Max margin markov networks

(Generalisation of Struct SVM framework using the language / semantics of graphical models)

http://wikipedia-miner.cms.waikato.ac.nz/demos/annotate/?

wikipedia

I love Matrix → movie ✓ / math ✗

It is a must-watch movie. Neo

The character Neo was great

alphabetical ✗
personality ✓
behaviour ✗

eg of ϕ_n

($\phi_{overlap}$ (x, Matrix_movie) = Overlap between x & the wikipedia defn of Matrix-movie

x → text doc

ϕ decides if x should be treated as a bag or vector of words

$$\underbrace{\phi_{\#\text{noun matches}}}_{\text{eg of } \phi_n}(x, \text{Matrix\_movie}) = \dots$$

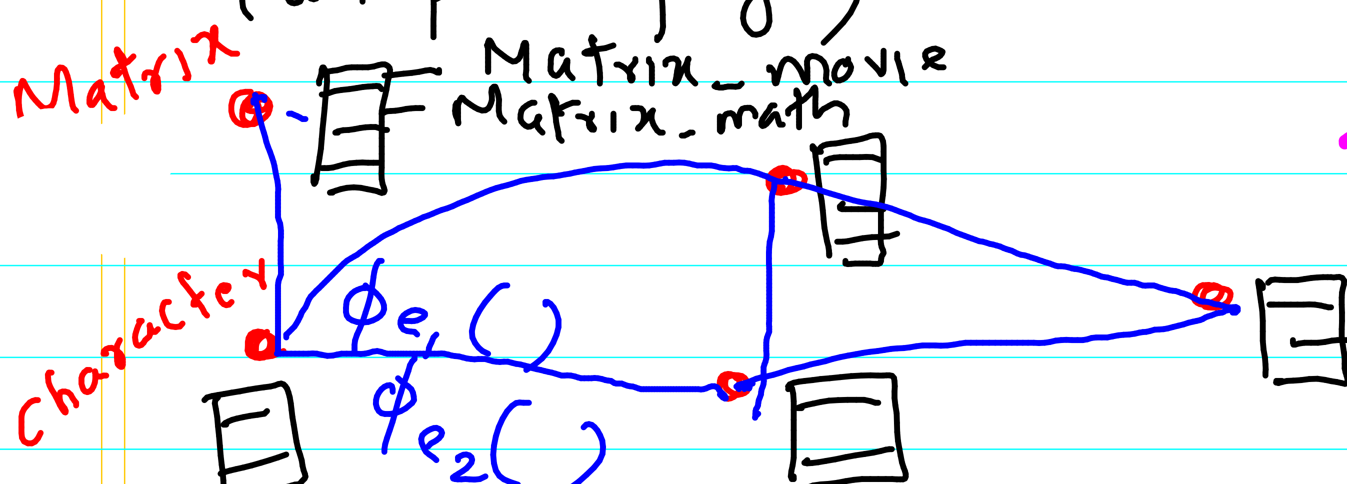
$$\underbrace{\phi_{\text{in\_link\_similarity}}}_{\text{eg of } \phi_e}(\text{Matrix\_movie}, \text{character\_personality}) = \#\text{ of common pages that link to Matrix\_movie \& character\_personality}$$

ϕ_n = feature for node

ϕ_e : feature for edge

● : Mentions in the doc

▤ : Possible labels (wikipedia pages)



Representation I

There is wt w_e & w_n for each ϕ_e & ϕ_n

In this representation, there is little logic to drop edges

Also note that unlike with generic undirected graphical models, where potentials are associated with all "maximal" cliques, we have associated potentials with

each node & each edge

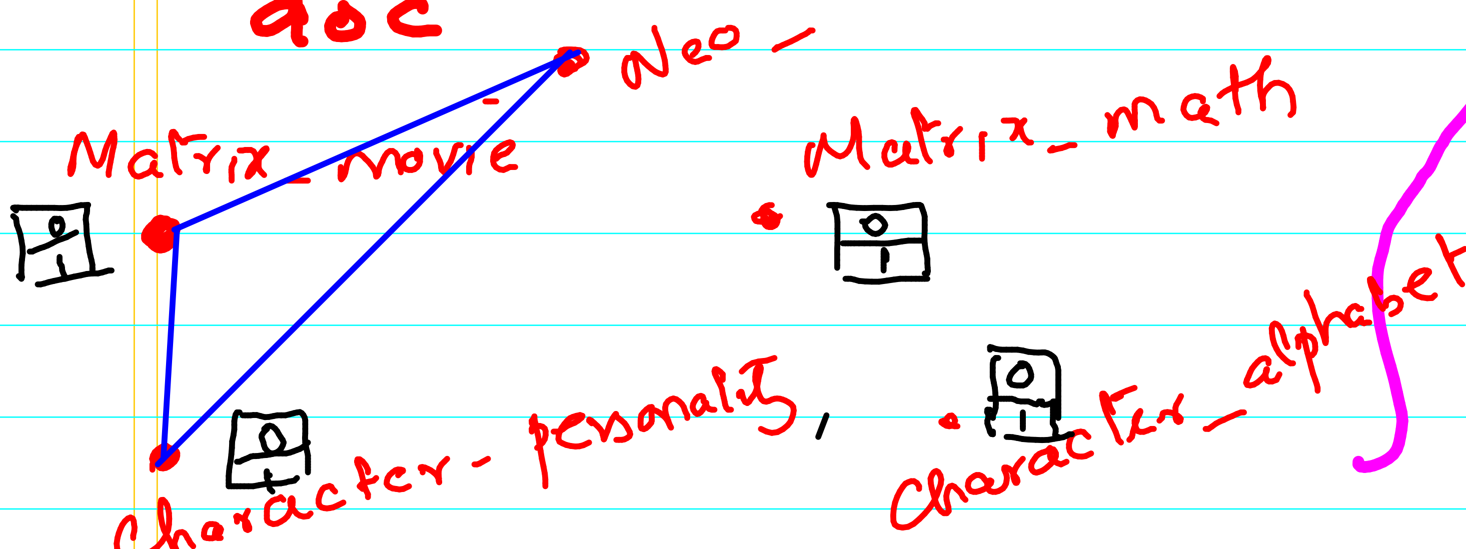
ϕ_n, ϕ_e: Very similar to previous case

H/W verify

$\boxed{0/1}$: Labels are binary (relevant/irrelevant)

• : **Node = all Wikipedia entities relevant to the doc**

Representation



Representation 1	Representation 2
(1) Association between mention & label directly retained ✓	(1) The association is lost (or needs to be maintained separately ✗
(2) Fewer nodes = m (mentions) ✓	(2) Large # nodes m x avg # labels per mention ✗
(3) Complete graph ✗	(3) Sparser & more meaningful graph ✓
(4) Learning & inference are hard ✗	(4) Learning & inference more tractable ✓

Learning w_n & w_e is actually a big gain or RHS

Connection between STRUCT SVM & CRF/Markov Models

1) In Structsvm

$$f(x,y) = w^T \phi(x,y)$$

In CRF / Markov models

$$p(y|x) = \exp\{w^T \phi(x,y)\}$$

It is found that a large family of prob distrs can be modeled using exponential family

eg. Poisson/Gaussian

ie $f(x,y) = \log p(y|x) = w^T \phi(x,y)$

The graphical model gives probabilistic interpretation/distribution

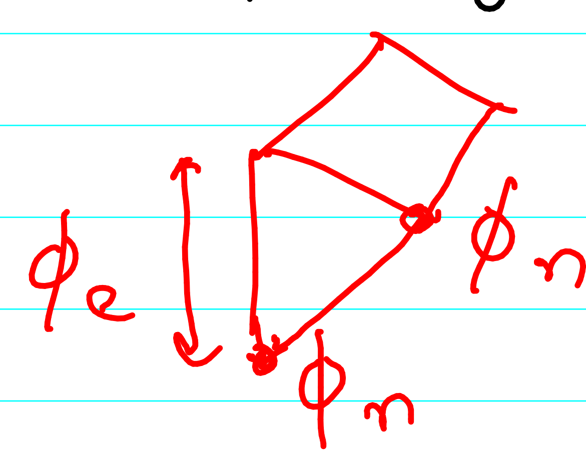
2) When markov networks are linear/Tree the max product (ie message passing) algo becomes

applicable (as in the case of
struct sum)

However, we get the additional
benefit of generalisation to
output spaces that are not
trees or chains!

Integer linear program formulation
for inference on max margin
markov network

$$\hat{y} = \arg\max_{y} \; w^T \phi(x^i, y) + \ell(y^i, y)$$



$Z_{n\ell}$ = A 0/1 variable associated with node n & label ℓ

$$y = \begin{bmatrix} \ell_1 \\ \ell_2 \\ \ell_n \end{bmatrix} \rightarrow \text{if } y(i) = \ell_i \text{ then } Z_{i\ell_i} = 1$$

$$\hat{Z} = \underset{Z}{\operatorname{argmax}} \sum_n \sum_{\ell_n} Z_{n\ell_n} \left(\sum_j w_{nj} \phi_{nj}(x^i, \ell^n) \right)$$

node features

makes it a QP

$$+ \sum_n \sum_{\ell_n} \sum_{n'} \sum_{\ell_{n'}} Z_{n\ell_n} Z_{n'\ell_{n'}} \cdot \sum_k w_k \phi_{e.k}(\ell_n, \ell_{n'})$$

set of edge features

H/W: Replace the QP with a LP

s.t. $\sum_{\ell_n} Z_{n\ell_n} = 1 \ \forall n$ & $Z_{n,\ell_n} \in \{0,1\}$

HINT: Look at slide # 39 onwards at

http://www.cse.iitb.ac.in/~cs717/notes/classNotes/StructuredOutput/mmmn.ppt