DS-GA 1008 Homework 3

Zian Jiang (zj444)

October 29, 2020

Q1

a

For a general case at one time step, $y = [0, ..., 1, ..., 0], \bar{y} = [\bar{y}_1, ..., \bar{y}_L]$ and suppose the correct class is l, then

$$NLLLoss(y, \bar{y}) = -\sum_{i=1}^{L} y_i \log(\bar{y}_i) = -\log(\bar{y}_l)$$

since every but one y_i in the sum is not 0. Now, for the y and $\bar{y}$ in this question, y will be a $(T,)$, and $\bar{y}$ will be (T, L). Suppose $y = [y_1, y_2, ..., y_T]$, so for example the correct class at time step i is $y_i \in \{1, 2, ..., L\}$. Then

$$NLLLoss(y, \bar{y}) = \frac{1}{T} \sum_{t=1}^{T} -\log(\bar{y}_{t,y_t}),$$

where $\bar{y}_{t,y_t} = P(y = y_t | x_t)$.

b

In language modeling, the goal is next to predict the next word given the previous words (everything towards the left). Thus if we use a bidirectional structure, the information of next word will be already coming from the right to left direction so training loss will just decrease to 0. On the other hand, for POS tagging task, context from both left to right and right to left can be helpful. Since our predicting target is not related to the input data, using bidirectional structure is fine.

Q2

Pros:

- It works when there is an infinite amount of suitable outputs so that using a softmax layer is not possible (predicting future video frames).

- It works when inference is more complex than vector mappings with an explicit function.

Cons:

- The line search process to the find the optimal y in $\breve{y} = \arg\min_y E(x, y)$ can be expensive or stuck at a local minimum if initialized poorly.

Q3

a

$$\breve{y} = \arg\min_y E(x, y),$$

which means finding the optimal sequence out of of 4^3 POS tags sequences that fits "the best" with input sentence x, or having the lowest energy. Thus we will need to look up E $4^3 = 64$ times in order to find the entry with the lowest energy.

b

i

There will be $50000^{15} \times 20^{15}$ many (x, y) pairs.

ii

20^{15} times.

Q4

a

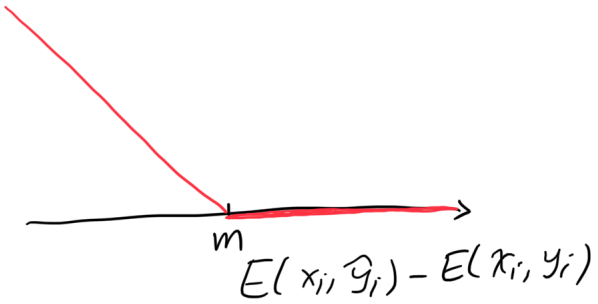
$E_\theta(x_i, y_i)$ is the energy at data points x_i and its gold standard output.
$\min_{y \neq y_i} E_\theta(x_i, y)$ is the worst incorrect output of x_i out of all the non-gold standard output.

b

The function aims to push down the energy of the correct answer (x_i, y_i) and push up the energy of the worst incorrect answer $\arg\min_{y \neq y_i} E_\theta(x_i, y)$. The margin m prevents the energy surface from being too flat since this implies $E_\theta(x_i, y_i) + m \leq \min_{y \neq y_i} E_\theta(x_i, y)$ instead of $E_\theta(x_i, y_i) \leq \min_{y \neq y_i} E_\theta(x_i, y)$. Then, $[\cdot]_+ = \max(0, \cdot)$ makes it a hinge loss so that there is no penalization if $E_\theta(x_i, y_i) + m \leq \min_{y \neq y_i} E_\theta(x_i, y)$ is satisfied or it gets penalized in a linear fashion. Thus whenever the difference between the energy of the incorrect answer with the lowest energy and the energy of the desired answer is less than the margin, the learning procedure should make that difference larger.

c

Let $x = E_\theta(x_i, \hat{y}_i) - E_\theta(x_i, y_i)$ and we want to draw $[-m+x]_+ = \max(0, -m+x)$.



d

$$\hat{y}_t \leftarrow \hat{y}_{t-1} - \eta \frac{\partial E_\theta(x, y)}{\partial y}$$

e

i

Precision for Verb, since Verb is the most frequent out of all tags.

ii

Out of every y from the output space, we want to find the energy of worst non-desired output y such that $E_\theta(x_i, y_i) - E_\theta(x_i, y)$ is the farthest away from margin $\Delta(y, y_i)$.

iii

While both functions are penalizing in a linear function given the distance of the error to a margin, the difference lies in the choice of margin. In function 1, the margin is m for all examples. On the other hand, in function 2 the margin is a different function of the worst incorrect output y, $\Delta(y, y_i)$ for each example x_i.

This flexibility in margin may allow the learning procedure to be more efficient because we are "customizing" the margin for each example so that the learning potential can be maximized. More importantly, the energy function may be flatter and smoother for function 2 because $\Delta(y, y_i)$ might be small.

Q5

a

Constrastive: pushing down energy at data points, pushing up everywhere else
Regularized: regularizing the volume of the low energy regions by using a regularization term
Architectural: bounding the volume of the low energy regions

b

$$E(y, z) = \|y - \texttt{Dec}(z)\|_2^2 + \lambda \|z\|_1$$

where λ is the regularization term and y is the input.

c

The energy function will have just a $L2$ norm term and now the latent variable z becomes too expressive so that for every input image y, there always will be a perfectly reconstructed image $\texttt{Dec}(z)$ and energy will be 0 everywhere because $\min \|y - \texttt{Dec}(z)\|_2^2$ is a convex optimization problem.

d

The regularizer limits the information capacity of the latent variable z and prevents the objective function from overfitting, or finding a solution where the energy is 0. Well-reconstructed images will have low energy and badly-reconstructed images will have high energy.

e

Given k centroids and d dimensions,

$$E(y, z) = \|y - wz\|_2^2$$

where z is an one-hot vector of k dimension, w is a matrix of size (d, k), and $y \in \mathbb{R}^d$. K-means is architectural because z is constrained to be an one-hot vector whose role is to select the corresponding column in w and it can take on only k discrete values.