4.2 Hidden Markov Models

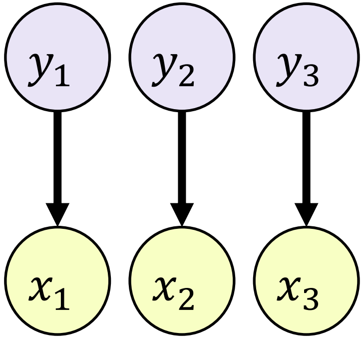
Edwin Simpson

Department of Computer Science,

University of Bristol, UK.

From Naïve Bayes to a Sequential Model: NB for Single Tokens

$$P(y_i|x\_i) \propto P(x_i|y_i)P(y_i)$$



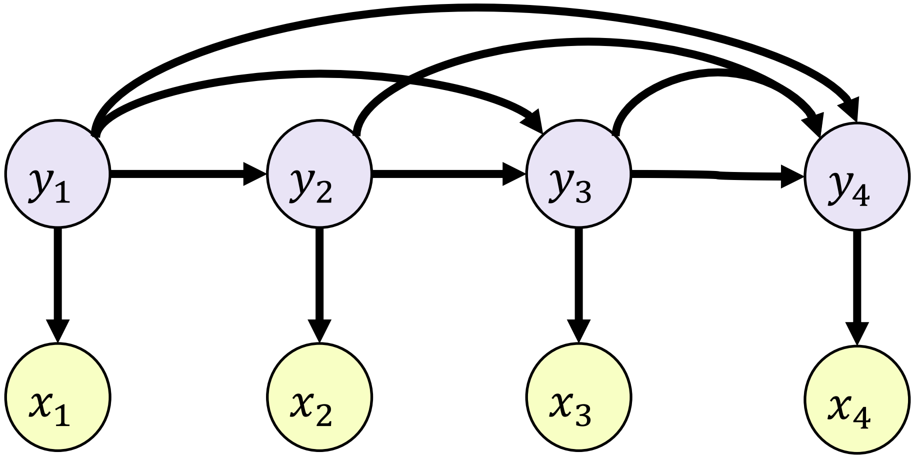
Maximum likelihood parameter estimates

$$P(y_i = c) = \frac{\text{num\_tokens\_with\_tag\_c}}{\text{total\_num\_toks}}$$

$$P(x_i = w|y_i = c) = \frac{count(w \mid c) + 1}{\sum_{w' \in V}(count(w' \mid c) + 1)}$$

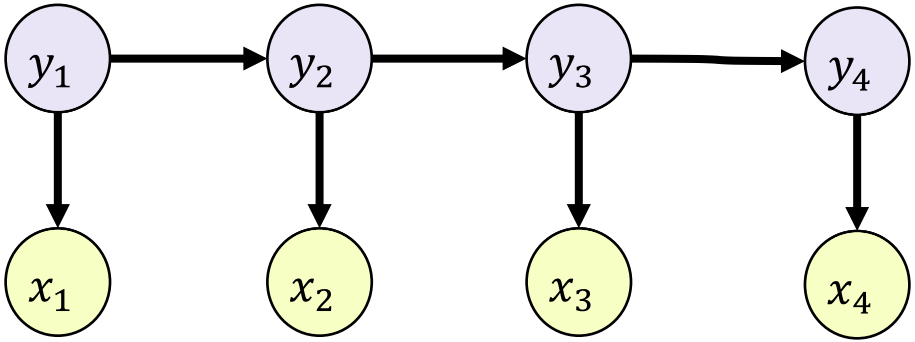
From Naïve Bayes to a Sequential Model:

$$P(y_i|x\_i) \propto P(x_i|y_i)P(y_i|\boldsymbol{y}_{-})$$



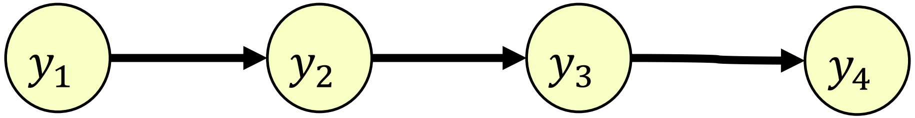
From Naïve Bayes to a Sequential Model:

$$P(y_i|x\_i) \propto P(x_i|y_i)P(y_i|\boldsymbol{y}_{-}) \approx P(x_i|y_i)P(y_i|y_{i-1})$$

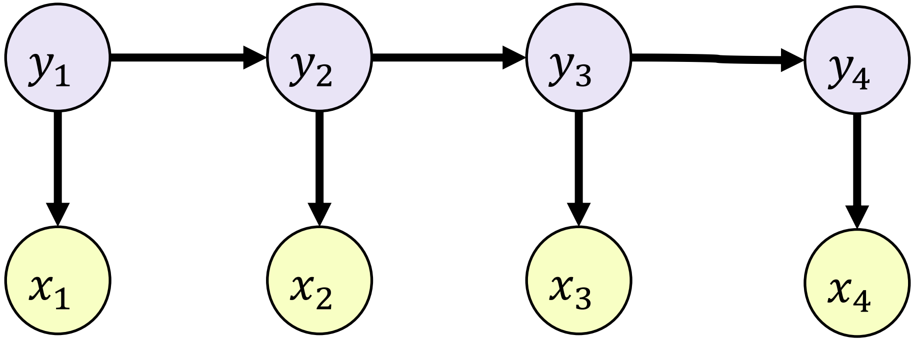


Markov Models and Hidden Markov Models

Markov assumption: $P(y_i|\boldsymbol{y}_{-}) \approx P(y_i|y_{i-1})$



Hidden Markov model (HMM): the states (y variables) are hidden and we observe x instead.



Transition Matrix

$$P(y_i|x\_i) \propto P(x_i|y_i)P(y_i|y_{i-1})$$

Transition matrix, *A*	**Current tag y_i**	
Previous tag y_{i-1}	0.5	0.5
	0.9	0.1

Initial probabilities π for p(y_1) instead of the transition matrix	
0.9	0.1

Maximum likelihood parameter estimate

$$P(y_i = c|y_{i-1} = d) = \frac{\text{num\_tokens\_with\_tag\_c\_preceded\_by\_d}}{\text{total\_num\_toks\_with\_tag\_d}}$$

Observation Model

$$P(y_i|x\_i) \propto P(x_i|y_i)P(y_i|y_{i-1})$$

- $P(x_i|y_i)$ is defined by the **observation** or **emission model, *B***;
- The HMM generalises the observation model to allow any type of word features:
  - Word embeddings – multivariate Gaussian as observation distribution;
  - *D* binary features, e.g., presence in sentiment lexicons -- $\prod_{d=1}^{D} P(x_{id}|y_i)$

Section 8.4, Speech and Language Processing (3rd edition draft), Jurafsky & Martin (2019).

Learning the Observation Model

HMMLearn: an Sklearn-style HMM library

- Words as observations (as in naïve Bayes):

Maximum likelihood parameter estimate

$$P(x_i = w|y_i = c) = \frac{count(w \mid c) + 1}{\sum_{w' \in V}(count(w' \mid c) + 1)}$$

- Word embeddings as Gaussian-distributed observations:

Maximum likelihood parameter estimates

$$P(\vec{\boldsymbol{x}}_{\boldsymbol{i}} = w|y_i = c) = \mathcal{N}(\vec{\boldsymbol{x}}_{\boldsymbol{i}}, \boldsymbol{\mu}_{\boldsymbol{c}}, \boldsymbol{\Sigma}_{\boldsymbol{c}})$$

Lin, C. C., et al. (2015, January). Unsupervised POS Induction with Word Embeddings. *HLT-NAACL*.

Summary

- The HMM models token likelihoods similarly to naïve Bayes.
- It also models the transition probabilities.
- To make learning and prediction tractable, we use the Markov assumption: the probability of each tag depends only on *m* predecessors in an order-m Markov model.