Conditional Random Fields:

Probabilistic Models for Segmenting and Labeling Sequence Data

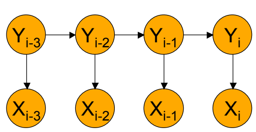
J. Lafferty, A. McCallum, F. Pereira. (ICML'01)

Presented by Kevin Duh
March 4, 2005
UW Markovia Reading Group

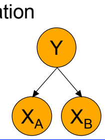
Outline

- Motivation:
  - HMM and CMM limitations
  - Label Bias problem
- Conditional Random Field (CRF) Definition
- CRF Parameter Estimation
  - Iterative Scaling
- Experiments
  - Synthetic
  - Part-of-speech tagging

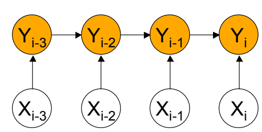
Hidden Markov Models (HMMs)



- Generative model p(X,Y)
  - Must enumerate all possible observation sequences ➔ Requires atomic representation
  - Assumes independence of features
    - same as Naïve Bayes



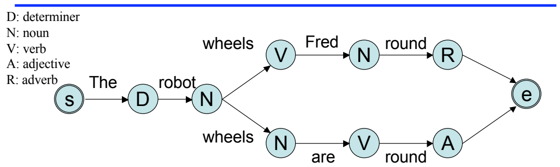
Conditional Markov Models (CMMs)



<u>Example</u>:
Maximum Entropy Markov Model (MEMM)

- Conditional model P(Y|X)
  - No effort wasted on modeling observations
  - Transition probability can depend on both past and future observations
  - Features can be dependent
- Suffers label-bias problem due to per-state normalization

Label Bias Example



Obs: "The robot wheels are round."

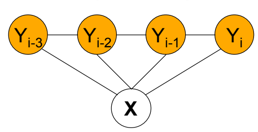
But if P(V|N,wheels) > P(N|N,wheels), then upper path is chosen regardless of obs.

After Wallach '02

Label Bias Problem

- The Problem: States with low-entropy next-state distributions ignore observations
  - Fundamental cause: Per-state normalization
    - "Conservation of score mass"
    - Transitions leaving a given state only compete against each other
- Solution
  - Model accounts for whole sequence at once
  - Prob. mass is amplified/dampened at individual transitions

Conditional Random Fields (CRFs)



- Single exponential model of joint probability of entire state sequence given observations
- Alternative view: Finite state model with un-normalized transition prob.

Definition of CRF

Def: A CRF is a undirected graphical model globally conditioned on the observation sequence

Graph: G=(V,E). V represents all Y.

(X,Y) is a CRF if, when conditioned on X, Y_v obeys the Markov property with respect to G:

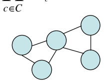
$$P(Y_v \mid X, Y_w, w \neq v) = P(Y_v \mid X, Y_w, w \sim v)$$

What does the distribution of a Random Field look like?

- Hammersley-Clifford Theorem:

  Conditional Independence Statements made by RF ⟷ $p(v_1, v_2, .., v_n) \triangleq \frac{1}{Z} \prod_{c \in C} \psi_{V_c}(v_c)$

  - Potential functions:
    - strictly positive and real value function
    - no direct probabilistic interpretation
    - represent "constraints" on configurations of random variables
      - An overall configuration satisfying more constraints will have higher probability



- Here, potential functions are chosen based on the Maximum Entropy principle

Maximum Entropy Principle

- MaxEnt says:
  - "When estimating a distribution, pick the max entropy distribution that respects all features *f(x,y)* seen in training data"
  - Constrained optimization problem

$$E_{\tilde{p}(x,y)}[f] = E_q[f]$$

i.e. $$\sum_{x,y} \tilde{p}(x,y) f(x,y) = \sum_{x,y} \tilde{p}(x) q(y \mid x) f(x,y)$$

  - Parametric form: $p_\lambda(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_k \lambda_k f_k(\mathbf{x}, \mathbf{y})\right)$

Parametric Form of CRF Distribution

Define each potential function as: $\psi_{Y_c}(\mathbf{y}_c) = \exp\left(\sum_k \lambda_k f_k(c, \mathbf{y}_c, \mathbf{x})\right)$

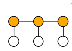
CRF distribution becomes:

$$p_\lambda(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{c \in C} \sum_k \lambda_k f_k(c, \mathbf{y}_c, \mathbf{x})\right)$$

Distinguish between two types of features:

$$p_\theta(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{e \in E, k} \lambda_k f_k(e, \mathbf{y}_e, \mathbf{x}) + \sum_{v \in V, k} \mu_k g_k(v, \mathbf{y}_v, \mathbf{x})\right)$$

Special Case of HMM-like Chain graph:



$$p_\theta(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{i,k} \lambda_k f_k(\mathbf{y}_{i-1}, \mathbf{y}_i, \mathbf{x}) + \sum_{i,k} \mu_k g_k(\mathbf{y}_i, \mathbf{x})\right)$$

CRF Parameter Estimation

- Iterative Scaling:
  - Maximizes likelihood $O(\theta) = \sum_{i=1}^{N} \log p_\theta(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)}) \propto \sum_{\mathbf{x},\mathbf{y}} \tilde{p}(\mathbf{x},\mathbf{y}) \log p_\theta(\mathbf{y} \mid \mathbf{x})$ by iteratively updating

$$\lambda_k \leftarrow \lambda_k + \delta\lambda_k \qquad \mu_k \leftarrow \mu_k + \delta\mu_k$$

  - Define auxilliary function A() s.t. $A(\theta', \theta) \le O(\theta') - O(\theta)$
    - Initialize each λ_k
    - Do until convergence:

      Solve $\frac{dA(\theta', \theta)}{d\delta\lambda_k} = 0$ for each $\delta\lambda_k$

      Update parameter: $\lambda_k \leftarrow \lambda_k + \delta\lambda_k$

CRF Parameter Estimation

- For chain CRF, setting $\frac{dA(\theta',\theta)}{d\delta\lambda_k}=0$ gives

$$\tilde{E}[f_k] \triangleq \sum_{\mathbf{x},\mathbf{y}} \tilde{p}(\mathbf{x},\mathbf{y}) \sum_{i=1}^{n+1} f_k(\mathbf{y}_{i-1},\mathbf{y}_i,\mathbf{x})$$
$$= \sum_{\mathbf{x},\mathbf{y}} \tilde{p}(\mathbf{x}) p(\mathbf{y}\mid\mathbf{x}) \sum_{i=1}^{n+1} f_k(\mathbf{y}_{i-1},\mathbf{y}_i,\mathbf{x}) \exp\left(\delta\lambda_k T(\mathbf{x},\mathbf{y})\right)$$

  - $T(\mathbf{x},\mathbf{y}) = \sum_{i,k} f_k(\mathbf{y}_{i-1},\mathbf{y}_i,\mathbf{x}) + \sum_{i,k} g_k(\mathbf{y}_i,\mathbf{x})$ is total feature count
- Unfortunately, T(**x,y**) is a global property of (**x,y**)
  - Dynamic programming will sum over sequences with potentially varying T. Inefficient exp sum computation

Algorithm S (Generalized Iterative Scaling)

- Introduce global slack feature s.t. T(**x,y**) becomes constant S for all (**x,y**)

$$S(\mathbf{x},\mathbf{y}) \triangleq S - \sum_{i,k} f_k(\mathbf{y}_{i-1},\mathbf{y}_i,\mathbf{x}) + \sum_{i,k} g_k(\mathbf{y}_i,\mathbf{x})$$

- Define forward and backward variables

$$\alpha_i(y\mid\mathbf{x}) = \alpha_{i-1}(y\mid\mathbf{x}) \exp\left(\sum_{i,k} f_k(\mathbf{y}_{i-1},\mathbf{y}_i,\mathbf{x}) + \sum_{i,k} g_k(\mathbf{y}_i,\mathbf{x})\right)$$
$$\beta_i(y\mid\mathbf{x}) = \beta_{i+1}(y\mid\mathbf{x}) \exp\left(\sum_{i,k} f_k(\mathbf{y}_{i+1},\mathbf{y}_i,\mathbf{x}) + \sum_{i,k} g_k(\mathbf{y}_{i+1},\mathbf{x})\right)$$

Algorithm S

The update equations become:

$$\delta\lambda_k = \frac{1}{S}\log\frac{\tilde{E}f_k}{Ef_k}, \quad \delta\mu_k = \frac{1}{S}\log\frac{\tilde{E}g_k}{Eg_k}$$

Rate of convergence governed by S

Where
$$Ef_k = \sum_{\mathbf{x}} \tilde{p}(\mathbf{x}) \sum_{i=1}^{n+1} \sum_{y',y} f_k(e_i, \mathbf{y}|_{e_i} = (y',y), \mathbf{x}) \times \frac{\alpha_{i-1}(y'\mid\mathbf{x})\, M_i(y',y\mid\mathbf{x})\, \beta_i(y\mid\mathbf{x})}{Z_\theta(\mathbf{x})}$$
$$Eg_k = \sum_{\mathbf{x}} \tilde{p}(\mathbf{x}) \sum_{i=1}^{n} \sum_{y} g_k(v_i, \mathbf{y}|_{v_i} = y, \mathbf{x}) \times \frac{\alpha_i(y\mid\mathbf{x})\, \beta_i(y\mid\mathbf{x})}{Z_\theta(\mathbf{x})}.$$

Note $p_\theta(\mathbf{Y}_i = y\mid\mathbf{x}) = \frac{\alpha_i(y\mid\mathbf{x})\,\beta_i(y\mid\mathbf{x})}{Z_\theta(\mathbf{x})}$ is like posterior as in HMM

Algorithm T (Improved Iterative Scaling)

The equation we want to solve
$$\tilde{E}[f_k] = \sum_{\mathbf{x},\mathbf{y}} \tilde{p}(\mathbf{x}) p(\mathbf{y}\mid\mathbf{x}) \sum_{i=1}^{n+1} f_k(\mathbf{y}_{i-1},\mathbf{y}_i,\mathbf{x}) \exp\left(\delta\lambda_k T(\mathbf{x},\mathbf{y})\right)$$
is polynomial in $\exp(\delta\lambda_k)$

So can be solved with Newton's method

Define $T(\mathbf{x}) \triangleq \max T(\mathbf{x},\mathbf{y})$ $\quad \tilde{E}[f_k] = \sum_{\mathbf{x},\mathbf{y}} \tilde{p}(\mathbf{x}) p(\mathbf{y}\mid\mathbf{x}) \sum_{i=1}^{n+1} f_k(\mathbf{y}_{i-1},\mathbf{y}_i,\mathbf{x}) \exp\left(\delta\lambda_k T(\mathbf{x})\right)$

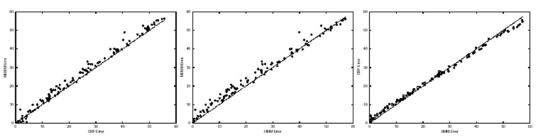
Then: $\sum_{t=0}^{T_{\max}} \left(\sum_{\{\mathbf{x},\mathbf{y}|T(\mathbf{x})=t\}} \tilde{p}(\mathbf{x}) p(\mathbf{y}\mid\mathbf{x}) \sum_{i=1}^{n+1} f_k(\mathbf{y}_{i-1},\mathbf{y}_i,\mathbf{x}) \exp\left(\delta\lambda_k\right)^t \right)$

Now, let $a_{k,t}$, $b_{k,t}$ be E[f_k|T(**x**)=t] $\quad a_{k,t} = \sum_{\mathbf{x},\mathbf{y}} \tilde{p}(\mathbf{x}) p(\mathbf{y}\mid\mathbf{x}) \sum_{i=1}^{n+1} f_k(\mathbf{y}_{i-1},\mathbf{y}_i,\mathbf{x}) \delta(t, T(\mathbf{x}))$

UPDATE: $\delta\lambda_k = \log\beta_k$, $\delta\mu_k = \log\gamma_k$ $\quad \sum_{i=0}^{T_{\max}} a_{k,t}\,\beta_k^t = \tilde{E}f_k, \quad \sum_{i=0}^{T_{\max}} b_{k,t}\,\gamma_k^t = \tilde{E}g_k$

Experiments with Synthetic Data

1. Modeling Label Bias:
   - Generated data by Fig 1 stochastic FSA
   - CRF: 4.6%, MEMM: 42% error rate
2. Modeling mixed-order sources
   - Generate data by $\alpha p(\mathbf{y}_i \mid \mathbf{y}_{i-1}, \mathbf{y}_{i-2}) + (1-\alpha) p(\mathbf{y}_i \mid \mathbf{y}_{i-1})$



POS tagging experiment

Wall Street Journal dataset; 45 POS tags

model	*error*	*oov error*
HMM	5.69%	45.99%
MEMM	6.37%	54.61%
CRF	5.55%	48.05%
MEMM+	4.81%	26.99%
CRF+	4.27%	23.76%

+Using spelling features

Training time:
Initial value is result of MEMM training (100 iter)
Convergence for CRF^+ took 1000 more iterations

Conclusion / Summary

- CRFs are undirected graphical models globally conditioned on observations
- Advantages of CRFs:
  - Conditional model
  - Allows multiple interacting features
- Disadvantage of CRFs:
  - Slow convergence during training
- Potential future directions:
  - More complex graph structures
  - Faster (approximate) Inference/Learning algorithms
  - Feature selection/induction algo for CRFs…

Useful References

- Hanna Wallach. **Efficient Training of Conditional Random Fields.** M.Sc. thesis, Division of Informatics, University of Edinburgh, 2002.
- Della Pietra, S., Della Pietra, V., & Lafferty, J. (1997). **Inducing features of random fields**. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19, 380–393.
- Berger, A. L., Della Pietra, S. A., & Della Pietra, V. J. (1996). **A maximum entropy approach to natural language processing**. Computational Linguistics, 22.