Homework 3

Guilherme Albertini

October 21, 2022

Theory

Problem 1.1: Energy Based Models Intuition

1. 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?

> We are mapping pairs (x, y) to a scalar energy value and find that the most likely values of y have a low $F(x, y)$. We then observe that, for each x, we can have several different values y that have this low energy.

> Note a good definition from DeepAI: "Energy-Based Models (EBMs) discover data dependencies by applying a measure of compatibility (scalar energy) to each configuration of the variables. For a model to make a prediction or decision (inference) it needs to set the value of observed variables to 1 and finding values of the remaining variables that minimize that "energy" level. In the same way, machine learning consists of discovering an energy function that assigns low energies to the correct values of the remaining variables, and higher energies to the incorrect values. A so-called "loss functional," that is minimized during training, is used to measure the quality of the energy functions. Within this framework, there are many energy functions and loss functionals allows available to design different probabilistic and non-probabilistic statistical models."
> Words from Alf: "We would like the energy function to be smooth and differentiable so that we can use it to perform the gradient-based method for inference. In order to perform inference, we search this function using gradient descent to find compatible y's. There are many alternate methods to gradient methods to obtain the minimum."

2. How do energy-based models differ from models that output probabilities?

> As is the key to their flexibility, we need not concern ourselves with normalization as EBMs output an unnormalized scalar (score) of $F(x, y)$ as opposed to conditional probabilities (i.e $\mathbb{P}(y|x)$ would later require an estimate of normalization).

3. How can you use energy function $F_W(x, y)$ to calculate a probability $\mathbb{P}(y|x)$?

> We can view energies as unnormalised negative log probabilities, and use Gibbs-Boltzmann distribution to convert from energy to probability (with normalization and calibrated β):
>
> $$\mathbb{P}(y|x) = \frac{\exp(-\beta F(x, y))}{\int_{y'} \exp(-\beta F(x, y'))}$$
>
> Note: β is postive constant and larger values produce models with more variance whereas smoother ones are produced with smaller values.

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

> The energy function is a measure of incompatibility between variables (for us, usually the input x and output y) whereas the loss function is used to mold the energy function (we minimize loss to end up with a well-behaved energy function). Note that the cost is how far prediction $\hat{y}$ is from target y. As Yann mentions: A loss functional, minimized during learning, is used to measure the quality of the available energy functions. A distinction should be made between the energy function, which is minimized by the inference process, and the loss functional, which is minimized by the learning process.

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

We may get a case of having energy be 0 everywhere, which is a valid minimization of the (flat) energy surface under this constraint. As this flat model can reach every location of the space, the distance between any two points (such as the length of the latent vector spanning the embedded model manifold's reconstructed $\tilde{y} = Wz$ to observed target y) is 0, hence at the minimum energy by default. To avoid this state of collapse case, we can augment $y = [1, y]^T$ to give an additional degree of freedom to dictionary $W = [1, W]^T$, so that we can now intersect any point in the 2D space but only at those points located at the specific height (here, 1) that gives the minimum energy near 0.

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

Regularization Methods: if the latent variable z is too expressive power in producing the final prediction $\tilde{y}$ then every true output y will be a perfect reconstruction from input x at the optimized latent $\check{z}$. We can then limit the volume of space of z (say, with L1 loss to promote sparsity) and therby reduce the regions of y with low energy, preventing the case of getting energy 0 everywhere.

Contrastive Methods: Push down the energy of training data points, $F(X_i, Y_i)$, while pushing up energy on everywhere else, $F(X_i, Y')$.

Architectural Methods: The manifold is of lower dimension than the ambient space so the data cannot be reconstructred perfectly. Autoencoders can reduce the dimensionality of the input in the hidden layer (under-complete) and thus cannot reconstruct data perfectly, preventing collapse. If the hidden space is over-complete, we can use encoding with higher dimensionality than the input to make optimization easier.

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

$$\ell_{NLL}(x, y, W) = -\log\left(\frac{\exp(-\beta F(x, y))}{\int_{y'} \exp(-\beta F(x, y'))}\right)$$

8. 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?

$$\check{y} = \arg\min_y F(x, y)$$
$$\check{z}, \check{y} = \arg\min_{y,z} G(x, y, z)$$

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

1. 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.

$$F_\beta(x, y) = -\frac{1}{\beta} \log \int_z \exp(-\beta G(x, y, z))$$

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

$$\begin{aligned}
\ell_{NLL}(x, y, W) &= -\log\left(\frac{\exp(-\beta F_W(x, y))}{\int_{y'} \exp(-\beta F_W(x, y'))}\right) \\
&= \log\left(\int_{y'} \exp(-\beta F_W(x, y'))\right) + \beta F_W(x, y) \\
&\implies \frac{1}{\beta} \log\left(\int_{y'} \exp(-\beta F_W(x, y'))\right) + F_W(x, y)
\end{aligned}$$

Where the last step was from dividing by β.

3. 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?

ya