Discriminative Training Methods for Hidden Markov Models

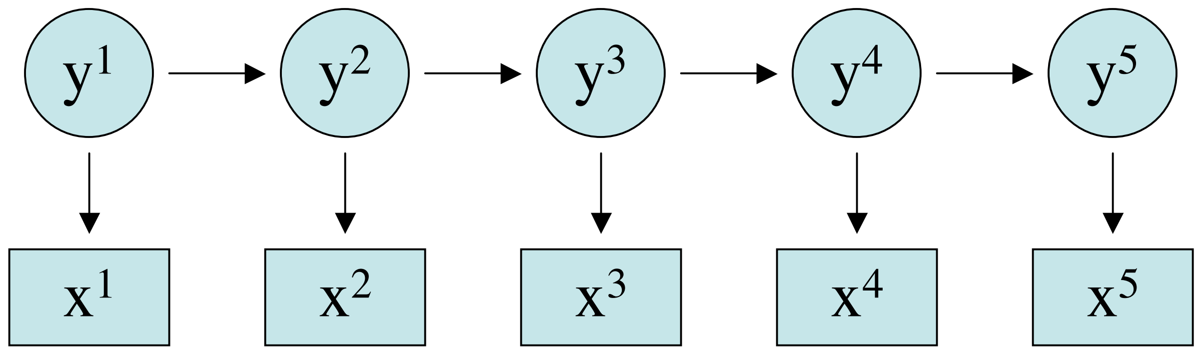
Michael Collins

Presenter: Alexandru Niculescu-Mizil

Sequence Prediction

- Input a sequence of observations $x = x^1 \ldots x^n$
  - e.g. x = the men saw the dog
- Output a sequence of labels $y = y^1 \ldots. y^n$
  - e.g. y = D N V D N
- In a probabilistic model, we want:
  - $\text{argmax}_y P(y|x) = \text{argmax}_y P(x,y)/P(x) =$
    $= \text{argmax}_y P(x,y)$

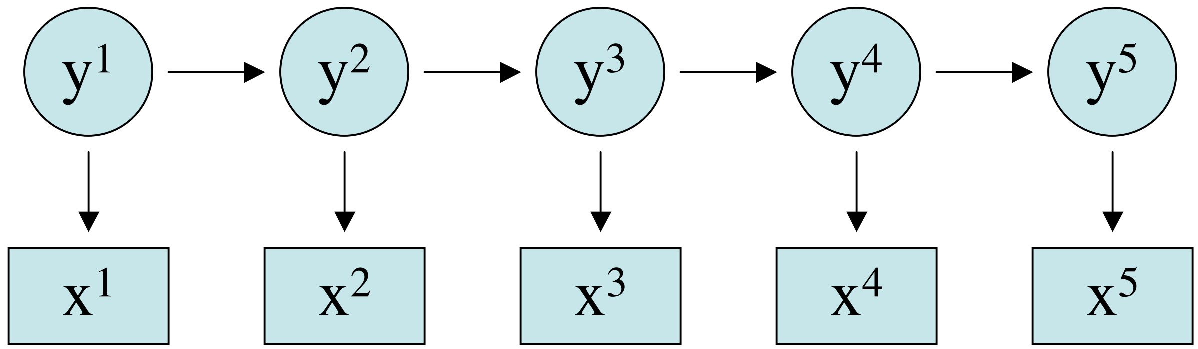
A probabilistic model for sequences



- $P(x,y) = P(x|y)P(y) = \prod P(x^i|y^i) \bullet P(y)$

  $= \prod_i P(x^i|y^i) \bullet \prod_i P(y^i|y^{i-1})$
- $\text{argmax}_y\, P(x,y)$ can be computed using Viterbi

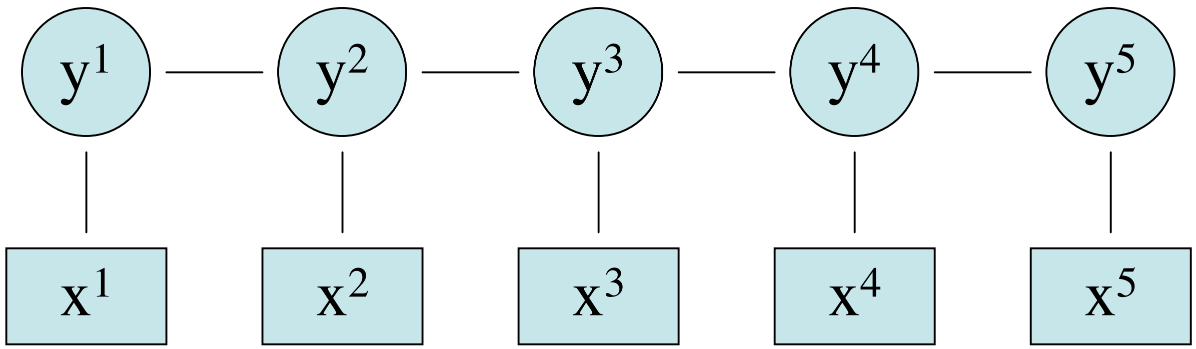
A probabilistic model for sequences



- $\log P(x,y) = \sum_i \log P(x^i|y^i) + \sum_i \log P(y^i|y^{i-1})$

$$= \sum_i (\sum_{w,t} \log P(w|t) * I(x^i = w, y^i = t)) +$$
$$+ \sum_i (\sum_{t,s} \log P(t|s) * I(y^i = t, y^{i-1} = s))$$
$$= \sum_{w,t} \log P(w|t) * \#(x^i = w, y^i = t) +$$
$$+ \sum_{t,s} \log P(t|s) * \#(y^i = t, y^{i-1} = s)$$

A non-probabilistic model for sequences



- score(x,y) = $\Sigma_{w,t}\, w_{a,t} * f_{a,t}(x,y) + \Sigma_{t,s}\, w_{t,s} * f_{t,s}(y)$
- given a train set (x_i, y_i), we want to find w s.t.

  argmax_z score(x,z) = y on future test examples (x, y)
- alternatively we want score(x, y) - score(x, z) > 0

  for all $z \neq y$

Perceptron Algorithm

Input: Training examples (x_i, y_i)

Initialization: $w = 0$

1. For $t = 1 \ldots T$, $i = 1 \ldots n$
    1. Calculate z_i the prediction for x_i given current w
    2. If there is a mistake, adjust w

Output: w

Perceptron for Classification

Input: Training examples (x_i, y_i)

Initialization: $w = 0$

1. For $t = 1 \ldots T$, $i = 1 \ldots n$
    1. Calculate $z_i = \text{sign}(w \cdot x_i)$
    2. If there is a mistake, $w = w + y_i * xi$

Output: w

Perceptron for Structured Outputs

Input: Training examples (x_i,y_i)

Initialization: $w = 0$

1. For $t = 1...T$, $i = 1...n$
    1. Calculate $z_i = \text{argmax}_z\, w \bullet f(x_i,z)$
    2. If there is a mistake, $w = w + f(x_i,y_i) - f(x_i,z_i)$

Output: w

Voted Perceptron, Averaged Perceptron

- $w_{1,1}, \ldots, w_{n,1}, \ldots, w_{n,T}$ - the parameters after each step of perceptron

- Voted perceptron:

  $$\text{majority}_{i,t}(\text{argmax}_z\, w_{i,t} \bullet f(x,z))$$

- Averaged perceptron:

  $$\text{argmax}_z\, (\, f(x,z) \bullet \text{avg}_{i,t}(w_{i,t})\,)$$

Convergence in Separable Case

- If there exist:
  - U s.t. $||U|| = 1$ and $U \bullet f(x_i,y_i) - U \bullet f(x_i,z) \geq d$ for all $z \neq y_i$ and for all i
  - R s.t. $||f(x_i,y_i) - f(x_i,z)|| \leq R$
- Then, number of mistakes $\leq R^2/d^2$

- Perceptron will converge to a solution that makes no mistakes on the training set.

Convervence in Inseparable Case

- For a given U and a desired margin d, let:
  - $m_i = U \bullet f(x_i, y_i)$ - $\text{argmax}_{z \neq yi}\ U \bullet f(x_i, z)$
    - the margin for example i
    - it can be negative if there is a mistake on example i
  - $e_i = \max\{0, d - m_i\}$
    - how far are we from achieving the desired margin
  - $D_{U,d} = (\sum e_i)^{1/2}$
- Then for one epoch,
  - Number of mistakes $\leq \min_{U,d}(R + D_{U,d})^2/d^2$

Generalization Error Bound

- For a sequence of n training examples,

- The probability that the voted perceptron makes a mistake on input n+1 is less than:

$$(2/n+1)E_{n+1}[\min_{U,d}(R + D_{U,d})^2/d^2]$$

Empirical Results

- Averaged perceptron is better than non-averaged one. - expected
- More, rare features better than less features for perceptron. -somewhat unexpected
- Averaged perceptron better than MaxEnt models.
- No statistical significance scores, only 2 problems.
- I didn’t find any thorough comparison with CRFs, but in the examples I found CRFs worked a little better than averaged perceptron.

Summary

- Discriminative learning for structured outputs.
  - does not require some of the independence assumptions
- The first "somewhat" max margin structured output learning
  - generalization bounds in terms of margin
- As long as inference is tractable, learning is tractable
  - convergence bounds
- Empirical results show improvements over MaxEnt models.