Homework 3: Theory

Yuxiang Chai

October 25, 2022

1.1 Energy Based Models Intuition

(a) Because instead of generating explicit result, EBM can generate implicit results, where several y can be compatible of one x

(b) Probabilistic model is a special case of energy based model. Energy based models are more flexible of choosing the loss function.

(c) We can use the Gibbs-Boltzmann Equation:

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

(d) Energy functions are used to do the inference. The lower the energy, the higher the compatibility of y and x. Loss functions are used to train the model, and it will press down the correct input and pull up the wrong input.

(e) Only pushing down the correct energy will cause the energy space flat, which can't give a distinguished output. We can use contrastive method and add wrong samples.

(f)
i. Maximum Likelihood. Use negative logarithm and Gibbs-Boltzmann equation to form a distribution to push down the certain probability.
ii. Denoising Auto-Encoder. Reconstruct samples to pull up the wrong energy.
iii. Latent-Variable EBM. Find the minimum of both y and z

(g)

$$\ell_{hingepairloss}(x, y, \hat{y}, W) = [F_W(x, y) - F_W(x, \hat{y}) + m(y, \hat{y})]^+$$

(h) $\check{y} = argmin_y F(x, y)$.
$\check{z} = argmin_z G(x, y, z)$ and $\check{y}$ is the result of decoding $\check{z}$.

1.2 Negative log-likelihood loss

(i)

$$p(y|x) = \frac{e^{-\beta F_W(x,y)}}{\sum_{y'=1}^{n} e^{-\beta F_W(x,y')}}$$

(ii)

$$\begin{aligned}
\ell &= -\frac{1}{\beta}\log(p(y|x)) \\
&= -\frac{1}{\beta}\log\frac{e^{-\beta F_W(x,y)}}{\sum_{y'=1}^{n} e^{-\beta F_W(x,y')}} \\
&= -\frac{1}{\beta}\log e^{-\beta F_W(x,y)} + \frac{1}{\beta}\log\sum_{y'=1}^{n} e^{-\beta F_W(x,y')} \\
&= F_W(x,y) + \frac{1}{\beta}\log\sum_{y'=1}^{n} e^{-\beta F_W(x,y')}
\end{aligned}$$

(iii)

$$\begin{aligned}
\frac{\partial \ell}{\partial W} &= \frac{\partial \left[F_W(x,y) + \frac{1}{\beta} \log \sum_{y'=1}^{n} e^{-\beta F_W(x,y')} \right]}{\partial W} \\
&= \frac{\partial F_W(x,y)}{\partial W} + \frac{\partial \left[\frac{1}{\beta} \log \sum_{y'=1}^{n} e^{-\beta F_W(x,y')} \right]}{\partial W} \\
&= \frac{\partial F_W(x,y)}{\partial W} + \frac{1}{\beta \sum_{y'=1}^{n} e^{-\beta F_W(x,y')}} \frac{\partial \sum_{y'=1}^{n} e^{-\beta F_W(x,y')}}{\partial W} \\
&= \frac{\partial F_W(x,y)}{\partial W} + \frac{1}{\beta \sum_{y'=1}^{n} e^{-\beta F_W(x,y')}} \sum_{y'=1}^{n} \frac{\partial e^{-\beta F_W(x,y')}}{\partial W} \\
&= \frac{\partial F_W(x,y)}{\partial W} + \frac{1}{\beta \sum_{y'=1}^{n} e^{-\beta F_W(x,y')}} \sum_{y'=1}^{n} (-\beta e_W^F(x,y')) \frac{\partial F_W(x,y')}{\partial W} \\
&= \frac{\partial F_W(x,y)}{\partial W} - \sum_{y'=1}^{n} \frac{-\beta e_W^F(x,y')}{\beta \sum_{y'=1}^{n} e^{-\beta F_W(x,y')}} \frac{\partial F_W(x,y')}{\partial W} \\
&= \frac{\partial F_W(x,y)}{\partial W} - \sum_{y'=1}^{n} p(y'|x) \frac{\partial F_W(x,y')}{\partial W}
\end{aligned}$$

if the label is continuous, then the equation would be

$$\frac{\partial \ell}{\partial W} == \frac{\partial F_W(x,y)}{\partial W} - \int_{y'} p(y'|x) \frac{\partial F_W(x,y')}{\partial W}$$

And the second term is an integration, which can't be computed in real world. So it may be intractable. We can use Monte Carlo methods to randomly generate sample to make it tractable.

(iv) For continuous label,

$$\ell = F_W(x,y) + \frac{1}{\beta} \log \int_{y'} e^{-\beta F_W(x,y')}$$

if y is correct, then the first term will have way larger impact than the second term on y, but the second term will pull up the energies of wrong y'. To make the loss smaller, $F_W(x,y)$ has to be pushed down to negative infinity and all other $F_W(x,y')$ should be pulled up to infinity so that the loss will be extremely small.

1.3 Comparing Contrastive Loss Functions

(a)

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

and for the first term

$$\frac{\partial [F_W(x,y)]^+}{\partial W} = \begin{cases} 0 & F_W(x,y) \le 0 \\ \frac{\partial F_W(x,y)}{\partial W} & F_W(x,y) > 0 \end{cases}$$

For the second term

$$\frac{\partial [m - F_W(x,\bar{y})]^+}{\partial W} = \begin{cases} 0 & F_W(x,\bar{y}) \ge m \\ -\frac{\partial F_W(x,\bar{y})}{\partial W} & F_W(x,y) < m \end{cases}$$

So add them together we get

$$\frac{\partial \ell_{simple}}{\partial W} = \begin{cases} 0 & F_W(x,\bar{y}) \geq m \text{ and } F_W(x,y) \leq 0 \\ -\frac{\partial F_W(x,\bar{y})}{\partial W} & F_W(x,\bar{y}) < m \text{ and } F_W(x,y) \leq 0 \\ \frac{\partial F_W(x,y)}{\partial W} & F_W(x,\bar{y}) \geq m \text{ and } F_W(x,y) > 0 \\ \frac{\partial F_W(x,y)}{\partial W} - \frac{\partial F_W(x,\bar{y})}{\partial W} & F_W(x,\bar{y}) < m \text{ and } F_W(x,y) > 0 \end{cases}$$

(b)

$$\begin{aligned} \frac{\partial \ell_{log}}{\partial W} &= \frac{\partial \log(1 + e^{F_W(x,y) - F_W(x,\bar{y})})}{\partial W} \\ &= \frac{1}{1 + e^{F_W(x,y) - F_W(x,\bar{y})}} \frac{\partial (1 + e^{F_W(x,y) - F_W(x,\bar{y})})}{\partial W} \\ &= \frac{e^{F_W(x,y) - F_W(x,\bar{y})}}{1 + e^{F_W(x,y) - F_W(x,\bar{y})}} \frac{\partial (F_W(x,y) - F_W(x,\bar{y}))}{\partial W} \\ &= \frac{e^{F_W(x,y) - F_W(x,\bar{y})}}{1 + e^{F_W(x,y) - F_W(x,\bar{y})}} (\frac{\partial F_W(x,y)}{\partial W} - \frac{\partial F_W(x,\bar{y})}{\partial W}) \end{aligned}$$

(c)

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

and for the first term

$$\frac{\partial ([F_W(x,y)]^+)^2}{\partial W} = \begin{cases} 0 & F_W(x,y) \leq 0 \\ 2\frac{\partial F_W(x,y)}{\partial W} & F_W(x,y) > 0 \end{cases}$$

For the second term

$$\frac{\partial ([m - F_W(x,\bar{y})]^+)^2}{\partial W} = \begin{cases} 0 & F_W(x,\bar{y}) \geq m \\ -2\frac{\partial F_W(x,\bar{y})}{\partial W} & F_W(x,\bar{y}) < m \end{cases}$$

So add them together we get

$$\frac{\partial \ell_{square-square}}{\partial W} = \begin{cases} 0 & F_W(x,\bar{y}) \geq m \text{ and } F_W(x,y) \leq 0 \\ -2\frac{\partial F_W(x,\bar{y})}{\partial W} & F_W(x,\bar{y}) < m \text{ and } F_W(x,y) \leq 0 \\ 2\frac{\partial F_W(x,y)}{\partial W} & F_W(x,\bar{y}) \geq m \text{ and } F_W(x,y) > 0 \\ 2\frac{\partial F_W(x,y)}{\partial W} - 2\frac{\partial F_W(x,\bar{y})}{\partial W} & F_W(x,\bar{y}) < m \text{ and } F_W(x,y) > 0 \end{cases}$$

(d) (i) NLL loss doesn't depend on margin, comparing with simple and square-square. And NLL can use many wrong $\bar{y}$s, comparing with log loss.

(ii) When the margin is infinity, we can treat the log loss as the soft version of the hinge loss. The advantage is that soft hinge loss is more smooth and the gradient is more smooth.

(iii) Margin in simple and square-square loss only participates in the wrong label, but the margin in the hinge/log loss is determined by both correct and wrong labels. Simple loss is used when the force of pushing down and pulling up is weak. And square-square loss is used when the force of pushing down and pulling up is strong.