University of BRISTOL

4.4 Conditional Random Fields

Edwin Simpson

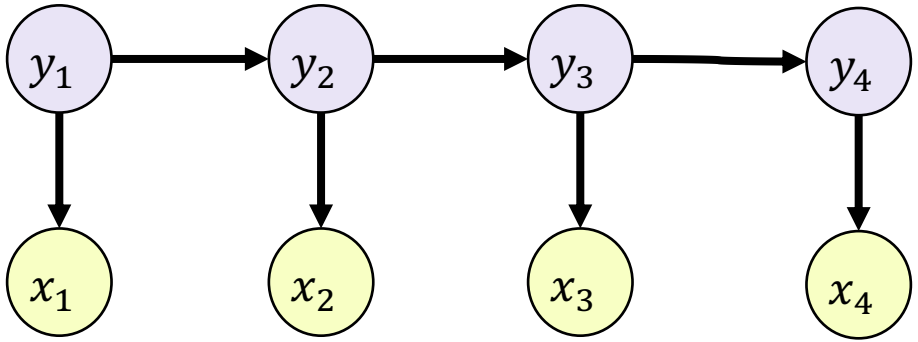
Department of Computer Science,

University of Bristol, UK.

bristol.ac.uk

HMM as a Generative Model

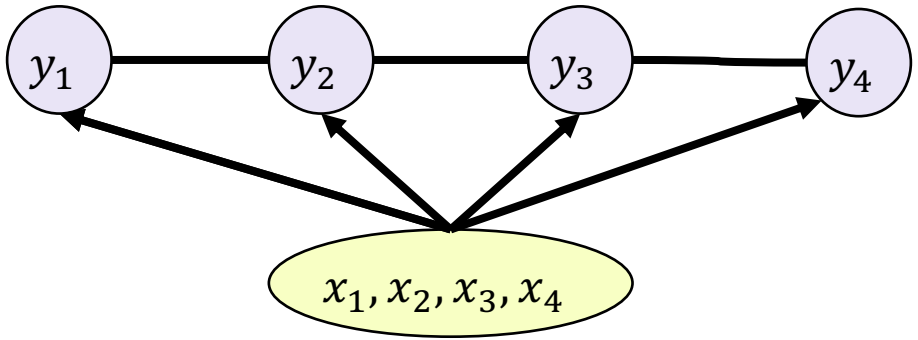
- HMM is a **generative** model:
  - Learns the likelihoods of observations then apply Bayes' rule to predict tags
  - Benefits: closed form maximum likelihood estimates, interpretable, modular
  - Related generative approach for classification: Naïve Bayes



Section 8.5, Speech and Language Processing (3rd edition draft), Jurafsky & Martin (2019).

Discriminative Models: CRF

- Conditional Random Field (CRF) is **discriminative**:
  - Optimises predictive distribution $p(\boldsymbol{y}|\boldsymbol{x})$
  - Related discriminative approach for classification: Logistic regression



Section 8.5, Speech and Language Processing (3rd edition draft), Jurafsky & Martin (2019).

CRF Prediction Function

Directly computes probability of the sequence

Weights like in logistic regression

$$P(\boldsymbol{y}|\boldsymbol{x}) \propto \exp\left(\sum_{k=1}^{K} \theta_k F_k(\boldsymbol{x}, \boldsymbol{y})\right)$$

Global feature function to compute the feature from the sequence $\boldsymbol{x}$

Section 8.5, Speech and Language Processing (3rd edition draft), Jurafsky & Martin (2019).

Global Feature Function

Function for feature *k*

Compute a local function for each token in the sequence

$$F_k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{N} f_k(y_{i-1}, y_i, \boldsymbol{x}, i)$$

Local function can use previous tag, current tag, whole token sequence, and current position.

Local Feature Functions

- We can use any function that can extract a feature from $y_{i-1}, y_i, \boldsymbol{x}, i$
- Examples for POS tagging: $f_k(y_{i-1}, y_i, \boldsymbol{x}, i) =$

$$[y_i = DET \text{ and } x_i = \text{“}the\text{”}]$$

$$[y_i = VB \text{ and } y_{i-1} = MD]$$

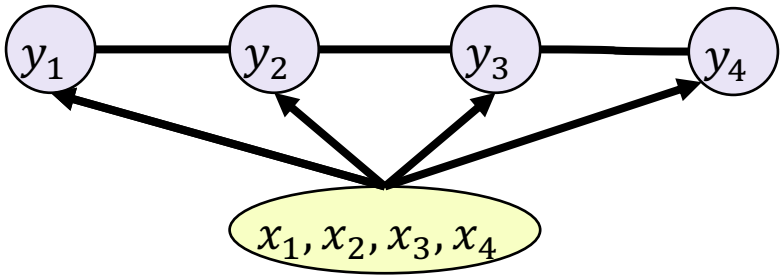
$$[x_{i-1} = \text{“}will\text{” and } x_{i+2} = \text{“}bill\text{”}]$$

$$[suffix(x_i) = \text{“}ed\text{”}]$$

‘[…]’ notation means the value is 1 if $y_i = DET \text{ and } x_i = \text{“}the\text{”}$ and 0 otherwise

Training

- Stochastic gradient descent
- Forward-backward algorithm needed to compute gradients
- Training is expensive with computational complexity $\mathcal{O}(C^2N)$.
- Often more accurate than HMM but less suitable for online learning.



Summary

- Like NB, HMM is a generative model while CRF is discriminative, like logistic regression
- CRF is undirected, so can find globally optimal sequences
- It has higher training cost than HMMs