Homework 3: Energy-Based Models

Anoushka Gupta

ag8733@nyu.edu

1 Theory (50 pts)

1.1 Energy Based Models Intuition (15 pts)

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?

Answer: In one to many mapping other models output an average of all possible outputs and hence the output is blurry. In cases where the mapping of x_i to output y_i is one to many, Energy based models output low energy for all possible y_i (targets) while all other points have high energy.

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

Answer: Probablisic models are special kind of EBM's where the energy are like un-normalized negative log probabilities. EBM gives more flexibility in choice of scoring function and in the choice of objective function for learning.

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

Answer:
Using the Gibbs-Boltzman distribution-

$P(y|x) = \frac{e^{-\beta F_W(x,y)}}{\int_{y'} e^{-\beta F_W(x,y')}}$

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

Answer: The energy function is used for inference and its also minimized during inference such that for an input x the correct target y has low energy. The loss function is minimized during training and measures the quality of the energy function on the training data. The loss function pushes down the energy of a correct answer and pushes up the energy of a incorrect answer.

(e) (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?

Answer: Pushing down energy of correct inputs only lead to a constant output where the energy is always 0. The architecture would collapse and the energy landscape/model would be flat. The energy function doesn't capture the dependency between x and y and good inference can't be done. This can be avoided by incorporating contrastive or regularized architecture methods during training. In regularized architecture methods the volume of points on the manifold which can take low energy are reduced while in the contrastive methods, additional data points are token and the energy of these points is pushed up while energy of training data is pushed down.

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

Answer:
i) Pushing down on positive examples
ii) Regularized methods where we restrict the volume of manifold which can have low energy
iii) Contrastive methods- adding data point other than training data and pushing up the energy of these points while pushing down energy of the training points.

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

Answer:

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

(h) (2pts) Say we have an energy function $F(x, y)$ with images x, classification for this image y. Write down the mathematical expression for doing inference given an input x. Now say we have a latent variable z, and our energy is $G(x, y, z)$. What is the expression for doing inference then?

Answer:
Inference-

$\check{y} = argmin_y F(x, y)$

Inference with latent variable z
$F_w(x, y) = E_w(x, y, \check{z})$
$\check{z} = argmin_{z \in \mathbb{Z}} E_w(x, y, z)$

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: $y \in \{1, ..., n\}$.

(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.

Answer: From Gibbs distribution -

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

Since y is discrete here and number of classes is finite we have-

$P(y \mid x) = \frac{e^{-\beta F_W(x,y)}}{\sum_{y'=1}^{n} e^{-\beta F_W(x,y')}}$

(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 (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 $\frac{1}{\beta}$.

Loss is given by-

$L(x, y, W) = \frac{-1}{\beta} log(P_W(y \mid x))$

$= \frac{-1}{\beta} log \frac{e^{-\beta F_W(x,y)}}{\sum_{y'=1}^{n} e^{-\beta F_W(x,y')}}$

$= \frac{-1}{\beta} \left[log\ e^{-\beta F_W(x,y)} - log \left(\sum_{y'=1}^{n} e^{-\beta F_W(x,y')} \right) \right]$

$= \frac{-1}{\beta} \left[-\beta F_W(x, y) - log \left(\sum_{y'=1}^{n} e^{-\beta F_W(x,y')} \right) \right]$

$= F_W(x, y) + \frac{1}{\beta} log \left(\sum_{y'=1}^{n} e^{-\beta F_W(x,y')} \right)$

(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?

$L(x, y, W) = \frac{-1}{\beta} log(P_W(y \mid x))$

$= F_W(x, y) + \frac{1}{\beta} log\left(\sum_{y'=1}^{n} e^{-\beta F_W(x,y')}\right)$

$\frac{\partial L}{\partial W} = \frac{\partial F_W(x,y)}{\partial W} + \frac{\partial\left(\frac{1}{\beta} log\left(\sum_{y'=1}^{n} e^{-\beta F_W(x,y')}\right)\right)}{\partial W}$

$= \frac{\partial F_W(x,y)}{\partial W} + \frac{1}{\beta} \times \frac{1}{\left(\sum_{y'=1}^{n} e^{-\beta F_W(x,y')}\right)} \times \frac{\partial\left(\sum_{y'=1}^{n} e^{-\beta F_W(x,y')}\right)}{\partial W}$

$= \frac{\partial F_W(x,y)}{\partial W} + \frac{1}{\beta} \times \frac{1}{\left(\sum_{y'=1}^{n} e^{-\beta F_W(x,y')}\right)} \times \left(-\sum_{y'=1}^{n} \beta e^{-\beta F_W(x,y')} \frac{\partial F(x,y')}{\partial W}\right)$

$= \frac{\partial F_W(x,y)}{\partial W} - \frac{1}{\left(\sum_{y'=1}^{n} e^{-\beta F_W(x,y')}\right)} \times \left(\sum_{y'=1}^{n} e^{-\beta F_W(x,y')} \frac{\partial F(x,y')}{\partial W}\right)$

$= \frac{\partial F_W(x,y)}{\partial W} - \sum_{y'=1}^{n} P(y' \mid x) \frac{\partial F_W(x,y)}{\partial W}$

If the label y was continous the gradient would be

$\frac{\partial L}{\partial W} = \frac{\partial F_W(x,y)}{\partial W} - \int_{y'} P(y' \mid x) \frac{\partial F_W(x,y)}{\partial W}$

The integral term sometimes can't be computed because its high dimension etc. And hence its intractable to compute it. This can be solved by using Monte Carlo methods where we sample y from $P(y \mid x)$.

(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).

Answer: $\frac{\partial L}{\partial W} = \frac{\partial F_W(x,y)}{\partial W} - \int_{y'} P(y' \mid x) \frac{\partial F_W(x,y)}{\partial W}$

The first term in the above equation pushes up the energy for correct and incorrect pairs. The second term in the equation is proportional to the probability of getting the incorrect y'. The probability of getting an incorrect y' is low while $F_W(x, y)$ is large and in totality NLL pushes the energy to positive infinity for incorrect pairs. Whereas for correct pairs the probability of getting the correct y' is high. This pushes down the energy to negative infinity.

1.3 Comparing Contrastive Loss Functions (15 pts)

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:

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

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.

Answer:

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

$$\frac{\partial l_{simple}(x,y,\bar{y},W)}{\partial W} = \begin{cases} 0 & \text{if } F_W(x,y) \leq 0 \text{ and } m \leq F_W(x,\bar{y}) \\ -\frac{\partial F_W(x,\bar{y})}{\partial W} & \text{if } F_W(x,y) \leq 0 \text{ and } m > F_W(x,\bar{y}) \\ \frac{\partial F_W(x,y)}{\partial W} & \text{if } F_W(x,y) > 0 \text{ and } m \leq F_W(x,\bar{y}) \\ \frac{\partial F_W(x,y)}{\partial W} - \frac{\partial F_W(x,\bar{y})}{\partial W} & \text{if } F_W(x,y) > 0 \text{ and } m > F_W(x,\bar{y}) \end{cases}$$

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

$$\ell_{log}(x, y, \bar{y}, W) = log\left(1 + e^{F_W(x,y) - F_W(x,\bar{y})}\right)$$

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_{log} with respect to W.

Answer:

$$\frac{\partial l_{log}(x,y,\bar{y},W)}{\partial W} = \frac{\partial log\left(1+e^{F_W(x,y)-F_W(x,\bar{y})}\right)}{\partial W}$$

$$= \frac{1}{\left(1+e^{F_W(x,y)-F_W(x,\bar{y})}\right)} \frac{\partial\left(1+e^{F_W(x,y)-F_W(x,\bar{y})}\right)}{\partial W}$$

$$= \frac{1}{\left(1+e^{F_W(x,y)-F_W(x,\bar{y})}\right)} \left(0 + \frac{\partial\left(e^{F_W(x,y)-F_W(x,\bar{y})}\right)}{\partial W}\right)$$

$$= \frac{e^{F_W(x,y)-F_W(x,\bar{y})}}{\left(1+e^{F_W(x,y)-F_W(x,\bar{y})}\right)} \frac{\partial(F_W(x,y)-F_W(x,\bar{y})))}{\partial W}$$

$$= \frac{1}{\left(1+e^{-(F_W(x,y)-F_W(x,\bar{y}))}\right)} \left(\frac{\partial(F_W(x,y))}{\partial W} - \frac{\partial(F_W(x,\bar{y}))}{\partial W}\right)$$

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

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

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 $\ell_{square-square}$ with respect to W.

Answer: $\frac{\partial l_{square-square}(x,y,\bar{y},W)}{\partial W} = \frac{\partial([F_W(x,y)]^+)^2}{\partial W} + \frac{\partial([m-F_W(x,\bar{y})]^+)^2}{\partial W}$

$$\frac{\partial l_{square-suare}(x,y,\bar{y},W)}{\partial W} = \begin{cases} \quad 0 & \text{if } F_W(x,y) \leq 0 \text{ and } m \leq F_W(x,\bar{y}) \\ -2(m - F_W(x,\bar{y}))\left(\frac{\partial F_W(x,\bar{y})}{\partial W}\right) & \text{if } F_W(x,y) \leq 0 \text{ and } m > F_W(x,\bar{y}) \\ 2(F_W(x,y))\left(\frac{\partial F_W(x,y)}{\partial W}\right) & \text{if } F_W(x,y) > 0 \text{ and } m \leq F_W(x,\bar{y}) \\ 2(F_W(x,y))\left(\frac{\partial F_W(x,y)}{\partial W}\right) - 2(m - F_W(x,\bar{y}))\left(\frac{\partial F_W(x,\bar{y})}{\partial W}\right) \\ \text{if } F_W(x,y) > 0 \text{ and } m > F_W(x,\bar{y}) \end{cases}$$

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

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

Answer:

NLL is not pair wise contrastive and it doesn't depened on the margin like simple loss an square square loss

(ii) (2pts) The hinge loss $[F_W(x,y) - F_W(x,\bar{y}) + m]^+$ has a margin parameter m, which gives 0 loss when the positive and negative examples have energy that are m apart. The log loss is sometimes called a "soft-hinge" loss. Why? What is the advantage of using a soft hinge loss?

Answer: When the margin becomes very big, the log loss is called soft-hinge. The soft-hinge is more smooth.

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

Answer: Simple loss can lead to vanishing gradient while square-square loss doesn't. The hinge loss takes in account the energy of both the correct and incorrect pairs while using margin and the other loss don't.