Homework 3: Energy-Based Models

CSCI-GA 2572 Deep Learning

Spring 2022

The goal of homework 3 is to test your understanding of Energy-Based Models, and to show you one application in structured prediction.

In the theoretical part, we'll mostly test your intuition. You'll need to write brief answers to questions about how EBMs work. In part 2, we will implement a simple optical character recognition system.

In part 1, you should submit all your answers in a pdf file. As before, we recommend using LaTeX.

For part 2, you will implement some neural networks by adding your code to the provided ipynb file.

As before, please use numerator layout.

The due date of homework 3 is 5:00pm 03/25. Submit the following files in a zip file `your_net_id.zip` through NYU classes:

- `hw3_theory.pdf`
- `hw3_practice.ipynb`

Note: we will subtract points for Campuswire posts containing solutions to problems. Campuswire shouldn't be a platform where you can get your solution checked, the goal is to help you with any misunderstandings associated with the homework.

The following behaviors will result in penalty of your final score:

1. 10% penalty for submitting your file without using the correct naming format (including naming the zip file, PDF file or python file wrong, adding extra files in the zip folder, like the testing scripts in your zip file).
2. 20% penalty for late submission within the first 24 hours after the deadline. We will not accept any late submission after the first 24 hours.
3. 20% penalty for code submission that cannot be executed following the steps we mentioned.

1 Theory (50pt)

1.1 Energy Based Models Intuition (15pts)

This question tests your intuitive understanding of Energy-based models and their properties.

(a) (1pts) How do energy-based models allow for modeling situations where the mapping from input x_i to output y_i is not 1 to 1, but 1 to many?

(b) (2pts) How do energy-based models differ from models that output probabilities?

(c) (2pts) How can you use energy function $F_W(x, y)$ to calculate a probability $p(y \mid x)$?

(d) (2pts) What are the roles of the loss function and energy function?

(e) (2pts) Can loss function be equal to the energy function?

(f) (2pts) What problems can be caused by using only positive examples for energy (pushing down energy of correct inputs only)? How can it be avoided?

(g) (2pts) Briefly explain the three methods that can be used to shape the energy function.

(h) (2pts) Provide an example of a loss function that uses negative examples. The format should be as follows $\ell_{\text{example}}(x, y, W) = F_W(x, y)$.

Solution:

(a)

Energy Based Model can give the energy estimation of different given y, thus for one to many mapping, we can set a threshold to get the y inference list.

(b)

Probability Model is a special case of Energy Based Model.

- The energy is such that the integral $\int_{y \in \mathcal{Y}} e^{-\beta E(W,y,X)}$ (partition function) converges.

- The model is trained by minimizing the negative log-likelihood loss.

(c)

How can you use energy function $F_W(x, y)$ to calculate a probability $p(y \mid x)$?

$$P(y \mid x) = \frac{e^{-\beta F(x,y)}}{\int_{y'} e^{-\beta F(x,y')}} \tag{1}$$

(d)

What are the roles of the loss function and energy function?

The loss function is used to minimizing the energy for target training data points and maximize the energy for the rest of the data points (if they are in the loss). The energy function is used to calculate the energy of the data points.

(e)

Can loss function be equal to the energy function?

If we only consider the target training data points, the loss function is equal to the energy function. We can say we want low energy/loss in the target points.

For EBM, final optimization (if we call the whole optimization target as loss) is the optimization about energy of target points and other areas, so they are different.

(f)

What problems can be caused by using only positive examples for energy (pushing down energy of correct inputs only)? How can it be avoided?

The model is not robust, the decision boundary is not that clear, and adversarial attack can ruin the model.

Push up other areas energy. Or contrastively make the energy gap between target points and other area larger.

(g)

Briefly explain the three methods that can be used to shape the energy function.

Contrastive methods:

1. Push down on energy of training samples. Pull up on energy of suitably-generated contrastive samples or anywhere else.

2. Train a function that maps points off the data manifold to points on the data manifold.

Regularized Methods: 1. Regularizer minimizes the volume of space that can take low energy.

2. build the machine so that the volume of low energy space is bounded

3. minimize the gradient and maximize the curvature around data points:

(h)

Provide an example of a loss function that uses negative examples. The format should be as follows $\ell_{\text{example}}(x, y, W) = F_W(x, y)$.

Hinge loss

$$\mathcal{L}(x, y, W) = \sum_{\hat{y} \in \mathcal{Y}} [F_W(x, y) - F_w(x, \hat{y}) + m(y, \hat{y})]^+ \tag{2}$$

1.2 Negative log-likelihood loss (20 pts)

Let's consider an energy-based model we are training to do classification of input between n classes. $F_W(x, y)$ is the energy of input x and class y. We consider n classes: $y \in \{1, \ldots, n\}$.

(i) (2pts) For a given input x, write down an expression for a Gibbs distribution over labels y that this energy-based model specifies. Use β for the constant multiplier.

(ii) (5pts) Let's say for a particular data sample x, we have the label y. Give the expression for the negative log likelihood loss, i.e. negative log likelihood of the correct label (don't copy expressions from the slides, show step-by-step derivation of the loss function from the expression of the previous subproblem). For easier calculations in the following subproblem, multiply the loss by $\frac{1}{\beta}$.

(iii) (8pts) Now, derive the gradient of that expression with respect to W (just providing the final expression is not enough). Why can it be intractable to compute it, and how can we get around the intractability?

(iv) (5pts) Explain why negative log-likelihood loss pushes the energy of the correct example to negative infinity, and all others to positive infinity, no matter how close the two examples are, resulting in an energy surface with really sharp edges in case of continuous y (this is usually not an issue for discrete y because there's no distance measure between different classes).

(i)

For a given input x, write down an expression for a Gibbs distribution over labels y that this energy-based model specifies. Use β for the constant multiplier.

$$P_w(y \mid x) = \frac{e^{-\beta F_w(x,y)}}{\int_{y'} e^{-\beta F_w(x,y')}} \tag{3}$$

$$= \frac{e^{-\beta F_w(x,y)}}{\sum_{y' \in 1,\dots n} e^{-\beta F_w(x,y')}} \tag{4}$$

(ii)

Let's say for a particular data sample x, we have the label y. Give the expression for the negative log likelihood loss, i.e. negative log likelihood of the correct label (don't copy expressions from the slides, show step-by-step derivation of the loss function from the expression of the previous subproblem). For easier calculations in the following subproblem, multiply the loss by $\frac{1}{\beta}$.

We have (1.2), according to the negative log likelihood loss definition,

$$\mathcal{L}(x,y,w) = -\frac{1}{\beta} \log P_w(y \mid x) \qquad \text{(multiply } \frac{1}{\beta} \text{ for convenience)} \tag{5}$$

$$= -\frac{1}{\beta} \log \frac{e^{-\beta F_w(x,y)}}{\sum_{y' \in 1,\dots n} e^{-\beta F_w(x,y')}} \tag{6}$$

$$= F_w(x,y) + \frac{1}{\beta} \log \underbrace{\left[\sum_{y' \in 1,\dots n} e^{-\beta F_w(x,y')} \right]}_{P} \tag{7}$$

The first term is the energy of the correct label, the second term is the log of the sum of the energy of all other labels. we would like to minimize the correct label energy and free energy over y.

(iii)

Now, derive the gradient of that expression with respect to W (just providing the final expression is not enough). Why can it be intractable to compute it, and how can we get around the intractability?

$$\frac{\partial \mathcal{L}(x,y,w)}{\partial w} = \frac{\partial F_w(x,y)}{\partial w} + \frac{1}{\beta}\frac{\partial \log\left[\sum_{y'\in 1,\dots n} e^{-\beta F_w(x,y')}\right]}{\partial w} \tag{8}$$

$$= \frac{\partial F_w(x,y)}{\partial w} + \frac{1}{\beta}\frac{\partial \log P}{\partial P}\sum_{y'}\frac{\partial P}{\partial F(x,y')}\frac{\partial F(x,y')}{\partial w} \tag{9}$$

$$= \frac{\partial F_w(x,y)}{\partial w} + \frac{1}{\beta}\frac{1}{P}(-\beta\sum_{y'\in\{1\dots n\}} e^{-\beta F_w(x,y')})\frac{\partial F(x,y')}{\partial w} \tag{10}$$

$$= \frac{\partial F_w(x,y)}{\partial w} - \sum_{y'\in\{1\dots n\}}\frac{e^{-\beta F_w(x,y)}}{\sum_{y'} e^{-\beta F_w(x,y')}}\partial F_w\left(x,y'\right)\partial w \tag{11}$$

$$= \frac{\partial F_w(x,y)}{\partial w} - \sum_{y'} P_w\left(y' \mid x\right)\frac{\partial F_w\left(x,y'\right)}{\partial w} \tag{12}$$

In the discrete classification case, we need to calculate all the energy of all the combination of x and possible label y, which is intractable. We can use Monte Carlo Methods to sample y from P(y | x) to approximate the integral.

1.3 Comparing Contrastive Loss Functions (15pts)

In this problem, we're going to compare a few contrastive loss functions. We are going to look at the behavior of the gradients, and understand what uses each loss function has. In the following subproblems, m is a margin, $m \in \mathbb{R}$, x is input, y is the correct label, $\bar{y}$ is the incorrect label. Define the loss in the following format: $\ell_{example}(x,y,\bar{y},W) = F_W(x,y)$.

(a) (3pts) **Simple loss function** is defined as follows:

$$\ell_{\text{simple}}(x,y,\bar{y},W) = [F_W(x,y)]^+ + [m - F_W(x,\bar{y})]^+$$

Assuming we know the derivative $\frac{\partial F_W(x,y)}{\partial W}$ for any x, y, give an expression for the partial derivative of the ℓ_{simple} with respect to W.

(b) (3pts) **Hinge loss** is defined as follows:

$$\ell_{\text{hinge}}(x,y,\bar{y},W) = [F_W(x,y) - F_W(x,\bar{y}) + m]^+$$

Assuming we know the derivative $\frac{\partial F_W(x,y)}{\partial W}$ for any x, y, give an expression for the partial derivative of the ℓ_{hinge} with respect to W.

(c) (3pts) **Square-Square loss** is defined as follows:

$$\ell_{\text{square-square}}(x,y,\bar{y},W) = \left([F_W(x,y)]^+\right)^2 + \left([m - F_W(x,\bar{y})]^+\right)^2$$

Assuming we know the derivative $\frac{\partial F_W(x,y)}{\partial W}$ for any x, y, give an expression for the partial derivative of the $\ell_{\text{square-square}}$ with respect to W.

(d) (6pts) **Comparison.**

(i) (2pts) Explain how NLL loss is different from the three losses above.

(ii) (2pts) What is the role of the margin in hinge loss? Why do we take only the positive part of $F_W(x,y) - F_W(x,\bar{y}) + m$?

(iii) (2pts) How are simple loss and square-square loss different from hinge loss? In what situations would you use simple loss, and in what situations would you use square-square loss?

(a)

$$\ell_{\text{simple}}(x,y,\bar{y},W) = [F_W(x,y)]^+ + [m - F_W(x,\bar{y})]^+$$

$$\frac{\partial \ell_{\text{simple}}(x,y,\bar{y},W)}{\partial W} = \frac{\partial [F_W(x,y)]^+}{\partial W} + \frac{\partial [m - F_W(x,\bar{y})]^+}{\partial W} \tag{13}$$

$$\frac{\partial [F_W(x,y)]^+}{\partial W} = \begin{cases} \frac{\partial F_W(x,y)}{\partial W} & \text{if } F_W(x,y) > 0 \\ 0 & \text{if } F_W(x,y) \le 0 \end{cases} \tag{14}$$

$$\frac{\partial [m - F_W(x,\bar{y})]^+}{\partial W} = \begin{cases} -\frac{\partial F_W(x,\bar{y})}{\partial W} & \text{if } F_W(x,\bar{y}) < m \\ 0 & \text{if } F_W(x,\bar{y}) \ge m \end{cases} \tag{15}$$

(b)

$$\ell_{\text{hinge}}(x,y,\bar{y},W) = [F_W(x,y) - F_W(x,\bar{y}) + m]^+$$

$$\frac{\partial \ell_{\text{hinge}}(x,y,\bar{y},W)}{\partial W} = \frac{\partial [F_W(x,y) - F_W(x,\bar{y}) + m]^+}{\partial W} \tag{16}$$

$$= \begin{cases} \frac{\partial F_W(x,y)}{\partial W} - \frac{\partial F_W(x,\bar{y})}{\partial W} & \text{if } F_W(x,y) - F_W(x,\bar{y}) + m > 0 \\ 0 & \text{if } F_W(x,y) - F_W(x,\bar{y}) + m \le 0 \end{cases} \tag{17}$$

(c)

$$\ell_{\text{square-square}}(x, y, \bar{y}, W) = \left([F_W(x, y)]^+\right)^2 + \left([m - F_W(x, \bar{y})]^+\right)^2$$

$$\frac{\partial \ell_{\text{square-square}}(x, y, \bar{y}, W)}{\partial W} = 2[F_W(x, y)]^+ \frac{\partial [F_W(x, y)]^+}{\partial W} + 2[m - F_W(x, \bar{y})]^+ \frac{\partial [m - F_W(x, \bar{y})]^+}{\partial W} \tag{18}$$

$$= \begin{cases} 2[F_W(x, y)] \frac{\partial [F_W(x, y)]}{\partial W} & \text{if } F_W(x, y) > 0 \text{ and } F_W(x, \bar{y}) > m \\ 2[F_W(x, y)] \frac{\partial [F_W(x, y)]}{\partial W} - 2[m - F_W(x, \bar{y})] \frac{\partial F_W(x, \bar{y})}{\partial W} & \text{if } F_W(x, y) > 0 \text{ and } F_W(x, \bar{y}) \leq m \\ 0 & \text{if } F_W(x, y) \leq 0 \text{ and } F_W(x, \bar{y}) > m \\ -2[m - F_W(x, \bar{y})] \frac{\partial F_W(x, \bar{y})}{\partial W} & \text{if } F_W(x, y) \leq 0 \text{ and } F_W(x, \bar{y}) \leq m \end{cases} \tag{19}$$

(d)

(i)

NLI loss is not pair-wise contrastive, it try to minimize the energy of target energy and the free energy of y.

The three pair wise loss try to maxmize the energy gap between target and other location energy.

(ii)

The margin term is the gap between the target (training data point) energy and the other location energy. For the negative part, $F_W(x, y) - F_W(x, \bar{y}) + m < 0$, which means $F_W(x, y) - F_W(x, \bar{y}) < m$, the gap between the target and the contrastive negative energy is larger than the margin, we stop to optimize the energy gap.

(iii)

Optimization Just like L1 and L2 norm, the optimization of square-square loss is more efficient in non-spase case. We will use square-square loss in non-spase loss calculation. In the meanwhile, square-square loss solution space is more stable, as illustrated below.

Relative Gap and Absolute Value The simple loss and hinge loss are two-piecewise loss, the threshold is based on the relative energy gap of the target and the other location, but square-square loss can be utilized to force the target energy less than zero and the other location energy larger than m. The square-square loss do not only optimize the relative energy gap, but also optimize the absolute energy range.

If we only would like to optimize the relative energy gap, we can use the simple

loss, but if we want to control the absolute energy range, we can use the square-square loss.

2 Implementation (50pt)

Please add your solutions to this notebook `hw3_practice.ipynb` . **Plase use your NYU account to access the notebook.** The notebook contains parts marked as `TODO`, where you should put your code or explanations. The notebook is a Google Colab notebook, you should copy it to your drive, add your solutions, and then download and submit it to NYU Classes. You're also free to run it on any other machine, as long as the version you send us can be run on Google Colab.