Deep Learning HW3

wbg231

October 15, 2023

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?*

- energy-based models can have multiple outputs for a single input because they have a defined energy function F which is used for inference.
- so that is there may be multiple solutions to $\check{y} = argmin_y F(x, y)$

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

- in short energy based models offer more flexibly in the choice of our scoring function
- proclitic models are a subset of energy based models
- further probabilistic models aim to have the energies of samples on the data manafold be infintly large, while all other dat is infinity high.

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

- energies can be thought of as un-normalized negative log probabilities
- so given a $\beta \in \mathbb{R} > 0$ an observation $x \in \mathbb{R}^n$ some set of potential inferences Y and an energy function $F(x, y) : (\mathbb{R}^n, \mathbb{R}) \to \mathbb{R}$ then for any $y \in Y$ we can calculate the conditional probability of y given x as follows:

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

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

- loss functions are used to learn an energy function
- energy functions are used for inference

(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 energy function can collapse (that is ignore the input and produce identical constant outputs)
- this is kind of a case of models having to many degrees of freedom
- for instance a Generative latent-variable Architecture can collapse if the degree of our latent variable z is greater than that of our target y in this case we can always find a z that will produce the same output y and thus result in zero energy regardless of the input
- another example is in auto-encoders where if the model learns the energy function the energy is always zero
- a final example could be joint embedding models which can collapse if the models learn the same embedding and thus will always produce 0 energy
- this problem can be avoided by using negative examples

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

1. max likelihood - ie probabilistic methods that only push down the energies of observed data points
2. regularized - learns the energy function by limiting the volume of low energy regions through regularization
3. contrastive - learn an energy function by pushing down the energy of positive examples(in the real data) and pushing up the energy of negative examples (simulated data points)

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

- an example of such a loss function is the simple loss function

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

(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?*

- for energy function F the expression for doing inference is

$$y^* = argmin_y F(x, y)$$

- for energy function G the expression for doing inference is

$$y^* = argmin_{y,z} G(x, y, z)$$

1.2 Negative Log-Likelihood Loss

Given:
- 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.
- n classes: $y \in \{1, \ldots, n\}$.

(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.

- we know that the set of labels Y has cardinality n so it is discrete
- thus we can write the Gibbs distribution as follows:

$$P(y) = \frac{e^{-\beta F_w(y)}}{\sum_{y' \in Y} F_w(y')}$$

  so more or less a softmax

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

- given a pair (x,y) we know the likelihood of the pair is given by $L(x, y, w) = P_w(y|x) = \int_{z'} P(y, z'|x) = \frac{\int_{z'} e^{-\beta E_w(x,y,z')}}{\int_{z'} \sum_{y'} e^{-\beta E_w(x,y',z')}}$
- item then we can get the negative log likclyhood as follows:

$$-\log(L(x, y, w)) = -\log(\frac{\int_{z'} e^{-\beta E_w(x,y,z')}}{\int_{z'} \sum_{y'} e^{-\beta E_w(x,y',z')}}) = \int_{z'} log(e^{-\beta E_w(x,y',z')}) + log(\int_{z'} \sum_{y'} e^{-\beta E_w(x,y',z')})$$

- this can be expressed as

$$-\beta F_w(x, y) - log(\sum_{y'} F_w(x, y'))$$

- and finally multiplying this expression by $-\frac{1}{\beta}$ we can get

$$L(x, y, w) = F_w(x, y) + \frac{1}{\beta} \sum_{y'} log(e^{-\beta F_w(x,y')})$$

(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?

- ok word. so we have from last question, the expression for our negative log likelihood (multiplied by $-(\frac{1}{\beta})$) as follows

$$L(x, y, w) = F_w(x, y) + \frac{1}{\beta} \sum_{y'} log(e^{-\beta F_w(x,y')})$$

- we can take the gradient of this in parts as differentiation is a linear operation
- so lets just focus finding the gradient of the second term

$$\frac{\partial}{\partial w}(\frac{1}{\beta} \sum_{y'} log(e^{-\beta F_w(x,y')}) = (\frac{1}{\beta})(\sum_{y'} \frac{e^{-\beta f_w(x,y')}}{\sum_{y''} e^{-\beta f_w(x,y'')}})(-\beta)(\frac{\partial f_w(x, y')}{\partial w})$$

$$= -\sum_{y'} \frac{e^{-\beta f_w(x,y')}}{\sum_{y''} e^{-\beta f_w(x,y'')}}(\frac{\partial f_w(x, y')}{\partial w})$$

- then we can marginalize over the unused latent variable z as

$$(\sum_{y'} \frac{e^{-\beta f_w(x,y')}}{\sum_{y''} e^{-\beta f_w(x,y'')}}(\frac{\partial f_w(x, y')}{\partial w}))$$

$$= (\int_{z'} \sum_{y'} \frac{e^{-\beta f_w(x,y',z')}}{\int_{z''} \sum_{y''} e^{-\beta f_w(x,y'',z'')}})(\frac{\partial f_w(x, y')}{\partial w}) = \sum_{y'} P(y'|x) \frac{\partial f_w(x, y')}{\partial w}$$

- so then we finally get

$$\frac{\partial L(x, y, w)}{\partial w} = \frac{\partial F_w(x, y)}{\partial w} - \sum_{y'} P(y'|x) \frac{\partial f_w(x, y')}{\partial w}$$

- finally computing the second term is intractable since we would have to compute over the domain of Y which could be quite large. we get around this by using monte carlo methods to sample from $p_w(y|x)$

(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).

- looking a the gradient we can see that negative log-likelihood will try to push the energy of negative examples down as far as possible, while at the same time raising the energy of the negative samples. So in other words the only low energy points will have been observed in the data, and all other points no matter how close will have high energies. this can create a really jagged loss surface, as even points which are close tighter are given vastly dirent energies

1.3 Comparing Contrastive Loss Functions

Given:
- m is a margin, $m \in \mathbb{R}$, x is input, y is the correct label, $\bar{y}$ is the incorrect label.
- loss: $\ell_{example}(x, y, \bar{y}, W) = F_W(s, y)$.

(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 $\frac{\delta F_W(x,y)}{\delta W}$ *for any* x, y, *give an expression for the partial derivative of the* ℓ_{simple} *with respect to* W.

- ok word $\nabla_w \ell = a + b$ where $a = \begin{cases} 0 & \text{if } F_w(x, y) < 0 \\ \frac{\partial F_w(x,y)}{\partial w} & \text{else} \end{cases}$ and $b = \begin{cases} 0 & \text{if } F_w(x, \bar{y}) < m \\ -\frac{\partial F_w(x,\bar{y})}{\partial w} & \text{else} \end{cases}$

(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 $\frac{\delta F_W(x,y)}{\delta W}$ *for any* x, y, *give an expression for the partial derivative of the* ℓ_{log} *with respect to* W.

- right so we can write this as $\nabla_w \ell = \frac{e^{F_W(x,y) - F_w(x,\bar{y})}}{1 + e^{F_W(x,y) - F_w(x,\bar{y})}} (\frac{\partial F_w(x,y)}{\partial w} - \frac{\partial F_w(x,\bar{y})}{\partial w})$

(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 $\frac{\delta F_W(x,y)}{\delta W}$ *for any* x, y, *give an expression for the partial derivative of the* $\ell_{square-square}$ *with respect to* W.

- we can write this as $\nabla_w \ell = a + b$ where $a = \begin{cases} 0 & \text{if } F_w(x, y) < 0 \\ 2\frac{\partial F_w(x,y)}{\partial w} & \text{else} \end{cases}$ and $b = \begin{cases} 0 & \text{if } F_w(x, \bar{y}) < m \\ \frac{\partial -2F_w(x,\bar{y})}{\partial w} & \text{else} \end{cases}$

(d) *Comparison.*

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

- NLL differs in that it uses the calculation of an integral over the domain of Y to calculate the loss, in effect pushing the energy of all other examples up while pushing the energy of the single positive sample down.
- the other three only look at one one negative example at a time, and thus at each evaluation only raise the energy of one negative point and lower the enrergy of one positive point.

(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?

- the log loss has a more smooth fall off compared to hinge loss. hinge loss either fires if $F_w(x, y) \geq F_w(x, \bar{y}) - m$ or does not fire at all.
- if we call $l = F_w(x, y) - F_w(x, \bar{y})$ then as $l \to \infty$ our log loss $\to \infty$, while as $l \to 0$ our hinge loss $\to 1$ but never stops firing to some extent

(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?

- the simple and sqaure loss look at signs of $F_w(x, y)$ and $m - F_w(x, \bar{y})$ limiting how low $F_w(x, y)$ and how high $m - F_w(x, \bar{y})$ could be
- the hinge and log loss look at the pair wise deference between $F_w(x, y)$ and $F_w(x, \bar{y})$ and thus ensures that the relative diference between these energies is low
- the simple loss is really computationally efficient so it could be good to use when we have really large datasets
- the square loss is good when we want to penalize outliers more than the simple loss