Sequential Models

$\mathbf{x} = (x_1, x_2, \cdots, x_n)$, $\mathbf{y} = (y_1, y_2, \cdots, y_n)$

$$\hat{\mathbf{y}} = \underset{\mathbf{y} \in \mathcal{Y}(\mathbf{x})}{\operatorname{argmax}} \ \psi(\mathbf{x}, \mathbf{y})$$

$\mathcal{Y}(\mathbf{x}) = \mathcal{Y}^n$

where $\mathcal{Y} = \{NN, VB, \cdots\}$

$|\mathcal{Y}(\mathbf{x})| = |\mathcal{Y}|^n$

scoring function on pairs of sequences.

$$V^n \times \mathcal{Y}^n \longrightarrow \mathbb{R}$$

vocabulary (V), label space ($\mathcal{Y}$), real number ($\mathbb{R}$)

$$\psi(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n+1} \psi(\mathbf{x}, y_i, y_{i-1}, i)$$

$\psi(\mathbf{x}, y_i, y_{i-1}, i) = W^T \cdot f(\mathbf{x}, y_i, y_{i-1}, i)$

linear model

feature-based

$$= W^T \cdot \sum_{i=1}^{n+1} f(\mathbf{x}, y_i, y_{i-1}, i)$$

decompose into a local scoring function

to make ~~making~~ the inference more tractable.

inference: making a series of interconnected labeling decisions.

HMM

$$\underset{\mathbf{y}}{\operatorname{argmax}} \ p(\mathbf{y}|\mathbf{x})$$

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} \ \log(p(\mathbf{x}, \mathbf{y}))$$

$$= \underset{\mathbf{y}}{\operatorname{argmax}} \ p(y_1) \prod_{i=2}^{n} p(y_i | y_{i-1}) \prod_{i=1}^{n} p(x_i | y_i)$$

CRF

normalizing constant (partition function)

$$P(y|x) = \frac{\exp(\psi(x,y))}{\boxed{\sum_{y' \in \mathcal{Y}(x)} \exp(\psi(x,y'))}} \qquad \boxed{\sum_{y' \in \mathcal{Y}(x)} \exp(\psi(x,y'))} = Z$$

almost identical to LR,
except that the label space is now sequence of tags.
requiring efficient algorithm for both:

- decoding : search for the best tag seq. y^*, given x and θ
- normalization : sum over all tag sequences $\mathcal{Y}(x)$ "y"

$$\hat{y} = \underset{y}{\operatorname{argmax}} \log P(y|x)$$

$$= \underset{y}{\operatorname{argmax}} \log \frac{1}{Z} \exp(\psi(x,y))$$

$$= \underset{y}{\operatorname{argmax}} \psi(x,y)$$

$$= \underset{y}{\operatorname{argmax}} w^T \cdot f(x, y_i, y_{i-1}, i)$$

LR training :

M training examples

$$L(x, y^*) = \sum_{j=1}^{M} \log P(y^{(j)*} \mid x^{(j)})$$

$$= \sum_{j=1}^{M} \left(w^T f(x^{(j)}, y^{(j)*}) - \log \sum_{y^* \in Y} \exp(w^T f(x^{(j)}, y^*)) \right)$$

$$\cancel{\frac{\partial L(x,y)}{\partial w}}$$

$$\frac{\partial}{\partial w} L(x, y^*) = f(x^{(j)}, y^{(j)*}) - \sum_{y^*} f(x^{(j)}, y^*) P_w(y^* \mid x^{(j)})$$

$$= f(x^{(j)}, y^{(j)*}) - \mathbb{E}_y [f(x^{(j)}, y)]$$

gold feature value

model's expectation of feature value.

CRF training :

M traing examples sequences sentences.

$$L(x, \mathbf{y}^*) = \sum_{j=1}^{M} \log P(\mathbf{y}^{(j)*} \mid \mathbf{x}^{(j)})$$

$$\frac{\partial}{\partial w} L(\mathbf{x}, \mathbf{y}^*) = f(\mathbf{x}^{(j)}, \mathbf{y}^{(j)*}) - \mathbb{E}_{\mathbf{y}} [f(\mathbf{x}^{(j)}, \mathbf{y})]$$

intractable

$$= f(\mathbf{x}^{(j)}, \mathbf{y}^{(j)*}) - \sum_{\mathbf{y}} f(\mathbf{x}^{(j)}, \mathbf{y}) P_w(\mathbf{y} \mid \mathbf{x}^{(j)})$$