Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequential Data

John Lafferty, Andrew McCallum, Fernando Pereira

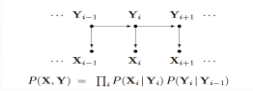
ICML 2001

Presented by Rohit Swarnkar and Guozhang Wang

Generative

- Model directly the joint distribution P(X, Y)
  - Need to numerate all the possible data points
- Enforce independence assumptions for learning efficiency
  - E.g., HMM

Standard tool is the hidden Markov Model (HMM).



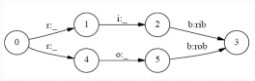
$$P(\mathbf{X}, \mathbf{Y}) = \prod_i P(\mathbf{X}_i \mid \mathbf{Y}_i)\, P(\mathbf{Y}_i \mid \mathbf{Y}_{i-1})$$

Discriminative

- Model conditional distributions P(Y|X)
- Relax strong independence assumptions made in generative models
  - Transition probability may depend on past and future observations
  - E.g., MEMM

The Label Bias Problem of Directed Models

- States might take little notice of observation
- Incorrect transitions cause unpredictability



Training Data:

"rob" "rob" "rob" "rib" "written" "wriggle" "bib" "nib"

Test Data: "rob"

P(1|0,r) * P(2|1,i) * P(3|2,b) = 0.75 * 0.5 * 0.5 = 0.1875

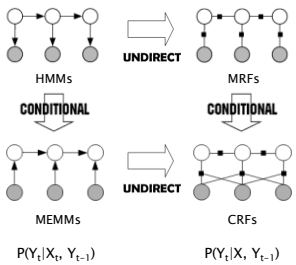
P(4|0,r) * P(5|4,i) * P(3|5,b) = 0.25 * 0.5 * 0.5 = 0.0625

Model will go from 0 to 1 after observing "r", then stuck there since no edge is from 1 to 5

Conditional Random Fields

- Conditional Random Fields (CRFs) is an undirected graphical models with two key features
  - It is a discriminative model relaxing the dependency assumptions between observations and labels
  - It is an undirected model whose conditional distribution is globally based on the whole observation sequence (thus no state bias problem)

Relationship With Other Models



$P(Y_t|X_t, Y_{t-1})$ $P(Y_t|X, Y_{t-1})$

Conditional Distribution in CRF

If the graph $G = (V, E)$ of Y is a tree, the conditional distribution over the label sequence Y = y, given X = x, by fundamental theorem of random fields has the form:

$$p_\theta(\mathrm{y} \mid \mathrm{x}) \propto \exp\left(\sum_{e \in E,k} \lambda_k f_k(e, \mathrm{y}|_e, \mathrm{x}) + \sum_{v \in V,k} \mu_k g_k(v, \mathrm{y}|_v, \mathrm{x})\right)$$

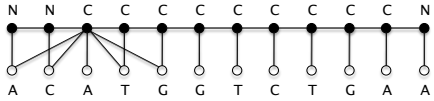
For every node and edge, take the weighted product of the features on that node or edge.

$$= \exp\left(< \boldsymbol{w}, \boldsymbol{\varphi}(\boldsymbol{X}, \boldsymbol{Y}) >\right)$$

Inference on CRF = solving argmax of this function

An Example

- CRF model for coding/noncoding genes



- Use Linear Chain CRF
- Labels Y_i: Coding/Noncoding DNA
- Data X_i: Gene sequence (A,T,C,G)
- Features: last 5 bases, adjacent 5 labels, frequency counts of last 50 bases, did start/stop codon (ATG/TGA) appear recently, (or any combination of these)

Learning CRF Models

- Need to maximize likelihood function

$$\arg\max_{\theta} \prod_{i=1}^{N} p_\theta(y^{(i)} \mid x^{(i)})$$

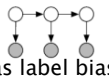
- We only have

$$p_\theta(y \mid x) \propto \exp\left(\sum_{e \in E,k} \lambda_k f_k(e, y|_e, x) + \sum_{v \in V,k} \mu_k g_k(v, y|_v, x) \right)$$

- Need to calculate normalization function $Z_\theta(x)$:
  - Use variant of forward-backwards algorithm to compute the product over all Y's

Experiment: Modeling Label Bias

- Generate Data from an HMM, with noise probability 3/32



- Classify using MEMM (which has label bias problem) and a CRF
- MEMM error: 42%
- CRF error: 4.6 %

Thanks

- Questions?