# Conditional Random Field and its Application

Keyu Wang

13. December 2008

# Outline

# Outline

# Conditional Random Fields (CRFs)

Conditional Random Fields (CRFs) is

- proposed in 2001 by John Lafferty
- a probabilistic framework for labeling and segmenting data
- a discriminative model based on undirected graph
- generalized Hidden Marcov Model(HMM)

# Notations

In what follows,

- $\mathcal{Y}$: a finite label alphabet
- $\mathbf{X}$: a random variable over data sequences to be labeled
- $\mathbf{Y}$: a random variable over corresponding label sequences
- $\boldsymbol{\lambda}$: the weight vector

# Further Notations for CRFs

Further Notations

$G$: an undirected graph which is relevant to the model

$C_p$: a clique in $G$

$\mathbf{Y}_{C_p}$: the set of components of $\mathbf{Y}$ associated with the vertices of $C_p$

$\mathbf{f}_{C_p}$: the feature function vector according to the clique $C_p$

$\boldsymbol{\lambda}_{C_p}$: the weight vector corresponding with $\mathbf{f}_{C_p}$

# General CRFs

Now we present the general definition of a Condtional Random Field [Lafferty et al., 2001].

## Definition

*A* **conditional random field** *is a distribution* $P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\lambda})$ *that takes the form*

$$P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\lambda}) = \frac{1}{Z(\mathbf{X})} \exp\left(\sum_{C_p} \langle \boldsymbol{\lambda}_{C_p}, \mathbf{f}_{C_p}(\mathbf{Y}_{C_p}, \mathbf{X}) \rangle\right),$$

*where* $Z(\mathbf{X}) = \sum_{\mathbf{Y}} \exp\left(\sum_{C_p} \langle \boldsymbol{\lambda}_{C_p}, \mathbf{f}_{C_p}(\mathbf{Y}_{C_p}, \mathbf{X}) \rangle\right)$ *which is a normalization factor.*

# Parameter Estimation

Maximum Conditional Likelihood Training

- choose parameter values such that the logarithm of the conditional likelihood (conditional log-likelihood) is maximized
- the training data $\{(x^{(k)}, y^{(k)})\}$ independently and identically distributed
- **Conditional likelihood**: the product of $P(\mathbf{y}^{(k)}|\mathbf{x}^{(k)}, \boldsymbol{\lambda})$, as a function of the parameters $\boldsymbol{\lambda}$

# Conditional Log-Likelihood for a CRF

For a CRF, the **conditional log-likelihood** is given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\lambda}) &= \log(\prod_k P(\mathbf{y}^{(k)}|\mathbf{x}^{(k)}, \boldsymbol{\lambda})) \\ &= \sum_k [\log \frac{1}{Z(\mathbf{x}^{(k)})} + \sum_{C_p} \langle \boldsymbol{\lambda}_{C_p}, \mathbf{f}_{C_p}(\mathbf{y}_{C_p}^{(k)}, \mathbf{x}^{(k)}) \rangle]\end{aligned}$$

This function is concave, guaranteeing convergence to the global maximum.

# Optimization

Optimization

- can be formulated as an unconstrained optimization problem
- in general, cannot be maximized in closed form
- numerical optimization: quasi-Newton, conjugate gradient, ...

# Using Viterbi just like HMM

CRFs use the well-known **Viterbi-method** to compute the conditional likelihoods.

- Viterbi variables: $i \in \{1, \ldots, n\}, l \in \mathcal{Y}$

$$\delta_{l,1} = \sum_k \lambda_k f_k(z_0, l, \mathbf{x}, 1)$$

$$\delta_{l,i+1} = \max_{z_i \in \mathcal{Y}} \{\delta_{z_i,i} + \sum_k \lambda_k f_k(z_i, l, \mathbf{x}, i+1)\}.$$

- The Viterbi-score:

$$\delta_{*,n} = \max_{l \in \mathcal{Y}} \delta_{l,n} \quad \propto \quad \max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}).$$

# Evaluation

The prediction has to fulfill two competing demands.

- TP: true positive label
- FP: false positive label
- FN: false negative label

Then the two demands are measured respectively

- coverage(recall): Cov=TP/(TP+FN)
- accuracy(precision): Acc=TP/(TP+FP)

# Outline

# NLP Problem

**P**art **O**f **S**peech **T**agging Problem (POST):

- There are commonly 8 parts of speech in English.
- Some words can represent more than one part of speech at different times.

**Example**

Books are made of ink, paper, and glue.
"books" noun.
Deborah waits patiently while Bridget books the tickets.
"books" verb.

# Addresse Classification

```
Soldotna, |5
AK |6
99669 |7

9112 |1
Mendenhall Mall Road, |3
Juneau, |5
AK |6
99801 |7

Mile K Beach Road # 1, |3
Kenai, |5
AK |6
99611 |7

Mile K Beach Road # 1, |3
Kenai, |5
AK |6
99611 |7

Mi K Beach Road # 2, |3
Kenai, |5
AK |6
99611 |7
```

Abbildung: Traing addresses

Training file

- 50 American post addresses
- segmented as house number, road, city name ...
- labeled 1, 2, 3 ...

# Addresse Classification

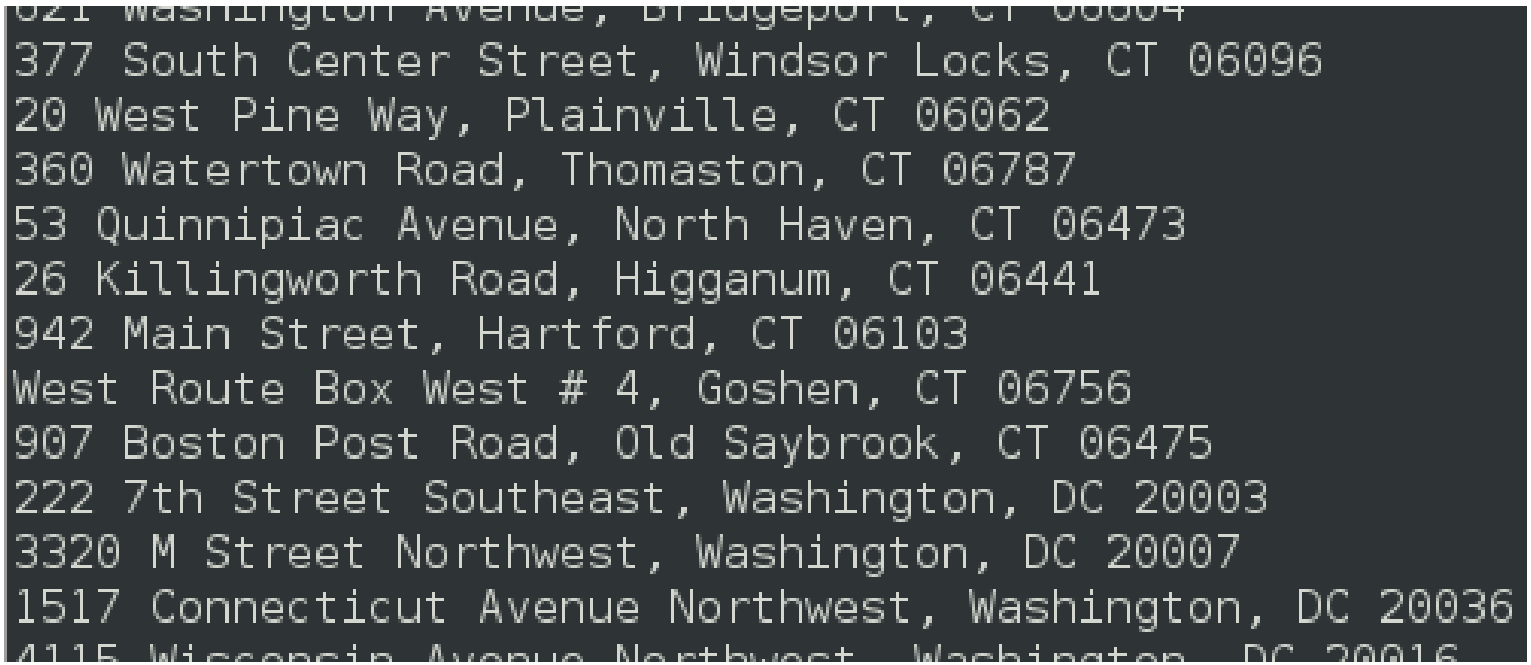



Abbildung: Test addresses

690 test addresses

# Results

```
run:
Number of features :220
Iter 0 log likelihood -858.1463757333934 norm(grad logli) 269.6606170
Number of training records51
Iter 1 log likelihood -642.7182853152169 norm(grad logli) 187.7821060
Iter 2 log likelihood -392.79123080713117 norm(grad logli) 301.597109
Iter 3 log likelihood -210.47823596105772 norm(grad logli) 246.593045

Iter 81 log likelihood -5.03380980019481 norm(grad logli) 0.0202677


Calculations:

Label   True+   Marked+ Actual+ Prec.   Recall  F1
0:      598     605     598     98.842  100.0   99.41762806650507
1:      0       0       20      0.0     0.0     NaN
2:      2263    2285    2443    99.037  92.632  95.72748210717435
3:      0       0       0       0.0     0.0     NaN
4:      1547    1675    1558    92.358  99.293  95.700026548257
5:      690     717     690     96.234  100.0   98.08086264357858
6:      690     717     690     96.234  100.0   98.08086264357858
-------------------------------------------------------------
0v:     5788    5999    5999    96.482  96.482  96.482
```

# Outline

# Biological Background

Protein and Protein Complex

- Proteins: linear polymers built from 20 different amino acids (an individual amino acid is called a residue)
- a group of two or more associated proteins
- formed by protein-protein interaction that is stable over time
- Protein Data Bank (PDB) *http://www.pdb.org*

# Biological Background

Protein-protein Interaction

- association of protein molecules and the study of these associations from the perspective of biochemistry
- important for many biological functions
- of central importance for virtually every process in a living cell

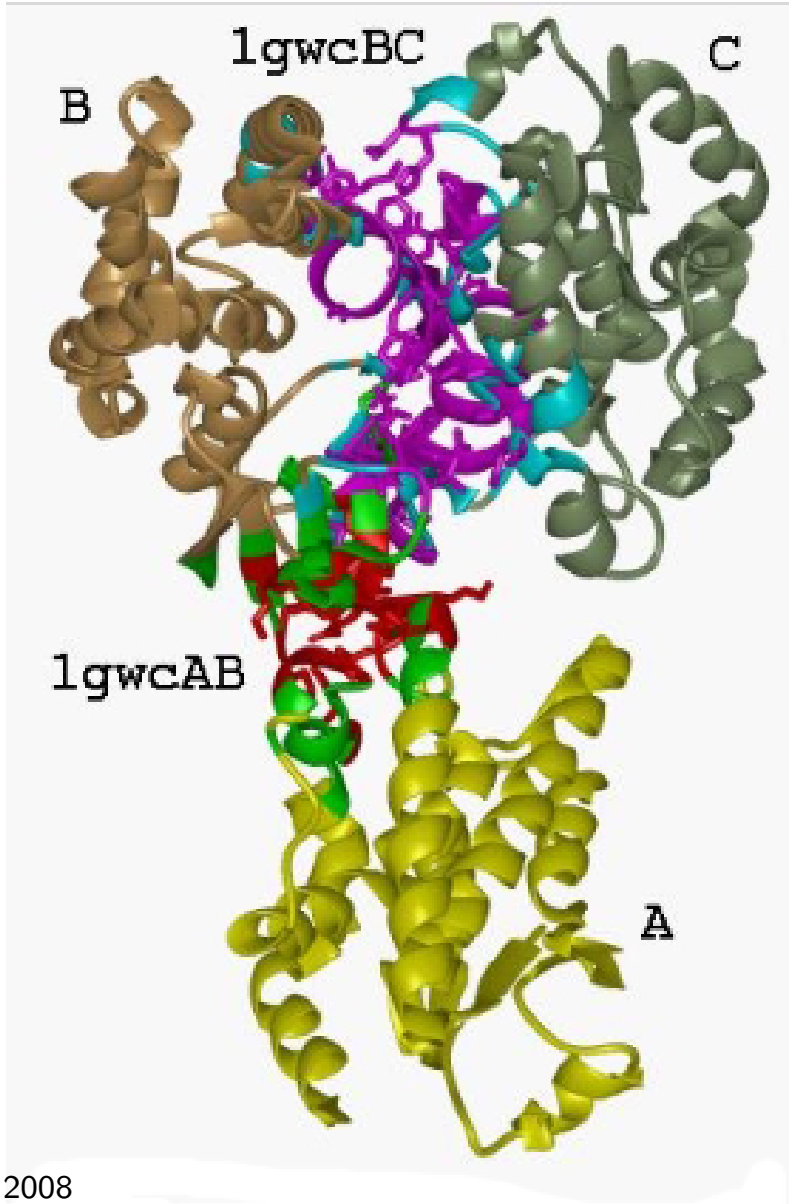
# Interface Residue

We define a interface residue according to a pair of interacting proteins.

**Definition**

*A residue is considered to be a* **interface residue** *if the distance between any of its atom and any atom of its interacting chains is* $<5\text{Å}$*. The* **interface** *of a protein is the set of all its interface residues.*

# Protein Complex 1gwc



# Protein-protein Interaction Sites Prediction (PPISP)

PPISP is a field combining bioinformatics and structural biology in an attempt to identify and catalog interactions between pairs or groups of proteins.

- experimentally can be proved but expensive
- up to now the accuracy of predictions cannot be satisfied

# Our Problem

We address this prediction as a sequence labeling task.
Each position along the protein sequence is assigned a state label, either I (interface residue) or N (noninterface residue).

- Given: a protein sequence with its 3D-structure
- Goal: to predict the most possible interface

# Challenges

Challenges in this project

- suitable graphs for protein
- biological features
- noisy data in database
- ...

# Literature


J. Lafferty, A. McCallum, and F. Pereira.
Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
*Proc. ICML-01*, pages 282–289, 2001.


C. Sutton and A. McCallum.
*An Introduction to Conditional Random Fields for Relational Learning*.
MIT Press, 2007.


H.X. Zhou and S. Qin.
Interaction-site prediction for protein complexes: a critical assesment.
*Bioinformatics*, pages 2203–2209, 2007.

Thank you for your attention

Merry Christmas and happy new year!