hw4$_t$*heory*

ak12016

March 2025

1 Energy Based Models Intuition

1.1 a

Energy Based Models allow for a mapping of 1 to many or 1 to inifinitely-many by introducing a latent (continuous) variable z. They define a probabilistic mapping where many different (or infinite since z can be continous) values of y are generated based on the different settings of z

1.2 b

Energy Based Models score how compatible an input x and output y by learning an Energy function that assigns it a scalar energy value and assign lower energies to more compatible pairs and higher to less compatible pairs, whereas other models directly compute normalized probabilities that sum to 1.

1.3 c

We use the Energy Function to calculate a probability by using the Boltzmann distribution (I am assuming y is continuous if its discrete we would just substitute the integral for summation).

$$p(x|y) = \frac{exp(-\beta F_w(x,y))}{\int_{y'} exp(-\beta F_w(x,y'))dy}$$

1.4 d

We can control the smoothness of the probabilities with our Beta parameter, when Beta is very small our distribution will be very smooth but as Beta increases the sharpness of our distribution also increases (and thus reducing the variance in the distribution as well).

1.5 e

The Energy function assigns an Energy score to each x, y pair determining that pairs' compatibility score. The loss function then determines how well that energy function is doing, and by taking its gradient we can update the weights to get better results

1.6 f

By using only positive examples for energy (pushing down the energy of correct inputs only), the model will only learn to push down energies, thus not giving it the ability to distinguish between good and bad, making the model collapse. We can prevent this by implementing contrastive learning techniques where the model punishes bad outputs by increasing their energies and implementing a margin-based loss function that explicitly enforces a margin between the energy of positive and negative examples we can also force negative samples into the training set with negative sampling.

1.7 g

3 methods to shape energy functions (i) Contrastive methods - comparing positive examples with negative examples and encouraging the model to lower energy for positive examples and increase it for negative examples we can "push and pull" the energy landscape (ii) Margin-based loss function - a loss that forces a gap between the energy of correct pairs and that of incorrect pairs. This also ensures that we can "push and pull" the energy landscape (iii) Negative sampling - we pretty much force negative samples into the training set and thus make sure the model doesnt collapse and shaping the energy landscape to be good.

1.8 h

$$\ell_{example}(x, y, W) = \max\Big(0,\ \alpha\ +\ F_W(x, y^+)\ -\ F_W(x, y^-)\Big),$$

where α ¿ 0 is the margin and y^+ represents a positive output and y^- represents a negative output

1.9 i

Given our energy function, then the expression for inference is just

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

If we now introduce a latent vatiable z and our Energy function becoes G(x,y,z) then the way we would be doing inference would be

$$F(x, y) = -log \int exp(-\beta G(x, y, z)]dz$$

$$\hat{y} = argminF(x, y)$$

as temperature goes to 0

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

2 Negative Log-Likelihood Loss

2.1 i

$$p(y \mid x) = \frac{\exp\left(-\beta\, F_W(x, y)\right)}{\sum_{y'=1}^{n} \exp\left(-\beta\, F_W(x, y')\right)}.$$

2.2 ii

Negative Log Likelihood loss is: $-log(p(y|x)$

$$\ell(x, y; W) = -log(\frac{\exp\left(-\beta\, F_W(x, y)\right)}{\sum_{y'=1}^{n} \exp\left(-\beta\, F_W(x, y')\right)})$$

$$= \beta F_W(x, y) + log(\sum_{y'=1}^{n} \exp\left(-\beta\, F_W(x, y')\right)$$

multiply by $1/\beta$

$$= F_W(x, y) + \frac{1}{\beta} log(\sum_{y'=1}^{n} \exp\left(-\beta\, F_W(x, y')\right)$$

2.3 iii

$$\nabla_W L(x, y; W) = \nabla_W F_W(x, y) + \frac{1}{\beta}\, \nabla_W \log\Big(Z(x)\Big), \qquad Z(x) = \sum_{y'=1}^{n} \exp\left(-\beta\, F_W(x, y')\right).$$

$$\nabla_W \log\Big(Z(x)\Big) = \frac{1}{Z(x)} \nabla_W Z(x), \quad \nabla_W Z(x) = \sum_{y'=1}^{n} \nabla_W \exp\left(-\beta\, F_W(x, y')\right)$$

$$= -\beta \sum_{y'=1}^{n} \exp\left(-\beta\, F_W(x, y')\right)\, \nabla_W F_W(x, y').$$

Thus,

$$\nabla_W \log\Big(Z(x)\Big) = -\beta \sum_{y'=1}^{n} \frac{\exp\left(-\beta\, F_W(x, y')\right)}{Z(x)}\, \nabla_W F_W(x, y') = -\beta \sum_{y'=1}^{n} p(y'|x)\, \nabla_W F_W(x, y').$$

$$\nabla_W L(x, y; W) = \nabla_W F_W(x, y) - \sum_{y'=1}^{n} p(y'|x)\, \nabla_W F_W(x, y').$$

2.4 iv

The reason negative log-likelihood loss pushes the energy of the correct example to negative infinity, and all others to positive infinity, no matter how close the two examples are is because the loss function is trying to make the probability of the correct label as close to one as possible. To achieve this, the correct energy must be so low that its exponential term dominates the other terms. The only way to have the fraction $p(x|y) = \frac{exp(-\beta F_w(x,y))}{\int_{y'} exp(-\beta F_w(x,y'))dy} \approx 1$ is if the exponential term is very large, i.e. the energy for the correct label goes to negative infinity and all the other energies go to positive infinity, making their exponential terms vanish in turn. Because the exponential function exaggerates even small differences in energy, the gradient forces a massive gap between correct and incorrect labels. This gives us an energy surface with very sharp edges.

3 Comparing contrastive loss functions

3.1 a

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

$$\frac{\partial \ell}{\partial W} = I_{F(x,y)>0} \frac{\partial F(x,y)}{\partial W} - I_{F(x,y)<m} \frac{\partial F(x,y)}{\partial W}$$

3.2 b

$$\ell_{hinge} = [m + F(x,y) - F(x,\bar{y})]^+$$

$$\frac{\partial \ell}{\partial W} = I_{F(x,y)-F(x,y)+m>0} * (\frac{\partial F(x,y)}{\partial W} - \frac{\partial F(x,\bar{y})}{\partial W})$$

3.3 c

$$\ell_{log} = log(1 + exp((F(x,y) - F(x,\bar{y})))$$

Let c = F(x,y) - $F(x,\bar{y})$

$$\frac{\partial \ell}{\partial W} = \frac{d}{dc}(1 + exp(c)) * \frac{d}{d(W)}c$$

$$\frac{d}{d(W)}c = \frac{\partial F(x,y)}{\partial W} - \frac{\partial F(x,\bar{y})}{\partial W}$$

$$\frac{d}{dc}(1 + exp(c)) = \frac{exp(c)}{1 + exp(c)}$$

$$\frac{\partial \ell}{\partial W} = \frac{exp((F(x,y) - F(x,\bar{y})}{1 + exp((F(x,y) - F(x,\bar{y}))} * (\frac{\partial F(x,y)}{\partial W} - \frac{\partial F(x,\bar{y})}{\partial W})$$

4 d

$$\ell_{square-square} = ([F(x,y)]^+)^2 + ([m - F(x,y)]^+)^2$$

$$\frac{\partial \ell}{\partial W} = 2 * [F(x,y)]^+ * \frac{\partial F(x,y)}{\partial W} - 2 * [m - F(x,y)]^+ * \frac{\partial F(x,y)}{\partial W}$$

4.1 e

4.1.1 i

NLL is a probabilistic loss that sums over all possible classes of y and pushes and pulls energies to neg and positive infinity, whereas these three compare correct and incorrect outputs of y and push or pull the energies by some margin/difference

4.1.2 ii

The log-loss is sometimes called a soft hinge loss because log-loss is a smooth approximation for the hinge loss function. This is beneficial in that log-loss is differentiable everywhere which means we have a smooth gradient and thus convergence can be more attainable

4.1.3 iii

The simple and square-square loss functions seperate the positive and negative energies and make distinct pushes and pulls whereas the hinge or log loss use a singular difference to make the choice

4.1.4 iv

Simple loss stops penalizing as soon as it's within the margin so if were ok with a constant amount of seperation then we should use simple loss, Square-square continues to penalize but less severely as you approach the margin but it is harder to tune so there is a tradeoff there.