Conditional Random Field

Zhiliang Tian

2016.04.30

1 Introduction

(TODO)

2 the calculate of Probability

2.1 unigram score and unigram probability

$$U_k = \sum_{i}^{dim} Wu_{ik} * x_i \tag{1}$$

$$Pu_k = \frac{\exp(U_k)}{\sum_{k=0}^{K} \exp(U_k)} \tag{2}$$

$$Pu_k = softmax(U_k) \tag{3}$$

x_i is the i-th dimension of input feature $\mathbf{x}$. U_k is the unigram score of k-th class, when input feature is $\mathbf{x}$. We assume there are K classes. Wu is the weight matrix between input feature and unigram score.Wu_{ik} is the connection between i-th feature and k-th class. Equation 2 is equal to Equation 3.

If CRF only use unigram feature and unigram weight, Pu is the final probability. And we can see that is Maximum Entropy.

2.2 bigram score and bigram probability

$$B_{kk'} = \sum_{i}^{dim} Wb_{ikk'} * x_i \tag{4}$$

$$Pb_{kk'} = \frac{\exp(B_{kk'})}{\sum_{k=0}^{K} \sum_{k'=0}^{K} \exp(B_{kk'})} \tag{5}$$

$$Pb_{kk'} = softmax(B_{kk'}) \tag{6}$$

x_i is the i-th dimension of input feature $\mathbf{x}$. $B_{kk'}$ is the bigram score when t-th time, tag is k-th class, when $t-1$ time, tag is k'-th class, and input feature is $\mathbf{x}$. We assume there are K classes. Wu is the weight matrix between input feature and unigram score.Wu_{ik} is the connection between i-th feature and k-th class. Equation 2 is equal to Equation 3.

2.3 probability in sequence

Firstly, we should combine unigram score and bigram score(why do that here?). Given previous state and the input of this time, we can represent this time's state.

$$\varphi_{kk'} = \exp(U_k) * \exp(B_{kk'}) = \exp(U_k + B_{kk'}) \tag{7}$$

$$S(t)_k = \sum_{k'}^{K} S(t-1)_{k'} * \varphi_{kk'} \tag{8}$$

$\varphi_{kk'}$ is the non-normalized probability of transforming last time k' to this time k. $S(t)_k$ is the non-normalized probability, means that from 0 to $t-1$ every time keep it class, and t-th time is k-class.

Secondly, add normalization term.

$$\begin{array}{l} P(t)_k = \frac{S(t)_k}{\sum_a^K S(t)_a} \\ S(0)_k = \exp(U_k) \end{array} \tag{9}$$

Finally, probability:

$$P(t)_k = \frac{\sum_{k'}^{K}(S(t-1)_{k'} * \exp(\sum_i^{dim}(Wu_{ik} * x_i) + \sum_i^{dim}(Wb_{ikk'} * x_i))}{\sum_k^K \sum_{k'}^{K}(S(t-1)_{k'} * \exp(\sum_i^{dim}(Wu_{ik} * x_i) + \sum_i^{dim}(Wb_{ikk'} * x_i))} \tag{10}$$

$$S(0)_k = \exp(\sum_i^{dim}(Wu_{ik} * x_i)) \tag{11}$$

$P_(t)k$ is the probability on k-th class at t-th time and previous time keep it state.

3 loss and gradient, not using bptt

3.1 introduction

Back propagation through time called bptt. It means now time's gradient will contribute on previous time. At this section, we ignore bptt. Our taget: $\frac{\alpha J}{\alpha W}$ and $\frac{\alpha J}{\alpha \mathbf{X}}$.

3.2 parameter estimation and loss function

For CRF model, we use log maximum likelihood to estimate unknown parameter.

$$L = \prod_{m}^{dataset} P = \prod_{m}^{dataset} \prod_{t=0}^{T} \prod_{k=0}^{K} (P_{mtk})^{y(mtk)} \tag{12}$$

if label is k, $y(k) = 1$; else, $y(k) = 0$. As ignoring bptt, we need only maximize every time's probability independently in a sequence.

$$J = log(L) = \sum_{m}^{dataset} \sum_{t=0}^{T} \sum_{k=0}^{K} (y(mtk) * P_{mtk}) \tag{13}$$

We introduce a SGD-based solution on CRF model. So, we need get the **loss** on a single sample. As ignoring bptt, we can treat a time as a sample imprecisely. (A sequence should indeed be a sample in CRF)

$$J = \sum_{k=0}^{K} (y(k)) * log(P_k)) \tag{14}$$

3.3 the gradient of unigram score

Our target: $\frac{\alpha(J)}{\alpha(U_k)}$. For a time of a sequence, we ignore the time t for convenient.

$$\frac{\alpha(J)}{\alpha(U_k)} = \frac{\alpha(J)}{\alpha(S_k)} * \frac{\alpha(S_k)}{\alpha(U_k)} = \sum_{a}^{K} (\frac{\alpha(J)}{\alpha(P_a)} * \frac{\alpha(P_a)}{\alpha(S_k)}) * \sum_{k'}^{K} (\frac{\alpha(S_k)}{\alpha(\varphi_{kk'})} * \frac{\alpha(\varphi_{kk'})}{\alpha(U_k)}) \tag{15}$$

$$\frac{\alpha(J)}{\alpha(P_a)} = \frac{y(a)}{P_a} \tag{16}$$

$$\frac{\alpha(P_a)}{\alpha(S_k)} = \frac{\alpha(\frac{S_a}{\sum_b (S_b)})}{\alpha(S_k)} = \frac{\alpha(\frac{S_a}{C_k + S_k})}{\alpha(S_k)} \tag{17}$$

Let $C_k = \sum_{b \neq k}^{K}(S_b)$, C_k is a constant variable respect to S_k. We divide Equation(17) into $a = k$ and $a \neq k$.

$$\frac{\alpha(P_a)}{\alpha(S_k)} = \begin{cases} \frac{\alpha(\frac{S_k}{C_k+S_k})}{\alpha(S_k)} = \frac{C_k}{(C_k+S_k)^2} = \frac{1-P_k}{\sum_b^K S_b} & a = k \\ \frac{\alpha(\frac{S_a}{C_k+S_k})}{\alpha(S_k)} = -\frac{S_a}{(C_k+S_k)^2} = -\frac{P_a}{\sum_b^K S_b} & a \neq k \end{cases} \tag{18}$$

$$\frac{\alpha(S(t)_k)}{\alpha(\varphi_{kk'})} = S(t-1)_{k'} \tag{19}$$

$$\frac{\alpha(\varphi_{kk'})}{\alpha U_k} = \exp(U_k + B_{kk'}) = \varphi_{kk'} \tag{20}$$

$$\frac{\alpha(\varphi_{kk'})}{\alpha B_k k'} = \exp(U_k + B_{kk'}) = \varphi_{kk'} \tag{21}$$

Then, merge equations above according to Equation(15), but we'd better calculate $\frac{\alpha(J)}{\alpha(U_k)}$ firstly. Let $Z(t) = \sum_a^K S(t)_a$, Z(t) is the normalization term at t-th time.

$$\begin{aligned} \frac{\alpha(J)}{\alpha(U_k)} &= \textstyle\sum_a^K (\frac{\alpha(J)}{\alpha(P_a)} * \frac{\alpha(P_a)}{\alpha(S_k)}) * \sum_{k'}^K (\frac{\alpha(S_k)}{\alpha(\varphi_{kk'})} * \frac{\alpha(\varphi_{kk'})}{\alpha(U_k)}) \\ &= (\frac{y(k)}{P_k}\frac{1-P_k}{Z(t)} + \textstyle\sum_{a \neq k} K \frac{y(a)}{P_a}\frac{-P_a}{Z(t)}) * (\sum(S(t-1)_{k'} * \varphi_{kk'})) \end{aligned} \tag{22}$$

Note that we can utilize $\sum_{a \neq k}^K y(a) = 1 - y(k)$, $P_k = \frac{S_k}{Z(t)}$. At last, Equation(22):

$$\frac{\alpha(J)}{\alpha(U_k)} = y(k) - P_k \tag{23}$$

3.4 the gradient of bigram score

Our target: $\frac{\alpha(J)}{\alpha(B_{kk'})}$

$$\frac{\alpha(J)}{\alpha(B_{kk'})} = \frac{\alpha(J)}{\alpha(S_{kk'})} * \frac{\alpha(S_{kk'})}{\alpha(B_{kk'})} = \sum_a^K (\frac{\alpha(J)}{\alpha(P_{ak'})} * \frac{\alpha(P_{ak'})}{\alpha(S_{kk'})}) * (\frac{\alpha(S_{kk'})}{\alpha(\varphi_{kk'})} * \frac{\alpha(\varphi_{kk'})}{\alpha(B_{kk'})}) \tag{24}$$

Note that $S_{kk'} = S(t)_{kk'} = S(t-1)_{k'} * \varphi_{kk'}$

$$\frac{\alpha(J)}{\alpha(P_{ak'})} = \frac{y(a)}{P_{ak'}} \tag{25}$$

$$\frac{\alpha(P_{ak'})}{\alpha(S_{kk'})} = \frac{\alpha(\frac{S_{ak'}}{\sum_b(S_{bk'})})}{\alpha(S_{kk'})} = \frac{\alpha(\frac{S_{ak'}}{C_{kk'}+S_{kk'}})}{\alpha(S_{kk'})} \tag{26}$$

Let $C_{kk'} = \sum_{b \neq k}^{K}(S_{bk'})$, $C_{kk'}$ is a constant variable respect to $S_{kk'}$. We divide into $a = k$ and $a \neq k$.

$$\frac{\alpha(P_{ak'})}{\alpha(S_{kk'})} = \begin{cases} \frac{\alpha(\frac{S_{kk'}}{C_{kk'}+S_{kk'}})}{\alpha(S_{kk'})} = \frac{C_{kk'}}{(C_{kk'}+S_{kk'})^2} = \frac{1-P_{kk'}}{\sum_b^K S_{bk'}} & a = k \\ \frac{\alpha(\frac{S_{ak'}}{C_{kk'}+S_{kk'}})}{\alpha(S_{kk'})} = -\frac{S_{ak'}}{(C_{kk'}+S_{kk'})^2} = -\frac{P_{ak'}}{\sum_b^K S_{bk'}} & a \neq k \end{cases} \tag{27}$$

$$\frac{\alpha(S_{kk'})}{\alpha(\varphi_{kk'})} * \frac{\alpha(\varphi_{kk'})}{\alpha(B_{kk'})} = S(t-1)_{k'} * \exp(U_k + B_{kk'}) = S(t-1)_{k'} * \varphi_{kk'} \tag{28}$$

Then, merge them:

$$\begin{aligned} &\tfrac{\alpha(J)}{\alpha(B_{kk'})} = (\tfrac{y(k)}{P_{kk'}} \tfrac{1-P_{kk'}}{Z(t)_{k'}} + \textstyle\sum_{a \neq k}^{K} \tfrac{y(a)}{P_{ak'}} \tfrac{-P_{ak'}}{Z(t)_{k'}}) * (S(t-1)_{k'} * \varphi_{kk'}) \\ &= \tfrac{S(t-1)_{k'} * \varphi_{kk'}}{Z(t)_{k'}} * (\tfrac{y(k)(1-P_{kk'})}{P_{kk'}} - (1 - y(k))) \\ &= \tfrac{S(t)kk'}{Z(t)_{k'}} * \tfrac{(y(k)-P_{kk'})}{P_{kk'}} \\ &= y(k) - P_{kk'} \end{aligned} \tag{29}$$

$S(t)_{kk'}$ is the non-normalized probability, means that from 0 to $t-2$ every time keep it class, t-th time is k-class, $t-1$-th time, it is a joint probability of t-th time and $t-1$-th time.

3.5 the gradient of weight and feature

Our final target is : $\frac{\alpha(J)}{\alpha(Wu_{ik})}$, $\frac{\alpha(J)}{\alpha(Wb_{ikk'})}$ and$\frac{\alpha(J)}{\alpha(x_i)}$. We can get easily :

$$\frac{\alpha(J)}{\alpha(Wu_{ik})} = \frac{\alpha(J)}{\alpha(U_k)} * \frac{\alpha(B_k)}{\alpha(Wu_{ik})} = (y(k) - P_k) * x_i \tag{30}$$

$$\frac{\alpha(J)}{\alpha(Wb_{ikk'})} = \frac{\alpha(J)}{\alpha(B_{kk'})} * \frac{\alpha(B_{kk'})}{\alpha(Wb_{ikk'})} = (y(k) - P_{kk'}) * x_i \tag{31}$$

In equation 33, we calculate x_i's gradient by unigram and bigram score, and "adding" them together. We use "multiple composite function derivation rule"(equation 32) here. For $z = f(x, y)$, $x = g(t)$, $y = h(t)$

$$\frac{\alpha z}{\alpha t} = \frac{\alpha z}{\alpha x}\frac{\alpha x}{\alpha t} + \frac{\alpha z}{\alpha y}\frac{\alpha y}{\alpha t} \tag{32}$$

$$\begin{aligned} &\frac{\alpha(J)}{\alpha(x_i)} = \sum_k^K \frac{\alpha(J)}{\alpha(U_k)} * \frac{\alpha(U_k)}{\alpha(x_i)} + \sum_k^K \sum_{k'}^K \frac{\alpha(J)}{\alpha(B_{kk'})} * \frac{\alpha(B_{kk'})}{\alpha(x_i)} = \\ &\sum_k^K ((y(k) - P_k) * W_{ik}) + \sum_k^K \sum_k^K ((y(k) - P_{kk'}) * W_{ikk'}) \end{aligned} \tag{33}$$

As we find a interesting thing in softmax, from final gradient, we can conclude that the gradient of CRF also is the different of expectation of observation and estimation.

4 some further discuss

4.1 Why give up tranditional softmax with loss and divide into unigram and bigram?

We use Equation 15 and 24 to calculate unigram and bigram gradient separately. But We have already know the gradient of "softmax with loss", why not using it?

Original softmax with loss just like equation 15, S_k only has connection with P_a, but in equation 24 $S_{kk'}$ has connection with $P_{ak'}$. Original softmax with loss can not connect with both P_a and $P_{ak'}$. So, we can not combine them to use original softmax with loss.

4.2 Why can unigram and bigram be merged for x_i's gradient

Why do we add the gradient from unigram score and bigram score when calculating x_i's gradient? Unigram score and bigram score is independent. Calculating x_i' gradient the should "adding" them by "multiple composite function derivation rule".

4.3 The Expectation of observation and estimation

As same as "softmax with loss", from equation 23 29, we can also find the same conclusion. For unigram score, we hope expectation of observation and $P_unigram$'s estimation nearly, for bigram score, we hope expectation of observation and P_bigram's estimation nearly. If they are totally equal, gradient is zero, we find the solution.

4.4 Why bigram using $S_{kk'}$

Equation 24 we just calculate gradient from $S_{kk'}$ and $P_{ak'}$ $a \in [0, K]$, but not S_{kb} and P_{ab} $a \in [0, K] b \in [0, K]$. Why?

CRF's output sequence is a linear chain Markov random field. It said that time t only can be influenced by time $t-1$ and $t+1$. That thought is worked by transform probability between $t-1$ and t, t and $t+1$, which are conditional probability, $P(t=k|t-1=k')$. So, we see the probability $P(t-1=k')$ is a constant. And the probability $P(t-1=k')$ will be update at $t-1$ time.

5 bi-directional

(TODO)

6 reference

https://zh.wikipedia.org/wiki/