Part-of-Speech Tagging

+

Neural Networks

CS 287

Quiz: ReLU

Last class we focused on standard hinge loss. Consider now the squared hinge loss,

$$L_{hinge} = \max\{0, 1 - (\hat{y}_c - \hat{y}_{c'})^2\}$$

What is the effect does this have on the loss? How do the parameters gradients change?

Contents

Penn Treebank

Hi! I am the ptb.

Penn Treebank

Statistics

Parse Tree

Dataset: Penn Treebank

Penn Treebank,

- Central dataset in NLP.
- 1M word tokens, collected from Wall Street Journal.
- Annotated with syntactic structure.

Shared Tasks

Tagset

Pass out examples

Linguistically

Why are tags important useful.

Tagging

How hard is this task?

rare words.

Tag Features: Word Properties

Representation can use specific aspects of text.

- $\mathcal{F}$; Spelling, all-capitals, trigger words, etc.
- $\mathbf{x} = \sum_i \delta(f_i)$

Example: Spam Email

`Your diploma puts a UUNIVERSITY JOB PLACEMENT COUNSELOR at your disposal.`

$$\mathbf{x} = v(\texttt{misspelling}) + v(\texttt{allcapital}) + v(\texttt{trigger:diploma}) + \ldots$$

$$\mathbf{x}^\top = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} = \begin{array}{c} \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} \end{array} \begin{array}{l} \texttt{misspelling} \\ \vdots \\ \texttt{capital} \\ \texttt{word:diploma} \end{array}$$

Features used in state of the art

What if we just used words and context?

Contents

Sentence Tagging

- $w_1, \ldots, w_n$; sentence words
- $t_1, \ldots, t_n$; sentence tags
- $\mathcal{C}$; output class, set of tags.

Window Model

Goal: predict t_5.

- Windowed word model.

$$w_1 \, w_2 \left[w_3 \, w_4 \, w_5 \, w_6 \, w_7 \right] w_8$$

- w_3, w_4; left context
- w_6, w_7; right context

Boundary Cases

Goal: predict t_2.

$$\left[< s > w_1 w_2 w_3 w_4 \right] w_5 w_6 w_7 w_8$$

Goal: predict t_8.

$$w_1 w_2 w_3 w_4 w_5 \left[w_6 w_7 w_8 < /s >< /s > \right]$$

k Symbols $< s >$ and $< /s >$ represent boundary padding.

The Role of Features

- Recall Zipf's law.
- Many words are ..
- Can capture patterns. example.

How much does this matter?

graph of tagging.

Sparse Tagging Model

- Create training data,

$$(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$$

- Each $\mathbf{x}_i$ includes features of window.
- Each $\mathbf{y}_i$ is the one-hot tag encoding.
- Prediction accuracy is measured identically.

Naive Bayes/Logistic Regression for Tagging

- Setup is identical to text classification.
- $$\hat{\mathbf{y}} = \mathbf{x}\mathbf{W} + \mathbf{b}$$

Contents

Collobert and Weston Natural Language Processing (almost) from Scratch

Two ideas

- Non-linear Models
- Dense Word embeddings

(1) Non-Linear Models for Classification

- Neural network represent any non-linear classifier, for example

$$NN_1 = f_1(\mathbf{x}\mathbf{W}^1 + \mathbf{b}^1))$$

$$\hat{\mathbf{y}} = f_2(NN_1\mathbf{W}^2 + \mathbf{b}^2)$$

- Where $\mathbf{W}^1 \in \mathbb{R}^{d_{\text{in}} \times dmid}$, $\mathbf{b}^1 \in \mathbb{R}^{1 \times dmid}$
- $\mathbf{W}^2 \in \mathbb{R}^{dmid \times dout}$, $\mathbf{b}^2 \in \mathbb{R}^{1 \times d_{\text{out}}}$
- Activation f_1 is non-linear.

Decision $\arg\max \hat{y}$

Can learn non-linear decision boundary. Diagram

For instance, f_1 Sigmoid and f_2 softmax

$$\frac{\partial L(y, \hat{y})}{\partial \hat{y}_j} = \frac{\mathbf{1}(y_j = 1)}{\hat{y}_j}$$

For instance, f_1 ReLU and $f_2$hinge-loss

Backpropagation

- Chain rule

Contents

(2) Dense Features

Instead of defining $\mathbf{x} = \sum_{i=1}^{n} \delta(f_i)$

Where $v : \mathcal{F} \mapsto \mathbb{R}^d$ for instance $v(f) = \delta(f)\mathbf{W}^0$

and define $\mathbf{x} = [v(f_1) \ldots v(f_k)]$

(For now we assume all examples have fixed length)

Dense Features for Tagging

Instead of defining $\mathbf{x} = \sum_{i=1}^{n} \delta(f_i)$

Where $v : \mathcal{F} \mapsto \mathbb{R}^d$ for instance $v(f) = \delta(f)\mathbf{W}^0$

and define $\mathbf{x} = [v(f_1) \ldots v(f_k)]$

(For now we assume all examples have fixed length)

Dense Features for Tagging

Instead of defining $\mathbf{x} = \sum_{i=1}^{n} \delta(f_i)$

Where $v : \mathcal{F} \mapsto \mathbb{R}^d$ for instance $v(f) = \delta(f)\mathbf{W}^0$

and define $\mathbf{x} = [v^1(f_1) \ldots v^1(f_k) \ldots v^2(f_k + 1) \ldots v^2(f_k)]$

(For now we assume all examples have fixed length)

Parameters

- With word features $|\mathcal{V}|$
- With all pair word features $|\mathcal{V}|^2$
- With word embedding features $d|\mathcal{V}|$

  Representation that allows parameter sharing.

Lookup layer is Learned too

results

Results Pretty good

objective

Diagram