Assignment 3: Energy-based Models

Shuang Gao

October 2020

1 Problem 1: feed-forward models

(a) The negative log-likelihood is

$$NLL = -\sum_{i=1}^{T} \tilde{y}_i[y_i]$$

where $\tilde{y}_i[y_i]$ is the $y_i th$ entry in vector $\tilde{y}_i$.

(b) Because in the task of POS tagging, we use both forward and backward contextual information for inference the tag of an associate word.

But in language modeling, we hope to predict the next word based on left part of a sentence, so a BiLSTM would introduce a data leakage issue.

2 Problem 2

Pros:

- EBM can be used for complex inference, where the problem requires a complex computation to produce its output. For example, the maximum likelihood in graphical models.
- EBM can be used to produce multi-modal outputs.

Cons:

- EBM is often more computationally expensive than feed-forward approaches when doing inference, so its inference speed is slow.

3 Problem 3

(a) Given x, the equation to obtain y is

$$y = \arg\min_{y_i} \ E(x, y_i)$$

y_i corresponds to a sequence of POS tags, and i is an integer between 1 and 64. For a given sentence, we look up for the possible sequence of tags to minimize energy of the given sentence and this tags sequence candidate. In the 27×64 table, given sentence x, we look up 64 times for the minimum value and its corresponding y in the row associated with sentence x.

(b)

(i) $50000^{15} \times 20^{15}$

(ii) We need to look up for 20^{15} times.

4 Problem 4

(a) Given energy function E_θ, $E_\theta(x_i, y_i)$ is the energy value of given example x_i and its gold-standard tags y_i; $\min_{y \neq y_i} E_\theta(x_i, y)$ is the free energy, which is the minimum energy value of given example x_i and other candidates of tags sequence except its gold-standard tags y_i.

(b) We want to minimize such a function because we want to push down the energy of given x_i and its gold-standard tags y_i, and push up elsewhere when the input pairs are negative samples.

Margin m is used to make the difference between $\min_{y \neq y_i} E_\theta(x_i, y)$ and $E_\theta(x_i, y_i)$ at least m.

$[\cdot]_+$ is used to stop minimizing when the $\min_{y \neq y_i} E_\theta(x_i, y) - E_\theta(x_i, y_i)$ already achieved the margin value m.

(c) $\hat{y}_i = \arg\min_{y \neq y_i} E_\theta(x_i, y)$, so $E_\theta(x_i, \hat{y}_i) = \min_{y \neq y_i} E_\theta(x_i, y)$.

$$[m + E_\theta(x_i, y_i) - \min_{y \neq y_i} E_\theta(x_i, y)]_+ = [m - (E_\theta(x_i, \hat{y}_i) - E_\theta(x_i, y_i))]_+$$

So the loss of instance i is a weakly decreasing function of $(E_\theta(x_i, \hat{y}_i) - E_\theta(x_i, y_i))$. For fixed m, when $(E_\theta(x_i, \hat{y}_i) - E_\theta(x_i, y_i)) \geq m$, loss is always zero; when $(E_\theta(x_i, \hat{y}_i) - E_\theta(x_i, y_i)) < m$, we update the parameters to make $(E_\theta(x_i, \hat{y}_i) - E_\theta(x_i, y_i))$ larger in order to decrease loss.

The plot is Figure 1 on next page.

(d)

$$y^{(t+1)} = y^{(t)} - \eta \frac{\partial E_\theta(x_i, y^{(t)})}{\partial y^{(t)}}$$

(e)

(i) We can define different scores for different types of mismatches, and use the sum of all mismatch scores in the same sentence as the distance metric.

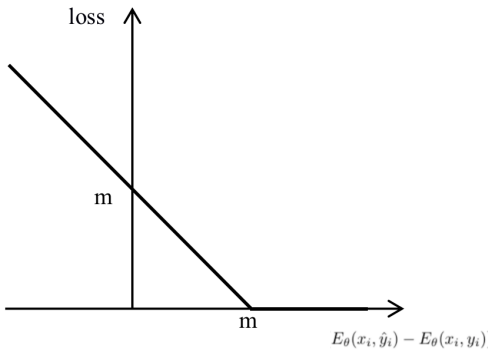

Figure 1: Problem 4(c)

In Table 1, there are 5 categories of tags ('Nominal, Nominal + Verbal', 'Other open-class words', 'Other closed-class words', 'Twitter/online-specific', 'Miscellaneous'). We can define the mismatch score of two different tags as 1 if they belong to different category, and as 0.5 if they belong to the same category.

(ii) The maximization-step picks y that is distant from y_i but has low energy.

(iii) The objective (2) can make the energy values vary along with the distance between the input tag sequence and the gold-standard tag sequence. Therefore, the more dissimilar the input y is to y_i, the higher energy it will have.

5 Problem 5

(a)

- Contrastive methods: push down the energy of correct samples, and simultaneously push up the energy of negative samples.
- Architectural methods: limit the information capacity of latent variable z by building the machine so that the volume of the low energy regions is bounded.
- Regularized methods: limit the information capacity of latent variable z by adding a regularized term to the Energy function.

(b)

$$F(y) = ||y - Dec(z)||^2 + \lambda |z|_{L_1}$$

where z is the latent variable, y is the target, $Dec(z)$ is the output of decoder.

(c) If we remove the regularizer in the objective function, there is no limitation to the information capacity of latent variable z. So for each y, there is always a latent z that can reconstruct it perfectly, and the energy function can achieve zero by choosing such a latent z.

(d) The regularizer limits the volume of low energy regions, so when doing inference, only some sparse latent z can minimize the energy (and other z would make it large by introduce in a large regularization term). With the regularization, the energy will be low for images similar to training set, and the energy will be high for images not similar to training set. For an image similar to the training set, the model can infer a latent variable z similar to the latent variables of the training set that appropriately encodes the image and reconstruct it; but for an image pretty different from the training set, the corresponding z could be largely different from latent variables of the training set, and therefore having a larger energy.

(e) The free energy function is

$$F(y) = \min_z ||y - wz||^2$$

Assume we have a 100 points in a two dimensional space, and we want to cluster the points to 5 different groups. Then for each point, $y \in R^2$, z is the latent variable and $z \in R^5$, indicating which group y should belong to. w is the parameters we need to learn, and in this case $w \in R^{2\times 5}$, with each column representing the center of a group.