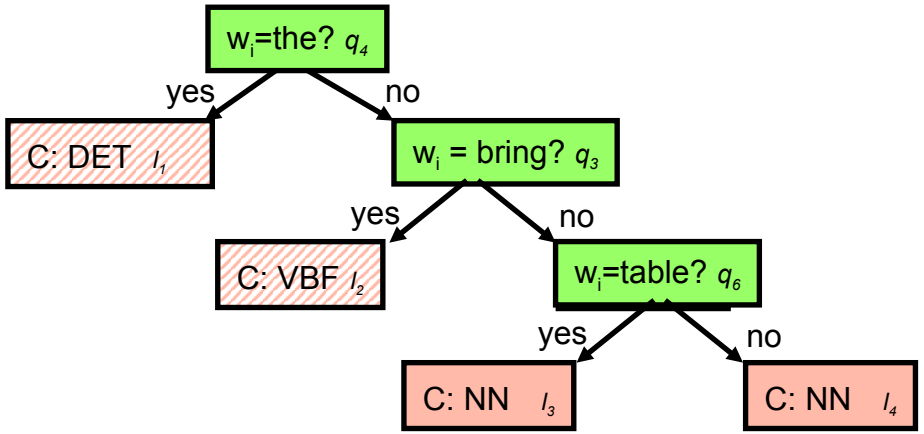
t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

t_{i-1}=PRE? q_{11}



$T_{3,6}$

Iteration 3 | Leaf 3 | Query 6 | i,j: 3,5 | $\Theta(T_{3,6},D) = 1/4$ | $\beta^{min}_3 = 1/8$ | $\beta_2 = 1/4$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=to? q_5

w_i=table? q_6

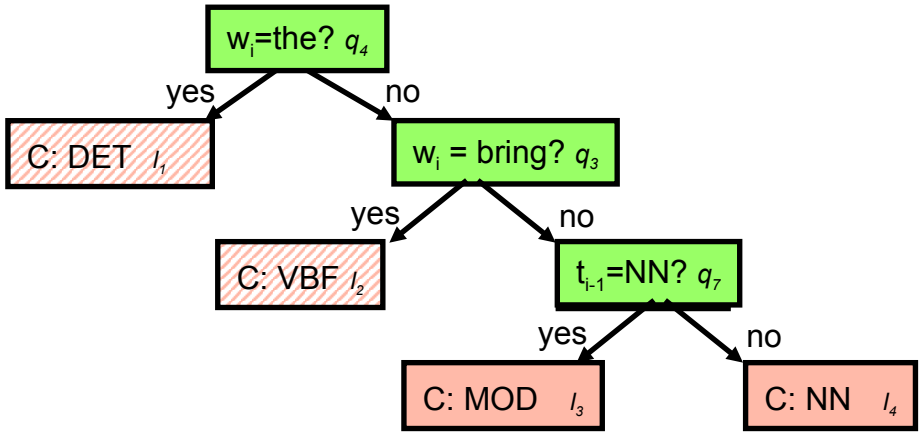
t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

t_{i-1}=PRE? q_{11}



$T_{3,7}$

Iteration 3 | Leaf 3 | Query 7 | i,j: 3,5 | $\Theta(T_{3,7},D) = 1/8$ | $\beta^{min}_3 = 1/8$ | $\beta_2 = 1/4$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=to? q_5

w_i=table? q_6

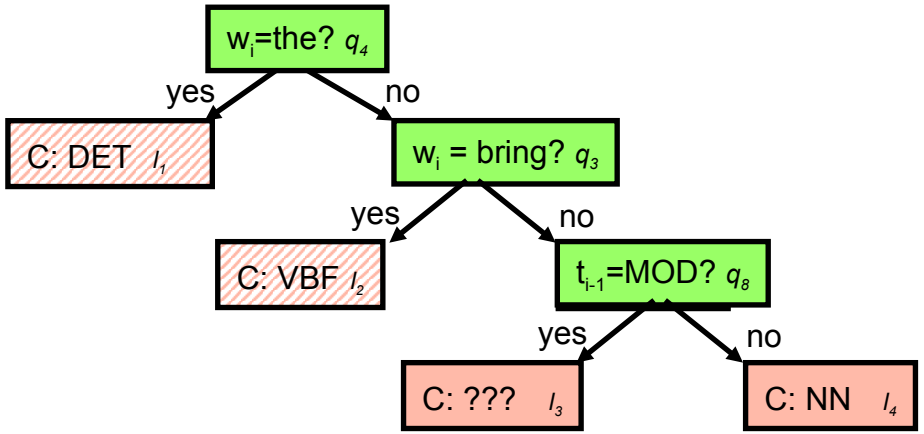
t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

t_{i-1}=PRE? q_{11}



$T_{3,8}$

Iteration 3 | Leaf 3 | Query 8 | i,j: 3,5 | $\Theta(T_{3,8},D) = 1/4$ | $\beta^{min}_3 = 1/8$ | $\beta_2 = 1/4$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=to? q_5

w_i=table? q_6

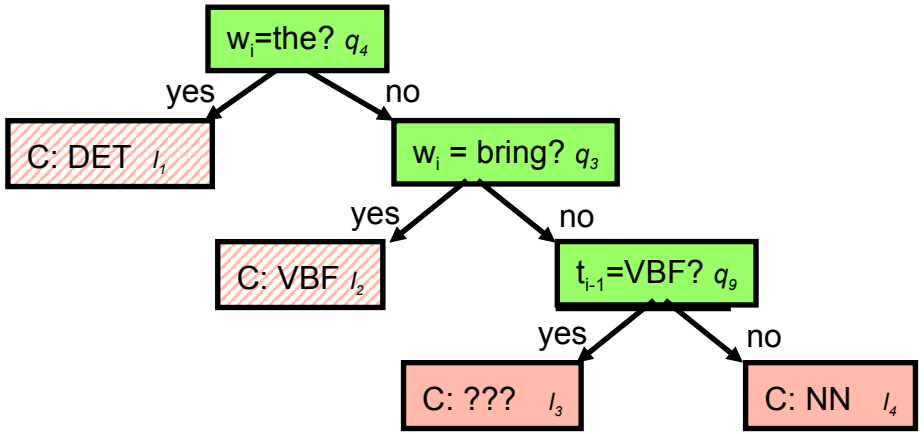
t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

t_{i-1}=PRE? q_{11}



$T_{3,9}$

Iteration 3 | Leaf 3 | Query 9 | i,j: 3,5 | $\Theta(T_{3,9},D) = 1/4$ | $\beta^{min}_3 = 1/8$ | $\beta_2 = 1/4$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=to? q_5

w_i=table? q_6

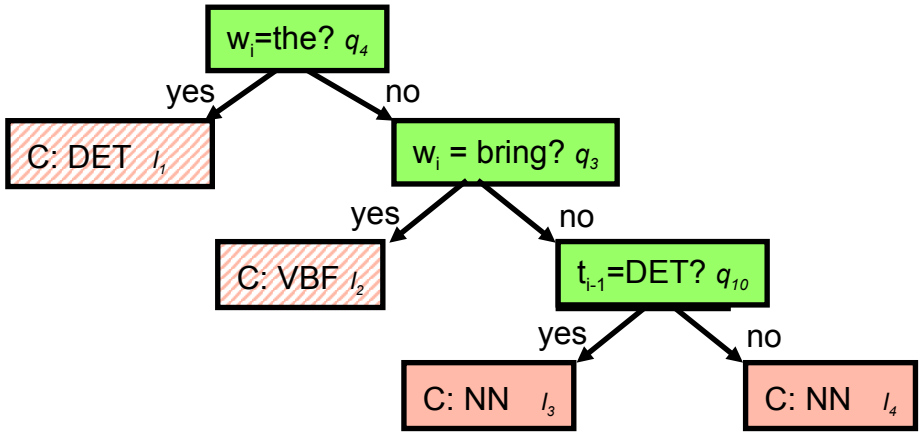
t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

t_{i-1}=PRE? q_{11}



$T_{3,10}$

Iteration 3 | Leaf 3 | Query 10 | i,j: 3,5 | $\Theta(T_{3,10},D) = 1/4$ | $\beta^{min}_3 = 1/8$ | $\beta_2 = 1/4$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=to? q_5

w_i=table? q_6

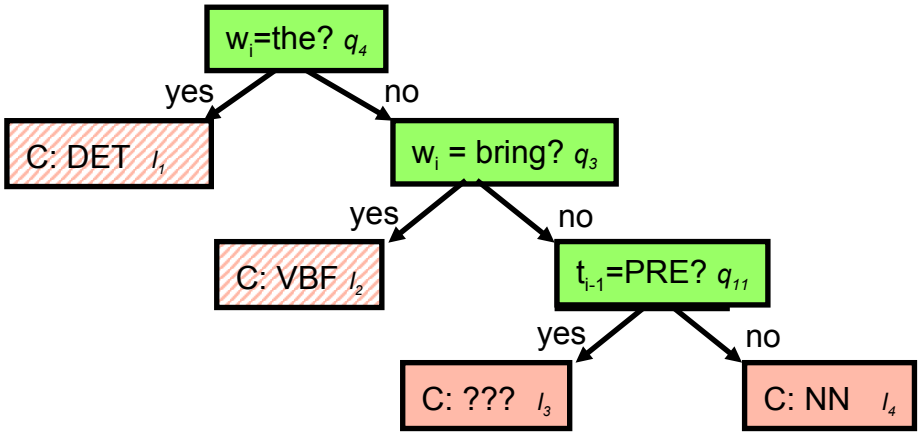
t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

t_{i-1}=PRE? q_{11}



$T_{3,11}$

Iteration 3 | Leaf 3 | Query 11 | i,j: 3,5 | $\Theta(T_{3,11},D) = 1/4$ | $\beta^{min}_3 = 1/8$ | $\beta_2 = 1/4$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=to? q_5

w_i=table? q_6

t_{i-1}=NN? q_7

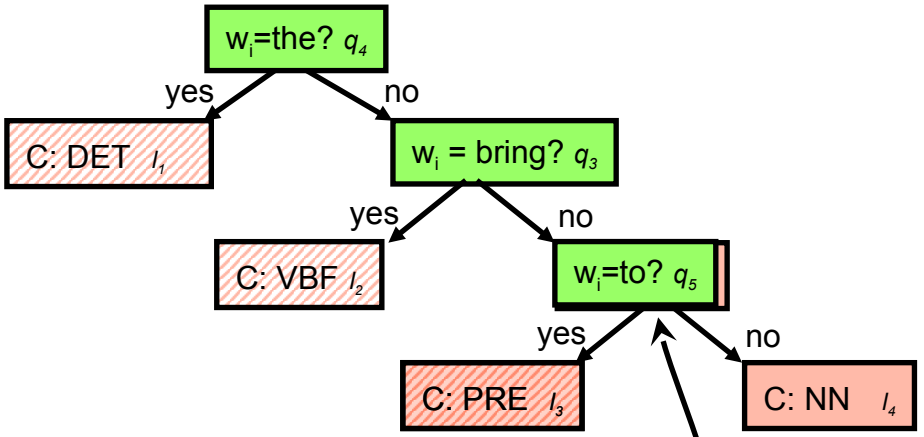
t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

t_{i-1}=PRE? q_{11}

Iteration 3



$T_{3,5}$

Start of iteration 4

Query q_5

Leaf l_3

i,j: 3,5

β^{min}_3 = 1/8

β_3 = 1/8

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=table? q_6

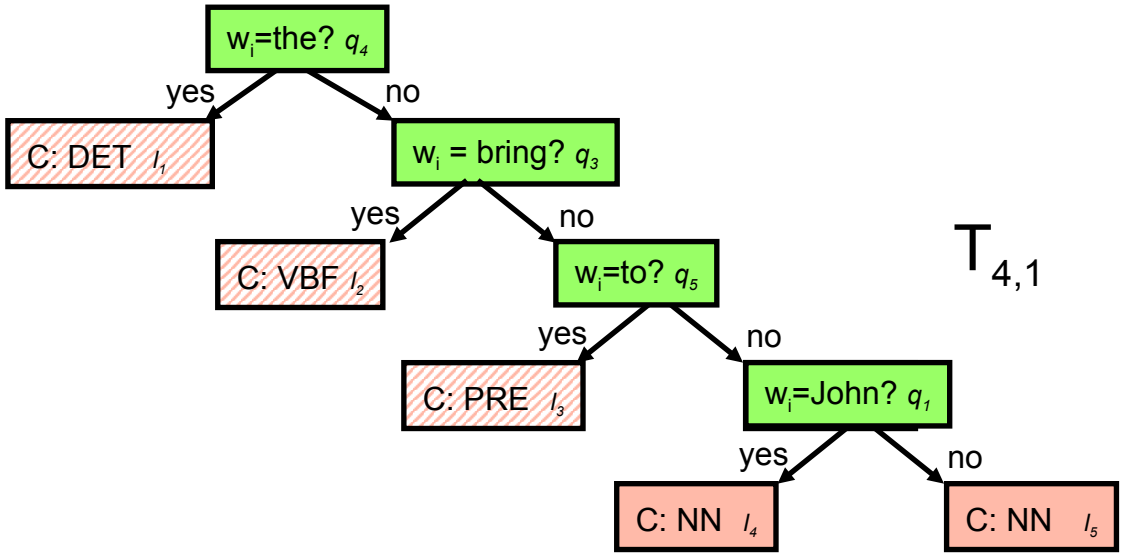
t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

t_{i-1}=PRE? q_{11}



$T_{4,1}$

Iteration 4 | Leaf 4 | Query 1 | i,j: ---- | $\Theta(T_{4,1},D) = 1/8$ | $\beta^{min}_4 = 1/8$ | $\beta_3 = 1/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=table? q_6

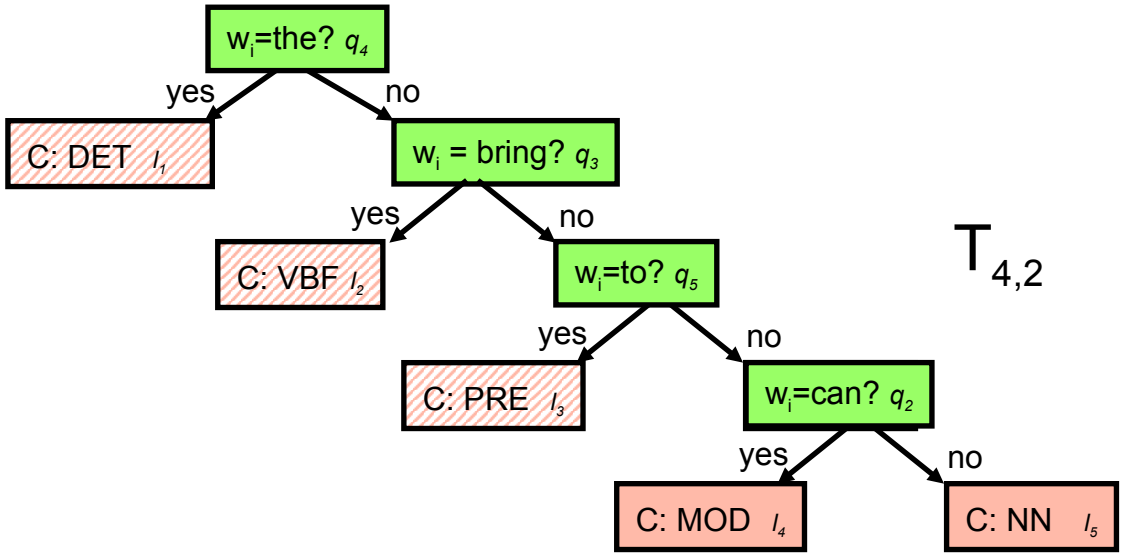
t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

t_{i-1}=PRE? q_{11}



$T_{4,2}$

Iteration 4 | Leaf 4 | Query 2 | i,j: ---- | $\Theta(T_{4,2},D) = 1/8$ | $\beta^{min}_4 = 1/8$ | $\beta_3 = 1/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=table? q_6

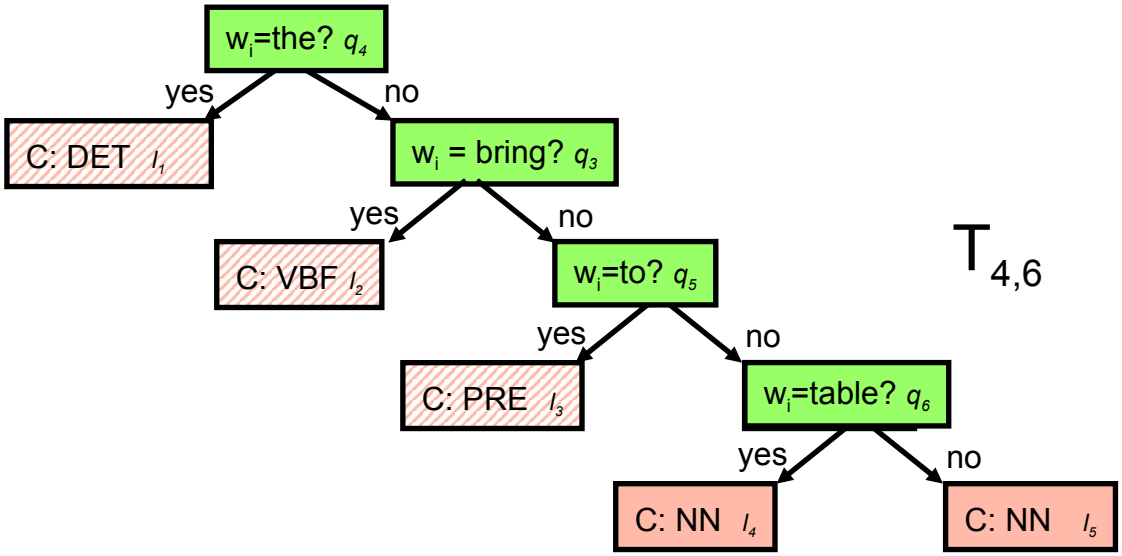
t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

t_{i-1}=PRE? q_{11}

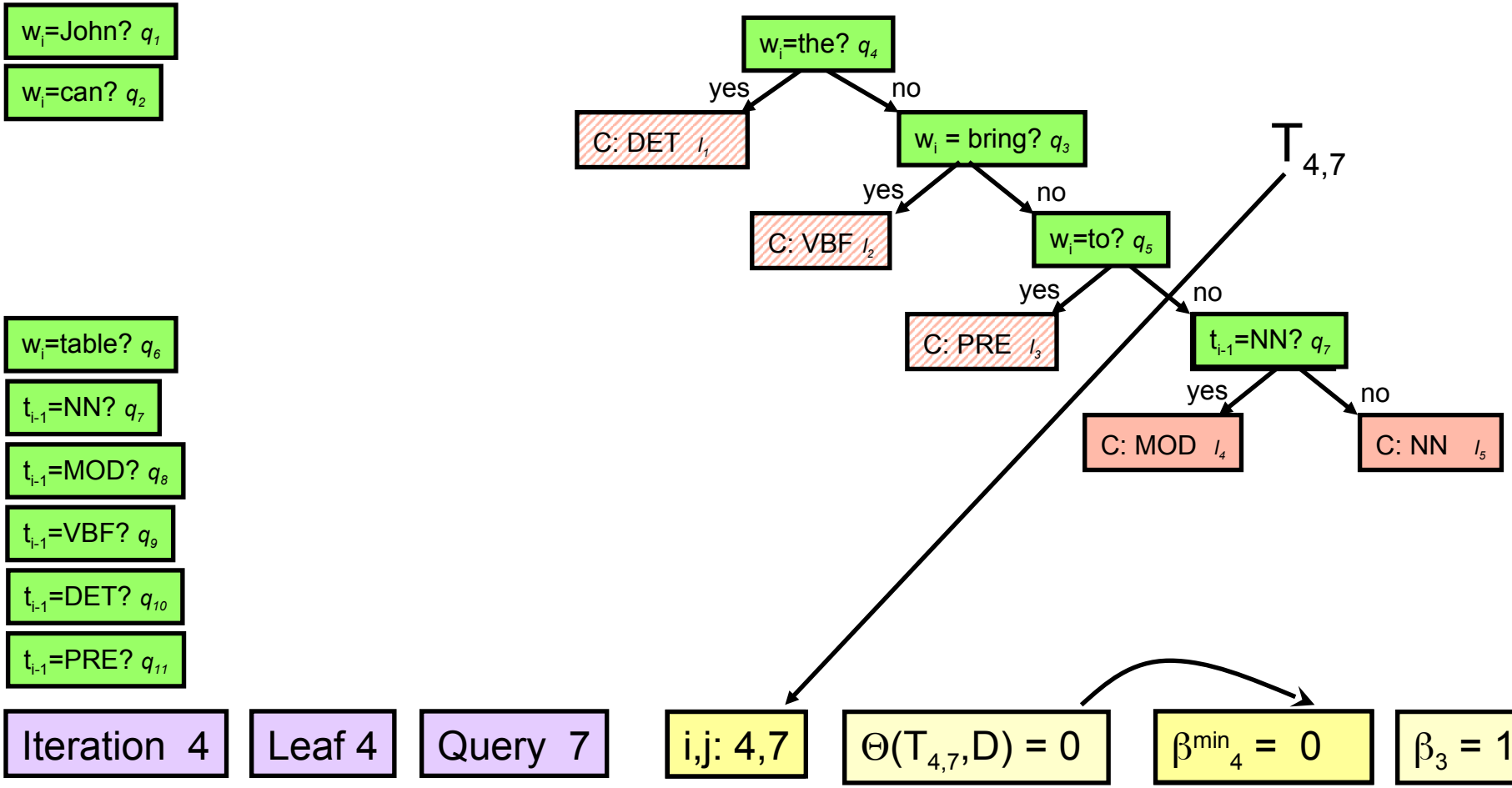


$T_{4,6}$

Iteration 4 | Leaf 4 | Query 6 | i,j: ---- | $\Theta(T_{4,6},D) = 1/8$ | $\beta^{min}_4 = 1/8$ | $\beta_3 = 1/8$

Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN



Example: POS tagging

1	2	3	4	5	6	7	8
X: John	can	bring	the	can	to	the	table
Y: NN	MOD	VBF	DET	NN	PRE	DET	NN

w_i=John? q_1

w_i=can? q_2

w_i=table? q_6

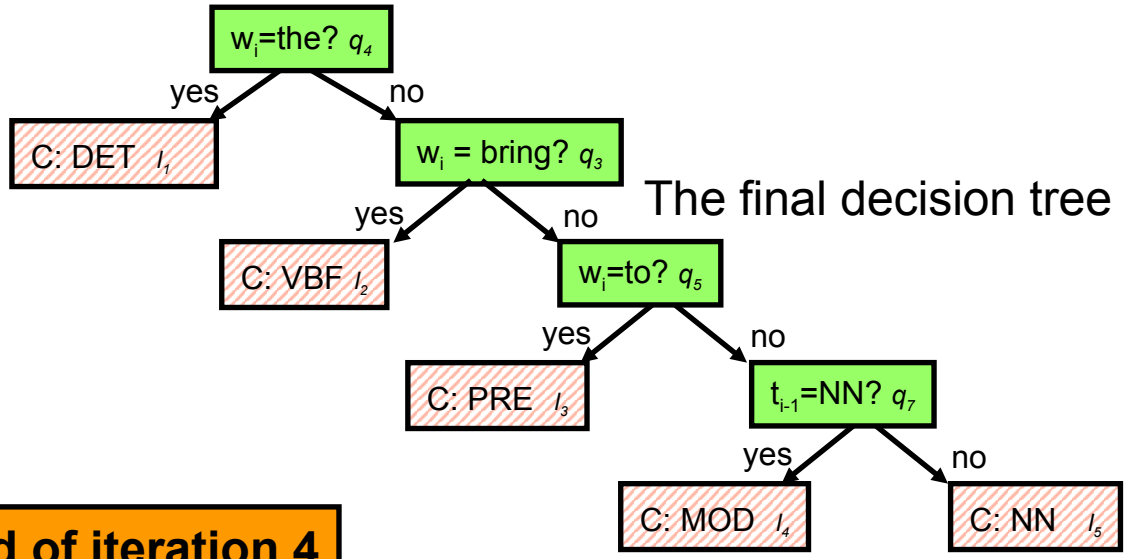
t_{i-1}=NN? q_7

t_{i-1}=MOD? q_8

t_{i-1}=VBF? q_9

t_{i-1}=DET? q_{10}

t_{i-1}=PRE? q_{11}



The final decision tree

End of iteration 4

End of training

Iteration 4

$\beta_4 = 0$

Generalized Case

- Same algorithm
- Objective function – entropy:

$$\text{argmin } \Theta(T,D) = -\sum_{y,x} p_T(y,x) \log(p_T(y|x))$$

$$= -1/|D| \sum_{i=1..|D|} \log(p_T(y_i|x_i))$$

- Effective computation
  - Compute only change of entropy at split node
    - Savings (much) greater towards end of computation
  - "Final" leaf heuristics:
    - Zero entropy at that leaf's distribution

Avoiding Overtraining (Overfitting)

- Stop growing the tree early
  - Set threshold for entropy gain at $\tau > 0$
- Smoothing (for the generalized case)
  - Keep distribution at every node (not just leaves) + weight (the "lambda")
  - Smooth along the path taken from root to leaf
  - Train weights by EM on heldout data
- Tree prunning
  - Use heldout data H for $\Theta(T,H)$ computation
  - Remove node if $\Theta(T,H)$ decreases (or stays the same ← minimal complexity principle)

Generalizations / Variants

- Historically: C4.5, C5.0
  - Conversion to rules at the end, prunning the rules at any "node" (not just leaves)
  - Ability to use incomplete data for training (unknown values of some attributes)
  - Using continuous-valued attributes
    - Key: finding (effectively) the threshold t for converting to query "is $value_A < t$?"
  - Effective runtime software for many programming languages
- Decision lists
  - "Narrow" graphs, allow multiple parents, special features
    - in certain cases, like WSD, avoid too fine data partitioning

Further Reading

Mitchell, T. M. (1997): *Machine Learning,* WCB/McGraw-Hill, ISBN 0-07-042807-7, Chapter 3

Quinlan, J. R. (1986): Induction of Decision Trees. *Machine Learning* 1(1), 81-106.

Quinlan, J. R. (1993): *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufman.

- http://rulequest.com (Quinlan's company)
- C5.0/2.09 (latest version)

Rokach, L., Maimon, O. (2005): Top-down induction of decision trees classifiers - a survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 35, no. 4, pp. 476-487, Nov. 2005

Google search ☺