Conditional Random Fields

Nurlanov Zhakshylyk, 574

Содержание

- Conditional Random Field, что это такое?
- Области применения
- Примеры

Определение CRFs

X – наблюдаемые величины, Y –ответы.

Пусть G = (V, E) – граф, такой что $Y = (Y_v)_{v \in V}$, то есть Y индексируется вершинами G.

Тогда (X, Y) – условное случайное поле (Conditional Random Field), если случайные величины Y_v, обусловленные X, удовлетворяют марковскому свойству относительно графа G:

$p(Y_v | X, Y_w, w \neq v) = p(Y_v | X, Y_w, w \sim v)$, где $w \sim v$ означает, что w и v соседние вершины в G.

Linear-chain CRFs

Definition 2.2. Let Y, X be random vectors, $\theta = \{\theta_k\} \in \Re^K$ be a parameter vector, and $\mathcal{F} = \{f_k(y, y', \mathbf{x}_t)\}_{k=1}^K$ be a set of real-valued feature functions. Then a *linear-chain conditional random field* is a distribution $p(\mathbf{y}|\mathbf{x})$ that takes the form:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}, \qquad (2.18)$$

where $Z(\mathbf{x})$ is an input-dependent normalization function

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}. \qquad (2.19)$$

Linear-chain CRFs

Notice that a linear chain CRF can be described as a factor graph over $\mathbf{x}$ and $\mathbf{y}$, i.e.,

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \Psi_t(y_t, y_{t-1}, \mathbf{x}_t) \tag{2.20}$$

where each local function Ψ_t has the special log-linear form:

$$\Psi_t(y_t, y_{t-1}, \mathbf{x}_t) = \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}. \tag{2.21}$$

Parameter Learning

Параметр θ находится максимизацией правдоподобия $p(Y_i|X_i;\ \theta)$.

Если все вершины G из экспоненциального семейства распределений и все вершины доступны во время обучения, то данная задача выпукла. Она может быть решена с помощью метода градиентного спуска, Квази-Ньютоновских методов.

Parameter Learning

$$\ell(\theta) = \sum_{i=1}^{N} \log p(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}; \theta). \qquad (5.1)$$

To compute the maximum likelihood estimate, we maximize $\ell(\theta)$, that is, the estimate is $\hat{\theta}_{\text{ML}} = \sup_\theta \ell(\theta)$.

After substituting in the CRF model (2.18) into the likelihood (5.1), we get the following expression:

$$\ell(\theta) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k(y_t^{(i)}, y_{t-1}^{(i)}, \mathbf{x}_t^{(i)}) - \sum_{i=1}^{N} \log Z(\mathbf{x}^{(i)}), \qquad (5.3)$$

Parameter Learning

- Regularization

$$\ell(\theta) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k(y_t^{(i)}, y_{t-1}^{(i)}, \mathbf{x}_t^{(i)}) - \sum_{i=1}^{N} \log Z(\mathbf{x}^{(i)}) - \sum_{k=1}^{K} \frac{\theta_k^2}{2\sigma^2}. \tag{5.4}$$

Области применения

- Text processing (NER, finding semantic roles in text, morphological analysis)
- Computer Vision (gesture recognition, semantic segmentation)
- Bioinformatics (RNA structural alignment, protein structure prediction)

Примеры: Named-Entity Recognition

- The New York Times, the White House.
- U.N. official Ekeus heads for Baghdad. U.N. – organization, Ekeus – person, Baghdad – location.
- BIO notation: B – first word of a mention, I – any subsequent words in the mention, O – words that do not reference any named entity.

Примеры: Named-Entity Recognition

- CoNLL 2003 data set

$\mathcal{Y} = \{\text{B-P}\textsc{er}, \text{I-PER}, \text{B-L}\textsc{oc}, \text{I-L}\textsc{oc}, \text{B-O}\textsc{rg}, \text{I-O}\textsc{rg}, \text{B-M}\textsc{isc}, \text{I-M}\textsc{isc}, \text{O}\}$

With this labeling, our example sentence looks like:

t	y_t	$\mathbf{x}_t$
0	B-ORG	U.N.
1	O	official
2	B-PER	Ekeus
3	O	heads
4	O	for
5	B-LOC	Baghdad

Примеры: Named-Entity Recognition

- $set\ F$ of feature function $f_k(y_t, y_{t-1}, x_t)$

$$f_{ij}^{\mathrm{LL}}(y_t, y_{t-1}, \mathbf{x}_t) = \mathbf{1}_{\{y_t=i\}}\mathbf{1}_{\{y_{t-1}=j\}} \forall i, j \in \mathcal{Y}. \tag{2.29}$$

For this problem, there are 9 different labels, so there are 81 label–label features. The second kind are *label-word* features, which are

$$f_{iv}^{\mathrm{LW}}(y_t, y_{t-1}, \mathbf{x}_t) = \mathbf{1}_{\{y_t=i\}}\mathbf{1}_{\{\mathbf{x}_t=v\}} \forall i \in \mathcal{Y}, v \in \mathcal{V}, \tag{2.30}$$

where $\mathcal{V}$ is the set of all unique words that appear in the corpus. For the CoNLL 2003 English data set, there are 21,249 such words, so there are 191,241 label-word features. Most of these features will not be very

Примеры: Named-Entity Recognition

- $set\ F$ of feature function $f_k(y_t, y_{t-1}, x_t)$

t	y_t	$\mathbf{x}_t$
0	B-ORG	(⟨START⟩, U.N., official)
1	O	(U.N., official, Ekeus)
2	B-PER	(official, Ekeus, heads)
3	O	(Ekeus, heads, for)
4	O	(heads, for, Baghdad)
5	B-LOC	(for, Baghdad, ⟨END⟩)

Примеры: Named-Entity Recognition

- $set\ F$ of feature function $f_k(y_t, y_{t-1}, x_t)$

t	y_t	$\mathbf{x}_t$
0	B-ORG	(⟨START⟩, U.N., official)
1	O	(U.N., official, Ekeus)
2	B-PER	(official, Ekeus, heads)
3	O	(Ekeus, heads, for)
4	O	(heads, for, Baghdad)
5	B-LOC	(for, Baghdad, ⟨END⟩)

Примеры: Named-Entity Recognition

- $set\ F$ of feature function $f_k(y_t, y_{t-1}, x_t)$

$$f_{iv}^{\mathrm{LW0}}(y_t, y_{t-1}, \mathbf{x}_t) = \mathbf{1}_{\{y_t=i\}}\mathbf{1}_{\{x_{t0}=v\}} \quad \forall i \in \mathcal{Y}, v \in \mathcal{V}$$
$$f_{iv}^{\mathrm{LW1}}(y_t, y_{t-1}, \mathbf{x}_t) = \mathbf{1}_{\{y_t=i\}}\mathbf{1}_{\{x_{t1}=v\}} \quad \forall i \in \mathcal{Y}, v \in \mathcal{V}$$
$$f_{iv}^{\mathrm{LW2}}(y_t, y_{t-1}, \mathbf{x}_t) = \mathbf{1}_{\{y_t=i\}}\mathbf{1}_{\{x_{t2}=v\}} \quad \forall i \in \mathcal{Y}, v \in \mathcal{V}.$$

However, we wish to use still more features than this, which will depend mostly on the word x_t. So we will add label-observation features, as decribed in Section 2.5. To define these, we will define a series of observation functions $q_b(x)$ that take a single word x as input. For each observation function q_b, the corresponding label-observation features f_{ib}^{LO} have the form:

$$f_{ib}^{\mathrm{LO}}(y_t, y_{t-1}, \mathbf{x}_t) = \mathbf{1}_{\{y_t=i\}} q_b(\mathbf{x}_t) \quad \forall i \in \mathcal{Y}. \qquad (2.31)$$

Примеры: Named-Entity Recognition

Table 2.2. A subset of observation functions $q_s(\mathbf{x}, t)$ for the CoNLL 2003 Engilsh named-entity data, used by Mccallum and Li [86].

W=v	$w_t = v$	$\forall v \in \mathcal{V}$
T=j	part-of-speech tag for w_t is j (as determined by an automatic tagger)	$\forall$POS tags j
P=I−j	w_t is part of a phrase with syntactic type j (as determined by an automatic chunker)	
Capitalized	w_t matches `[A-Z][a-z]+`	
Allcaps	w_t matches `[A-Z][A-Z]+`	
EndsInDot	w_t matches `[^\.]+.*\.`	
	w_t contains a dash	
	w_t matches `[A-Z]+[a-z]+[A-Z]+[a-z]`	
Acro	w_t matches `[A-Z][A-Z\\.]*\\.[A-Z\\.]*`	
Stopword	w_t appears in a hand-built list of stop words	
CountryCapital	w_t appears in list of capitals of countries	
⋮	many other lexicons and regular expressions	
$q_k(\mathbf{x}, t+\delta)$ for all k and $\delta \in [-1, 1]$		

Примеры: Image Labelling

- Пусть вектор $x = (x_1, x_2, \dots, x_T)$ представляет изображение размера $\sqrt{T} \times \sqrt{T}$. Пусть изображение серое, то есть каждый пиксель принимает значения от 0 до 255, означающих яркость.
- Задача: предсказать вектор $y = (y_1, y_2, \dots, y_T)$, где y_i равно метке класса. Пусть y_T может быть 0 или 1, означающие передний и задний план на изображении.

Примеры: Image Labelling

- Let $q(x_i)$ be a vector of features based on a region of the image around x_i, for example, using color histograms or image gradients.

$$f_m(y_i, x_i) = \mathbf{1}_{\{y_i = m\}} q(x_i) \quad \forall m \in \{0,1\}$$

$$g_{m,m'}(y_i, y_j, x_i, x_j) = \mathbf{1}_{\{y_i = m\}} \mathbf{1}_{\{y_j = m'\}} \nu(x_i, x_j) \quad \forall m, m' \in \{0,1\}$$

$$f(y_i, x_i) = \begin{pmatrix} f_0(y_i, x_i) \\ f_1(y_i, x_i) \end{pmatrix}$$

$$g(y_i, y_j, x_i, x_j) = \begin{pmatrix} g_{00}(y_i, y_j, x_i, x_j) \\ g_{01}(y_i, y_j, x_i, x_j) \\ g_{10}(y_i, y_j, x_i, x_j) \\ g_{11}(y_i, y_j, x_i, x_j) \end{pmatrix}$$

Примеры: Image Labelling

- N define the neighborhood relationship among pixels.

$$\nu(x_i, x_j) = \exp\{-\beta(x_i - x_j)^2\}$$
$$g(y_i, y_j, x_i, x_j) = \mathbf{1}_{\{y_i \neq y_j\}} \nu(x_i, x_j). \qquad (2.33)$$

Putting this all together, the CRF model is

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left\{ \sum_{i=1}^{T} \theta^\top f(y_i, x_i) + \sum_{(i,j)\in\mathcal{N}} \lambda^\top g(y_i, y_j, x_i, x_j) \right\}, \qquad (2.34)$$