CSCI-GA 2572 Deep Learning - Homework 3

Due: October 15 @ 11:59pm

Name: Stephen Spivack (ss7726)

1 Theory

1.1 Energy Based Models Intuition

(a) 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?

By assigning low energy values to (x, y) pairs that are correct or desirable and high energy values to those that are incorrect or undesirable; inference is based on finding the minimal energy value, and there can be more than one of these which would allow for such a mapping.

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

These models do not directly output probabilities; they output energy values, which can be converted to probabilities using some kind of normalization, i.e., Gibbs-Boltzmann distribution.

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

Use Gibbs-Boltzmann: $P(y|x) = \frac{e^{-\beta F(x,y)}}{\int_{\hat{y}} e^{-\beta F(x,\hat{y})}}$

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

The loss function computes the discrepancy between the predicted and true (x, y) pair, whereas the energy function actually defines the compatibility of that (x, y) pair. The loss is minimized during training, which forces low energy for correct configurations and high energies for incorrect ones.

(e) 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 will produce identical, constant outputs. Therefore, negative examples are necessary to produce a varied energy surface with distinct local minima corresponding to given configuration of (x, y).

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

Collapse - adding negative training examples will increase variation in energy surface to guard against identical, constant inputs.

Contrastive method - train model to push down on the energy of correct inputs while pulling up on the energy of well-chosen negative example inputs.

Regularized methods - prevent collapse, allow for more variation in the energy surface

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

For example, we can use pairwise margin loss: $L(x, y, W) = max(0, \Delta + F_W(x, y^+) - F_W(x, y^-))$

(h) 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?

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

$(\hat{y}, \hat{z}) = argmin_{y,z} G(x, y, z)$

1.2 Negative log-likelihood loss

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) 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.

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

(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 (show step-by-step derivation of the loss function from the expression of the previous sub-problem). For easier calculations in the following sub-problem, multiply the loss by $\frac{1}{\beta}$.

$\ell_{NLL} = -logP(y|x) = -log(\frac{e^{-\beta F_W(x,y)}}{\sum_{\bar{y}} e^{-\beta F_W(x,\bar{y})}}) = \beta F_W(x, y) + log(\sum_{\bar{y}} e^{-\beta F_W(x,\bar{y})}) = F_W(x, y) + \frac{1}{\beta} log(\sum_{\bar{y}} e^{-\beta F_W(x,\bar{y})})$

(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?

$\frac{\partial \ell_{NLL}}{\partial W} = \frac{\partial F_W(x,\bar{y})}{\partial W} - \frac{1}{\beta} \frac{\sum_{\hat{y}} e^{-\beta F_W(x,\bar{y})} \frac{\partial F_W(x,\bar{y})}{\partial W}}{\sum_{\bar{y}} e^{-\beta F_W(x,\bar{y})}}$

Given the summation term, if there are several classes, this can be very computationally expensive to compute. We can use SGD to overcome this limitation.

(iv) Explain why negative log-likelihood loss pushes the energy of the correct example to $-\inf$, and all others to $+\inf$, 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).

The negative log-likelihood pushes energy of correct example to negative infinity because as energy of correct class decreases, probability approaches 1. For examples that are pushed toward infinity the probability approaches 0, which results in a sharp edge in the error surface.

1.3 Comparing Contrastive Loss Functions

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: $l_{example}(x, y, \bar{y}, W) = F_W(x, y)$.

(a) 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.

$\frac{\partial F_W(x,y)}{\partial W}$ if $F_W(x, y) > 0$;

$-\frac{\partial F_W(x,\bar{y})}{\partial W}$ if $m - F_W(x, \bar{y}) > 0$;

0 otherwise.

(b) Log loss is defined as follows:

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

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_{log} with respect to W.

Take $z = F_W(x, y) - F_W(x, \bar{y})$

Now we can update the loss in terms of z: $\ell_{log}(x, y, \bar{y}, W) = log(1 + e^z)$

$\frac{\partial \ell_{log}}{\partial W} = frac11 + e^z \cdot \frac{\partial}{\partial W} e^z$

$\frac{\partial \ell_{log}}{\partial W} = \frac{e^z}{1+e^z} \cdot (\frac{\partial F_W(x,y)}{\partial W} - \frac{\partial F_W(x,\bar{y})}{\partial W})$

(c) 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.

Take $y = [F_W(x, y)]^+$ and $z = [m - F_W(x, \bar{y})]^+$; then

$\frac{\partial \ell_{square-square}}{\partial W} = 2y \frac{\partial F_W(x,y)}{\partial W} - 2z \frac{\partial F_W(x,\bar{y})}{\partial W}$

(d) Comparison

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

NLL maximizes the probability assigned to the correct class

(ii) 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?

It is smooth and differentiable, so we can use SGD. Soft hinge loss provides a differentiable approximation to hinge loss and provides a probabilistic interpretation.

(iii) 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?

Simple loss penalizes positive part of correct class and positive part of margin between incorrect class and a given value; square-square loss squares these before summing. Simple loss is good for a linear penality where square-square is good for a squared/quadratic penalty.