Homework 3: Energy-Based Models

Yue Zhou

Spring 2021

1 Theory

1.1 Energy Based Models Intuition

(a) For the energy-based models, the output y_i not only depends on x_i, but an extra variable z_i, which called the latent variable that we do not know the value of it. By varying the latent variable over a set, we can make the prediction output y_i vary over the manifold of possible predictions, which allows the mapping from input x_i to output y_i is not 1 to 1, but 1 to many.

(b) The difference is that in probabilistic models, you basically don't have the choice of the objective function you're going to minimize, and you have to stay true to the probabilistic framework in the sense that every object you manipulate has to be a normalized distribution. But if you're going to train a system end-to-end to make decisions, then whatever scoring function you use is fine, as long as it gives the best score to the best decision. Energy-based models give you way more choices in how you handle the model, way more choices of how you train it, and what objective function you use. Meanwhile, it's no need for proper normalization.

(c) We can look at the energies as unnormalized negative log probabilities, and use Gibbs-Boltzmann distribution to convert from energy to probability after normalization is:

$$P(y|x) = \frac{exp(-\beta F_w(x,y))}{\int_{y'} exp(-\beta F_w(x,y'))} \tag{1}$$

where β is positive constant and needs to be calibrated to fit your model.

(d) The loss function is to calculate the difference between the true value and the predicted value, which means that is the probability of your result being wrong. The energy function is used to capture the dependency between x and y, which can be interpreted as whether or not it matches. The lower the energy is, the higher the match is.

(e) In some cases, loss function can be equal to energy function. When your energy model is still used to be probabilistic, maximum likelihood should be used in your energy function. In this case, your energy function and loss function are almost equivalent.

(f) Because when it only pushes down energy of correct inputs, in some cases it will even make all the energy of inputs zero and then we can't find the answers we wanted.

(g) PCA: build the machine so that the volume of low energy stuff is constant.

Max likelihood: push down of the energy data points, push up everywhere else.

Contrastive divergence: push down of the energy of data points, push up on chosen locations.

(h) Generalized Margin Loss is the loss function that uses a negative example. This loss function is looking for the Most offending incorrect answer, which has the lowest energy among all possible answers that are incorrect. And then if the energy function is able to ensure that energy of the Most offending incorrect answer is higher than the correct answer by some margin, this energy function should work well. Hinge Loss is an example:

$$L_{hinge}(X^i, Y^i, W) = \max(0, m + F_W(X^i, Y^i)) - F_W(X^i, \bar{Y}^i) \quad (2)$$

Where $\bar{Y}^i$ is the most offending incorrect answer. This loss enforces that the difference between the correct answer and the most offending incorrect answer be at least m.

1.2 Negative log-likelihood loss

(i)

$$P(y|x) = \frac{e^{-\beta F_w(x,y)}}{\int_{y' \in y} e^{-\beta F_w(x,y')}} \quad (3)$$

(ii) Assuming that the samples are independent, and denoting by $P(y^i|x^i, W)$ the conditional probability of y^i given x^i that is produced by our model with parameter W, the conditional probability of the training set under the model is a simple product over samples:

$$P(y^1, \ldots, y^P|x^1, \ldots, x^P, W) = \prod_{i=1}^{P} P(y^i|x^i, W) \quad (4)$$

Applying the maximum likelihood estimation principle, we seek the value of W that maximizes the above product, or the one that minimizes the negative log of the above product:

$$-\log \prod_{i=1}^{P} P(y^i|x^i, W) = \sum_{i=1}^{P} -\log P(y^i|x^i, W) \tag{5}$$

Using the Gibbs distribution (the previous subproblem), we get:

$$-\log \prod_{i=1}^{P} P(y^i|x^i, W) = \sum_{i=1}^{P} \beta F_W x^i, y^i + \log \int_{y' \in y} e^{-\beta F_W(x^i, y')} \tag{6}$$

The final form of the negative log-likelihood loss:

$$L_{nll}(W, x^i, y^i) = \frac{1}{P} \sum_{i=1}^{P} (F_W x^i, y^i + \frac{1}{\beta} \log \int_{y' \in y} e^{-\beta F_W(x^i, y')}) \tag{7}$$

(iii) The gradient of that expression with respect to W is:

$$\frac{\partial L_{nll}(W, x^i, y^i)}{\partial W} = \frac{\partial F_W x^i, y^i}{\partial W} - \int_{y' \in y} \frac{\partial F_W x^i, y'}{\partial W} P(y'|x^W) \tag{8}$$

Why it can be intractable tocompute is because it has an integral with log and exp operations. Multiply the loss by $\frac{1}{\beta}$ makes it easier to calculate.

(iv) The first term of loss function will push down on the energy of the correct answer while the second term pushing up on the energies of all answers in proportion to their probabilities. The energy of the correct answer is also pushed up, but not as hard as it is pushed down by the first term. For the correct example, when its energy goes down, the second term will become larger, the first term shoul become smaller as a result, which means that it's pushed to negative infinity. And for the others, when their energies are pushed up, the second term for them will be smaller, which minimize when the energy go to positive infinity.

1.3 Comparing Contrastive Loss Functions

(a)

$$\frac{\partial l_{simple}(x, y, \bar{y}, W)}{\partial W} = [\frac{\partial F_W(x, y)}{\partial W}]^+ + [-\frac{\partial F_W(x, \bar{y})}{\partial W}]^+ \tag{9}$$

(b)

$$\frac{\partial l_{hinge}(x, y, \bar{y}, W)}{\partial W} = [\frac{\partial F_W(x, y)}{\partial W} - \frac{\partial F_W(x, \bar{y})}{\partial W}]^+ \tag{10}$$

(c)

$$\frac{\partial l_{square-square}(x, y, \bar{y}, W)}{\partial W} = [2F_W(x, y)\frac{\partial F_W(x, y)}{\partial W}]^+ + [-2F_W(x, \bar{y})\frac{\partial F_W(x, \bar{y})}{\partial W}]^+ \tag{11}$$

(d) (i) These three losses above are Generalized Margin Loss that uses a negative example, which look for the Most offending incorrect answer that has the lowest energy among all possible answers that are incorrect. On the contrary, NLL loss mainly pushes down on the energy of the correct answer while pushing up on the energies of all answers in proportion to their probabilities.

(ii) The difference between the energies of the correct answer and the most offending incorrect answer is penalized linearly when larger than margin.

The positive part of $F_W(x, y) - F_W(x, \bar{y}) + m$ means that the energies of the incorrect answer is margin larger than the correct answer, which needs to be amended. And other cases means that their energies are right.

(iii) Unlike the hinge loss, the simple loss and square-square loss treat the energy of the correct answer and the most offending answer separate. The energy of the correct answer and the energy of the wrong answer below margin are penalized, which will push the energy of the correct answer to zero and the wrong answer to above m.

In the ordinary case you can choose to use simple loss, and when want the energy to be more discrete or allow the model to work better, you can use square-square loss.