Case study: Training Neural Translation Models.

$\underline{\underline{x}} \equiv$ Sentence in English

$x_1, x_2 \ldots x_m$

$y \equiv$ sentence in hindi

$y_1\ y_2\ y_3 \ldots y_{\underline{\underline{n}}}$

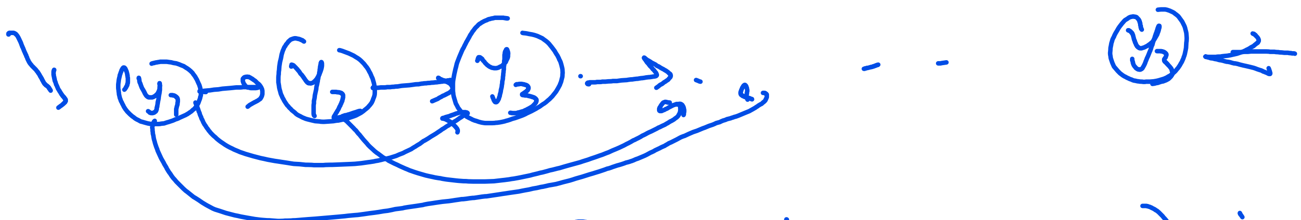
$y_j \in$ Hindi dictionary.

30K

$(30{,}000)^n$!!

$$P(y \mid x) = \prod_{j=1}^{n} P(y_j \mid \underbrace{y_1 \ldots y_{j-1}, x}_{})$$

a BN



Parametrization of $P(y_j \mid y_1 \ldots y_{j-1}, x)$ is using an neural network that can handle variable length inputs: eg: RNNs & Transformers.

Eg: RNN: $s_t \leftarrow$ embedding of $y_1 \ldots y_{j-1}$.

computed recursively.

Case study: Training Neural Translation Models.

$s_0 \leftarrow$ initial state

$s_t \leftarrow \text{LSTM\_cell}(\theta, s_{t-1}, y_{t-1})$

$v_t \leftarrow$ embedding of x

$$P(y_j \mid y_1 \ldots y_{j-1}, x) = \text{Softmax}(\{y_j\}, NN_\theta[s_t, v_t])$$

classification problem.

During inference:

Given a x, find the $\vec{y}$ for which $\boxed{P(Y \mid x)}$ is maximized. intractable for chain graphs.

In practice, people use greedy inference + algorithms such as beam-search

Training Non-linear parameters in CRFs

To be discussed later under Energy-based models.

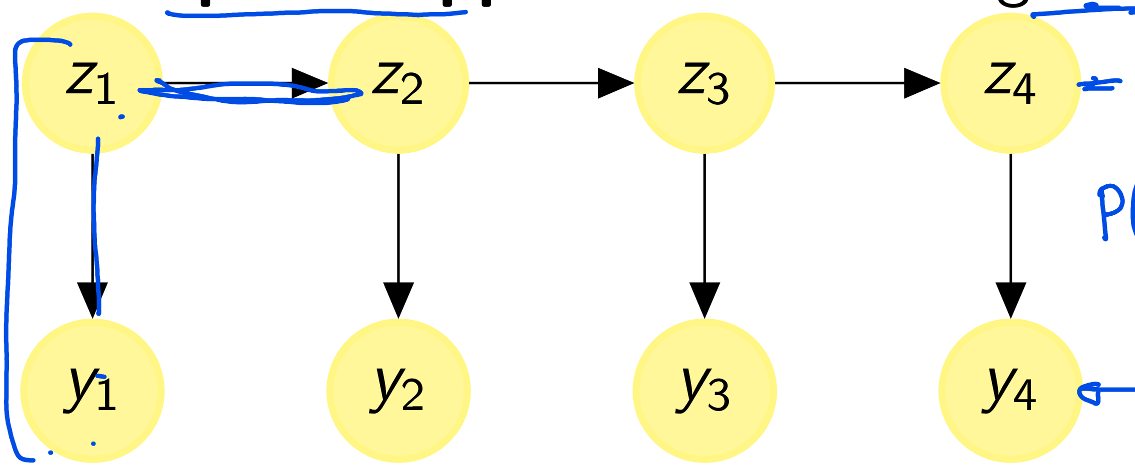
Learning with hidden ~~parameters~~ variables

Suppose only a subset of variables are observed. Other variables are hidden variables. How to learn the parameters of the graphical model?

$P(y|\cancel{x})$

GM $\sim y_1, y_2 \dots y_n$

Example of application: Training HMM



$$P(y_1 \dots y_n, z_1, \dots z_n) = \exp\left(\sum_c F_\theta(z_{c-1}, z_c)\right) \exp\left(\sum_c F_\theta(y_c, z_c)\right)$$

In CRF, we try to learn $\Pr(Y|X)$ with $D = \{\mathbf{x}^i, \mathbf{y}^i\}$, where all variables $y_1^i, y_2^i, \dots, y_n^i$ are present in the dataset. Here, in addition some variables $z_1^i, z_2^i, \dots, z_m^i$ are not present in D but is present in the graphical model.

Framework for learning

Let θ be the parameters of the graphical model.

$$P_{\theta,G}(y_1, y_2, \ldots, y_n, z_1, z_2, \ldots, z_m|\mathbf{x})$$
$$= \frac{1}{Z_\theta(\mathbf{x})} exp(\sum_C F_\theta(\mathbf{y}_C, \mathbf{z}_C, \mathbf{x}))$$

where C is the set of cliques in the graph. Suppose $D = \{(\mathbf{x}^i, \mathbf{y}^i) : i = 1, \ldots N\}$ is our dataset
Our goal during training is to find,

$$\theta^{ML} = arg \max_\theta \sum_{i=1}^{N} log P_\theta(\mathbf{y}^i|\mathbf{x}^i)$$
$$= arg \max_\theta \sum_{i=1}^{N} log \sum_{\mathbf{z}} P_{\theta,G}(\mathbf{y}^i, \mathbf{z}|\mathbf{x}^i)$$

The summation over $\mathbf{z}$ within the log make optimization difficult. Hence we approximate this objective. We will apply ideas from variational approximation to solve this problem. We will see that it will give rise to the well-known EM algorithm.

Variational Approach

We rewrite the original optimization in terms of new auxillary variables that we introduce.

$$\max_{\theta} \sum_{i=1}^{N} \log \sum_{\mathbf{z}:z_{1,\ldots,m}} P(\mathbf{y}^i, \mathbf{z}|\theta, \mathbf{x}^i)$$

$$\equiv \max_{\theta} \sum_{i=1}^{N} \max_{q_{i,\mathbf{z}}:\sum_{\mathbf{z}} q_{i,\mathbf{z}}=1} \sum_{\mathbf{z}} q_{i,\mathbf{z}} \log P(\mathbf{y}^i, \mathbf{z}|\theta, \mathbf{x}^i) - \sum_{\mathbf{z}} q_{i,\mathbf{z}} \log q_{i,\mathbf{z}}$$

≥ 0

The advantage of this rewriting is that now we do not have summation within the log.

We have two maximization problems to solve: over θ and over q variables.

The inner one can be solved in closed form for fixed value of θ.

The outer one can be solved like normal MLL training without hidden variables.

Example: CRFs

Example: CRFs

Variational approach (Proof)

We will show that:

$$\log \sum_{z=1}^{k} g(y,z) = \max_{q_1,q_2,\ldots,q_k} \boxed{\sum_{z=1}^{k} q_z \log g(y,z) - \sum_{z} q_z \log q_z}$$

$$s.t. \sum_{z=1}^{k} q_z = 1 \text{ and } q_z \geq 0$$

where $q_1, \ldots, q_k$ are auxiliary variables and

$$Q(q,g) = \sum_{z=1}^{k} q_z \log g(y,z) - \sum_{z} q_z \log q_z^{\,2}$$

$\max_{[q_1, q_2, \ldots q_k]} \quad \max Q(\vec{q}, g)$

s.t $\quad q_z \geq 0 \quad \sum_{z=1}^{k} q_z = 1 \Rightarrow \sum_z q_z - 1 = 0$