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 energy surface under this constraint. To avoid this degenerate case, we pull up negative examples by $-\nabla_{F(x,\hat{y})} = softargmin_\beta[F(x,Y)=0]^T\hat{y} \rightarrow \frac{1}{K}$ for K classes in the initial normal distribution and push down the positive examples by $-\nabla_{F(x,y)}$ of height 1 for sufficiently large Y (Aside: $-\nabla_{F(x,y)} = -1 + \frac{\exp(-\beta F(x,y))}{\sum_{y' \in Y}\exp(-\beta F(x,y'))}$)

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

   ye

7. Provide an example of a loss function that uses negative examples.The format should be as follows ‘example(x, y,W) = FW (x, y).