Learning Algorithms, Project 2

Mohsen Malmir, Erfan Sayyari

February 11, 2014

1 Abstract

2 Introduction

In this paper, we provide the results and details of implementation of a Conditional Random Field (CRF) model for predicting English language text punctuations. CRFs are a variant of undirected graphical models that are well suited for predicting structured labels. We train our model by maximizing the log conditional likelihood of the training data. During our experiments, we found that the model could predict individual tags by more than 94% accuracy. However, as we performed more and more experiments, we discovered the over fitting to some prevalent tags in the training set. Using several experiments and methods that will be described later in this paper, we could overcome the over-fitting problem to some degrees.

3 Design and Analysis of Algorithm

3.1 The general log-linear model

3.2 Feature functions

3.2.1 Part of Speech tagging

3.3 Conditional random fields

3.4 The Collins perceptron

3.5 Contrastive divergence

3.5.1 Gibbs sampling

4 Design of Experiments

4.1 Initialization

4.2 Preprocessing

4.3 Performance Measure

5 Results of Experiments

5.1 Collins perceptron

5.2 Gibbs sampling

6 Findings and Lessons Learned

6.1 Numerical Issues and Preprocessing

6.2 Overfitting

6.3 Model Selection

6.4 Future works