Conditional Random Fields (CRF)

Hasala Marakkalage
Ph.D. Candidate
Singapore University of Technology and Design

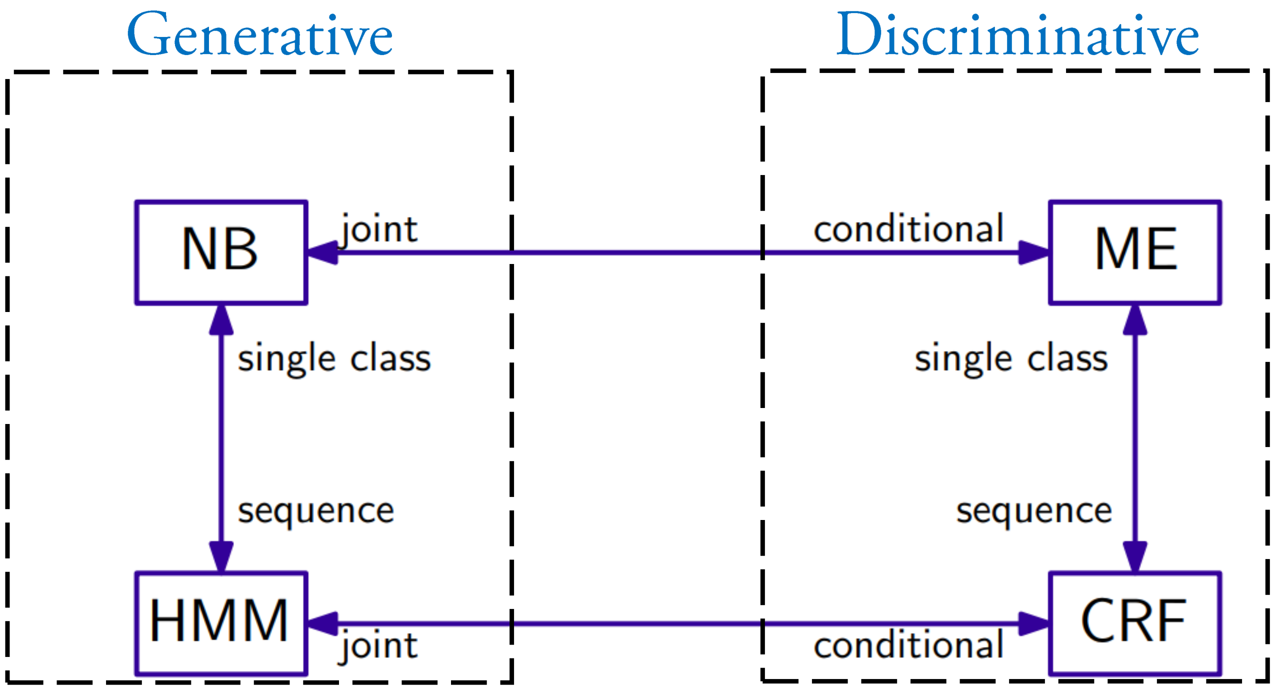
Date: 19th October, 2017

Introduction

Conditional Random Fields (CRFs) are a probabilistic framework for labeling and segmenting sequential data.

A CRF is a form of undirected graphical model that defines a single log-linear distribution over label sequences, given a particular observation sequence.

Overview of probabilistic models



Introduction

Motivation: The need to segment and label sequences arises in many problems

Hidden Markov Models (HMMs) are widely used in such problems
e.g.: part-of-speech (POS) tagging, align biological sequences

Generative models

HMMs are generative models, assigning joint probability to observation and label sequences

$$p(\boldsymbol{X},\boldsymbol{Y}) \text{ where } \underset{\substack{\uparrow \\ \text{Observation sequence}}}{\boldsymbol{X}} \text{ and } \underset{\substack{\uparrow \\ \text{Label sequence}}}{\boldsymbol{Y}} \text{ are random variables}$$

But, there are few drawbacks in generative models

Introduction

Drawbacks of generative models

To define joint probability over observation and label sequences, a generative model needs to enumerate all possible observation sequences (e.g.: all possible words in speech tagging)

It is not practical to represent long range dependencies of observations, hence inference problem for such models is intractable (hard to deal with)

This difficulty is the main motivation to look at conditional models as an alternative

Conditional (discriminative) models

A conditional model specifies the probabilities of possible label sequences given an observation sequence

$$\text{conditional probability } p(\boldsymbol{Y}|\boldsymbol{x})$$

Conditional probability of the label sequence can depend on arbitrary, non-independent features of the observation sequence, without forcing the model to account for the distribution of those dependencies.

Introduction

The probability of transition between labels not only depend on current state, but also on past and future observations (e.g.: POS). In contrast generative models make strict independence assumptions on the observations.

Maximum entropy (ME) models are conditional probabilistic models with above advantages

Weakness of ME models: label bias problem

Introduction

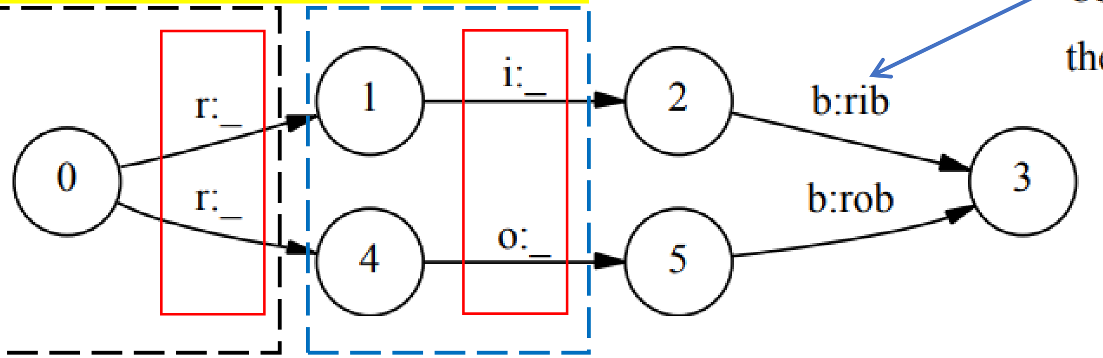
The Label Bias Problem

Simple finite-state model to distinguish between the two words ***rib*** and ***rob***

Suppose that the observation sequence is *r i b*

In the 1st time step, r matches both transitions from start state

Probability mass gets distributed equally among both those two transitions



observation-label pairs $o : l$

the symbol '_' represents the null output label.

Next we observe i.
Both states **1 and 4 have only one outgoing** transition
State 1 has seen this observation in training
State 4 has never seen this observation
But both state 1 and 4 will pass all their mass to their single outgoing transition (conservation of score mass)
Because they do not generate the observation, only conditioning it.
This way, states with single outgoing transition ignore their observations.

The top path and bottom path will be equally likely, independently of the observation sequence.
If one of the two words is more common in the training set , the transitions (out of start state) will prefer its corresponding transition, and that word's state sequence will always win.

CRF has all advantages of MEs and solves label bias problem too

Introduction

Proper solution to label bias problem

Models that account for whole state sequences at once by letting some transitions “vote” more strongly than others depending on the corresponding observations.

Meaning that score mass will not be conserved, but instead individual transitions can “amplify” or “dampen” the mass they receive.

From previous example, transitions from the start state would have a very weak effect on the path score, while the transitions from states 1 and 4 would have much stronger effects.

Conditional Random Fields

Set of all possible states $\mathcal{S}$

Set of all possible state sequences $\mathcal{S}^m$

Need a model of conditional distribution

$$p(s_1 \ldots s_m | x_1 \ldots x_m) = p(\underline{s}|\underline{x})$$

Sequence of states

Input sequence

Idea is to define a **global feature vector** $\underline{\Phi}(\underline{x}, \underline{s}) \in \mathbb{R}^d$

Map entire input sequence paired with an entire state sequence to d-dimensional feature vector

And build a log-linear model

$$\boxed{p(\underline{s}|\underline{x};\underline{w})} = \frac{\exp\left(\underline{w} \cdot \underline{\Phi}(\underline{x}, \underline{s})\right)}{\boxed{\sum_{\underline{s}' \in \mathcal{S}^m} \exp\left(\underline{w} \cdot \underline{\Phi}(\underline{x}, \underline{s}')\right)}}$$

Normalization constant

Conditional probability of s given x, under parameters w

Conditional Random Fields

Global feature vector

$$\underline{\Phi}(\underline{x}, \underline{s}) = \sum_{j=1}^{m} \underline{\phi}(\underline{x}, j, s_{j-1}, s_j)$$

Local feature vector

Sum up the local feature vectors in m different state transitions

Need to look into

- Decoding
- Parameter Estimation

Conditional Random Fields

Decoding

Input sequence → $\underline{x} = x_1, x_2, \ldots x_m$

Find the most likely underlying state sequence under the model → $\arg\max_{\underline{s}\in\mathcal{S}^m} p(\underline{s}|\underline{x};\underline{w})$ → $\arg\max_{\underline{s}\in\mathcal{S}^m} \sum_{j=1}^{m} \underline{w}\cdot\underline{\phi}(\underline{x}, j, s_{j-1}, s_j)$

Each transition from state s_{j-1} to state s_j has an associated score $\underline{w}\cdot\underline{\phi}(\underline{x}, j, s_{j-1}, s_j)$ which calculates plausibility of sequence s

The decoding problem is to find an entire sequence of states such that the sum of transition scores is maximized. Viterbi algorithm is used to find the maximum score

Parameter estimation

A measure that how well w explains the labelled examples. Dynamic programming is used for this estimation. Cont. in next talk.

Applications

Part-of-Speech (PoS) Tagging

Goal: Label a sentence (a sequence of words or tokens) with tags like ADJECTIVE, NOUN, PREPOSITION, VERB, ADVERB, ARTICLE

Given: Bob drank coffee at Starbucks

Labelled: Bob (NOUN) drank (VERB) coffee (NOUN) at (PREPOSITION) Starbucks (NOUN)

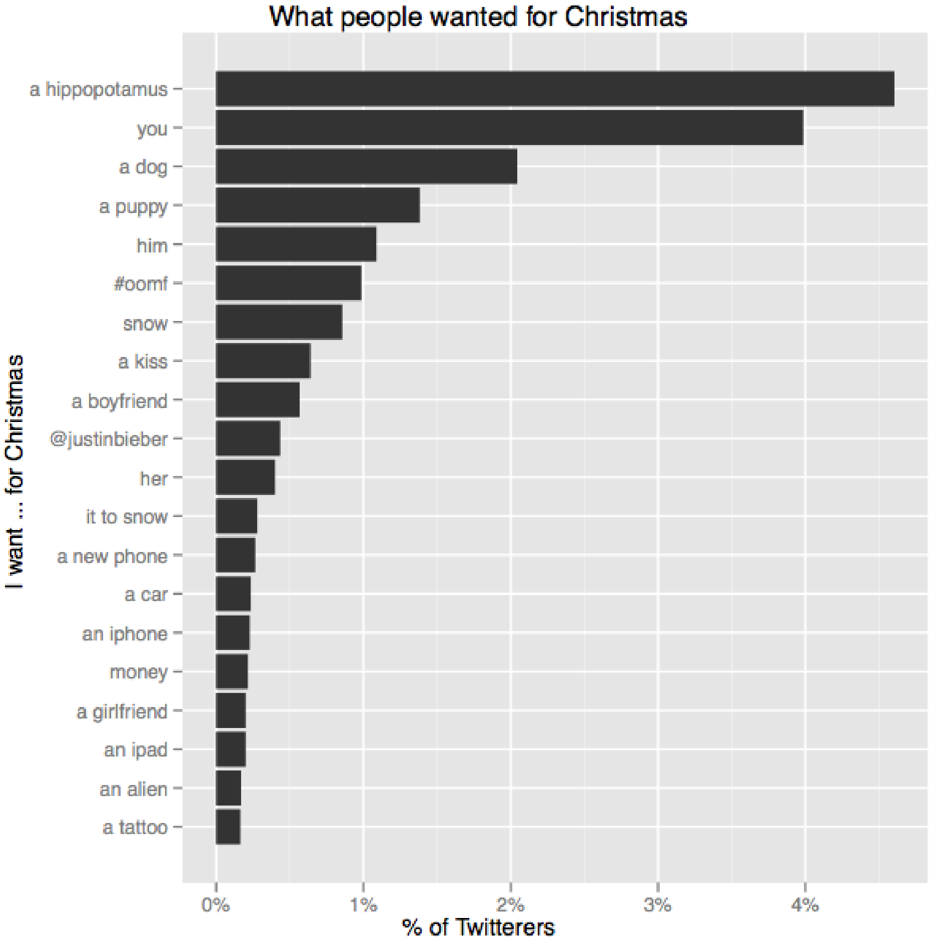
Can be used in a sequence of observed activities to label them

Applications

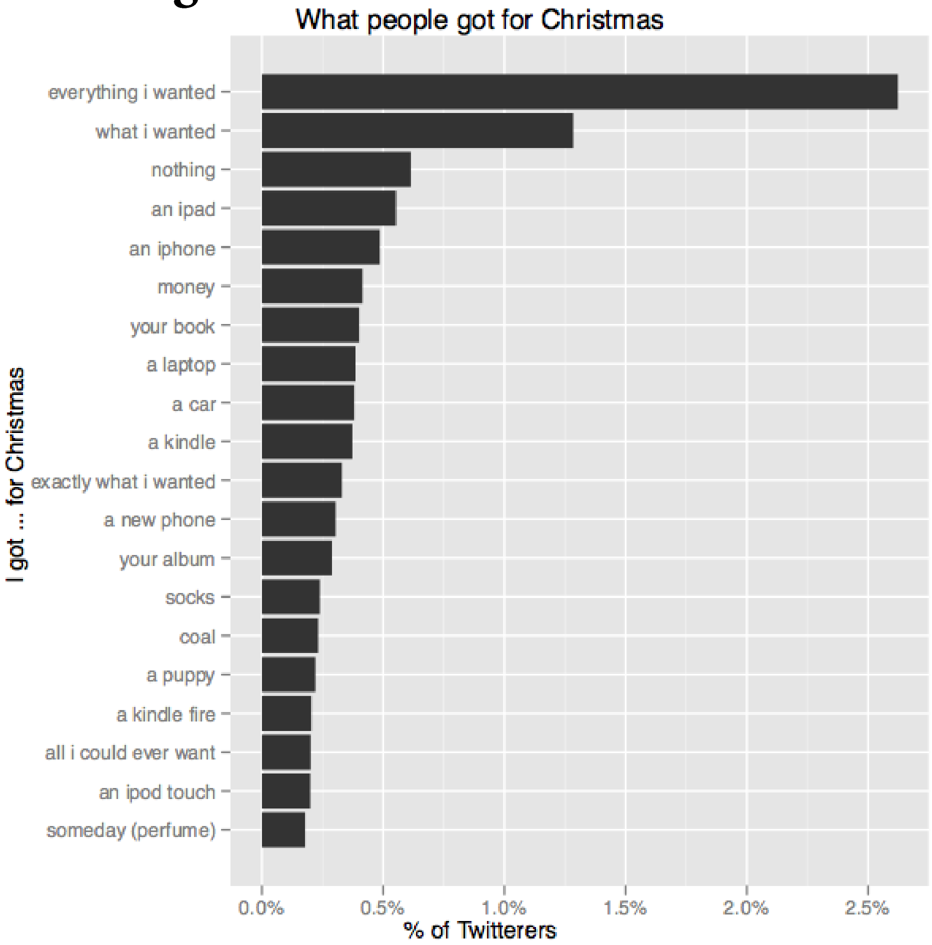
Part-of-Speech (PoS) Tagging

Types of presents people received for Christmas

“I want --- for Christmas”



“I got --- for Christmas”



References

http://www.eng.utah.edu/~cs6961/papers/klinger-crf-intro.pdf

http://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers

http://blog.echen.me/2012/01/03/introduction-to-conditional-random-fields