Lecture 4: Word window classification and NN

2018年7月5日 星期四　　上午11:27

Simple Softmax classifier

1.Parameter

- X :concatenated context word vectors
- W:center word vectors

… museums in Paris are amazing … .

$$X_{window} = [\ x_{museums} \quad x_{in} \quad x_{Paris} \quad x_{are} \quad x_{amazing}\ ]^T$$

2.Prediction

Details of the softmax classifier

$$p(y|x) = \frac{\exp(W_{y\cdot}x)}{\sum_{c=1}^{C}\exp(W_{c\cdot}x)}$$

We can tease apart the prediction function into two steps:

1. Take the y'th row of W and multiply that row with x:

$$W_{y\cdot}x = \sum_{i=1}^{d} W_{yi}x_i = f_y$$

Compute all f_c for c=1,…,C

2. Apply softmax function to get normalized probability:

$$p(y|x) = \frac{\exp(f_y)}{\sum_{c=1}^{C}\exp(f_c)} = softmax(f)_y$$

6　　1/18/18

3.Training:

- Loss function:Softmax—Cross-Entropy

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{f_{y_i}}}{\sum_{c=1}^{C} e^{f_c}}\right) = \frac{1}{N}\sum_{y} -\log P(y|x)$$

$$H(p|q) = H(\underset{\text{(one hot)}}{p}) + D_{KL}(p\|q)$$
$$= D_{KL}(p\|q)$$
$$= \sum_{c=1}^{C} p(c)\log\frac{p(c)}{q(c)}$$

?

形式上与J(θ)并不一样

如何理解J(θ)即为cross Entropy？

- Gradient:

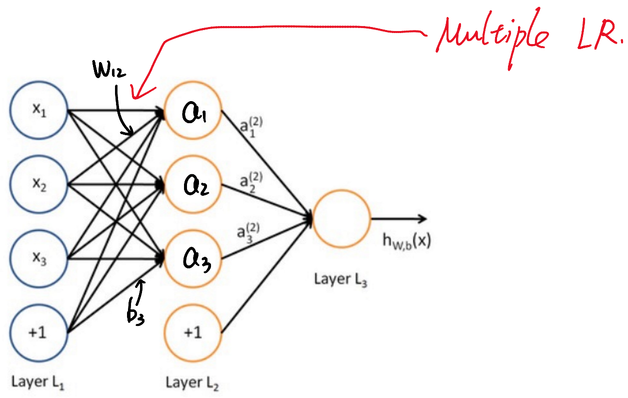
$$\nabla_\theta J(\theta) = \begin{bmatrix} \nabla_{W_{\cdot 1}} \\ \vdots \\ \nabla_{W_{\cdot d}} \\ \nabla_{x_{aardvark}} \\ \vdots \\ \nabla_{x_{zebra}} \end{bmatrix} \in \mathbb{R}^{Cd+Vd}$$

- SGD update

NN with max-margin

1.Neural Network structure

- Notations:
  - $z = Wx + b$
  - $a = f(z)$ Activation function



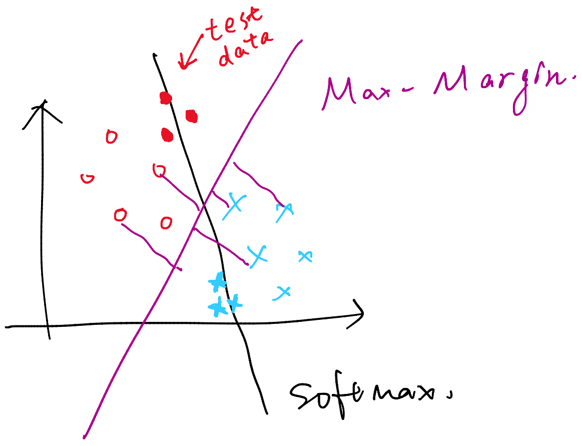
2.Max-Margin loss

Make score of true window larger and corrupt windows lower.

$S = U^T f(Wx + b)$

$a = f(Wx + b)$

$z = Wx + b$



3.Training with BP

buffer Area

- **Loss function:**

$$minimize J = \max(\Delta + s_c - s, 0)$$

Neg. case.

- **Back propagate:**
  - For Weight Matrix W

    U_i only appears with a_i

    $$\frac{\partial}{\partial (W_{ij})} U^T a \rightarrow \frac{\partial}{\partial (W_{ij})} U_i a_i$$

    $$\frac{\partial (s)}{\partial (W_{ij})} = U_i f'(z_i) x_i = \delta_i x_j$$

    Matrix form: Outer product

    $$\frac{\partial (s)}{\partial (W)} = \delta x^T$$

  - For biases b
    - $U_i \frac{\partial}{\partial b_i} a_i = U_i f'(z_i) = \delta_i$

      $$\frac{\partial (s)}{\partial (b)} = \delta$$