Energy-Based Models

Notes from Yann Lecun's lectures

Energy function

The energy function $F(x, y)$ measures the incompatibility between variables x and y. The lower the energy, the better the match between x and y.

EBMs provide more flexibility in the choice of the objective function. While we can set our loss function to be identical to the energy function, we can also choose something else.

EBM: unconditional version

Conditional EBM: $F(x, y)$

Unconditional EBM: $F(y)$

- measures the incompatibility between the components of y

EBM Inference

For prediction, we find the $\hat{y}$ that minimizes the energy: $\hat{y} = \arg\min_y F(x, y)$.

To estimate the distribution, we use the Boltzmann distribution: $P(y|x) = \frac{e^{-\beta F(x,y)}}{\int_{y'} e^{-\beta F(x,y')}}$, where β is a constant related to temperature in the physical system analogy.

EBM Training

How do we design the loss to prevent collapse?

1. Contrastive methods
    - Push down on energy of training samples
    - Pull up on energy of suitably-generated contrastive samples
    - Examples: $m \in \mathbb{R}$ is a margin
        - $l_{\text{simple}}(x, y, \bar{y}, W) = [F_W(x, y)]^+ + [m - F_W(x, \bar{y})]^+$
        - $l_{\log}(x, y, \bar{y}, W) = \log(1 + e^{F_W(x,y) - F_W(x,\bar{y})})$
        - $l_{\text{square-square}}(x, y, \bar{y}, W) = ([F_W(x, y)]^+)^2 + ([m - F_W(x, \bar{y})]^+)^2$
2. Regularized & architectural methods
    - Push down on energy of training samples
    - Regularizer minimizes the volume of space that can take low energy

Contrastive methods scale very badly with dimension. When y is in a high-dimensional space, it may require a very large number of contrastive samples to ensure that the energy is higher in all dimensions unoccupied by the local data distribution.

Therefore, regularized methods are much more promising in the long run.

Latent Variable EBMs

- Latent variable z: captures the information in y that is not available in x
- Computed by minimizing the energy function: $\hat{z} = \arg\min_z E_w(x, y, z)$

  then we have $F_w(x, y) = E_w(x, y, \hat{z})$ and $P(y|x) = \frac{\int_{z'} e^{-\beta E_w(x,y,z')}}{\int_{y'} \int_{z'} e^{-\beta E_w(x,y',z')}}$