Energy-based models

Energy-based models for distributions

- Distribution as: $p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\int_{\boldsymbol{x}} g_{\boldsymbol{\theta}}(\boldsymbol{x}) d\boldsymbol{x}} g_{\boldsymbol{\theta}}(\boldsymbol{x})$

- $p_{\boldsymbol{\theta}}$ as known probability distributions (Gaussian, exp.) can be restrictive
  - Maybe I want to encode domain knowledge of how variables interact

- We can also define an energy function and divide by its volume

$$g_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp(f_{\boldsymbol{\theta}}(\boldsymbol{x})) \Rightarrow p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})} \exp(f_{\boldsymbol{\theta}}(\boldsymbol{x}))$$

Energy-based models for distributions

$$g_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp(f_{\boldsymbol{\theta}}(\boldsymbol{x})) \Rightarrow p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})} \exp(f_{\boldsymbol{\theta}}(\boldsymbol{x}))$$

- $-f_{\boldsymbol{\theta}}(\boldsymbol{x})$ is the energy function
- Partition function is the hard bit

$$Z(\boldsymbol{\theta}) = \int_{\boldsymbol{x}} \exp(f_{\boldsymbol{\theta}}(\boldsymbol{x}))\, d\boldsymbol{x}$$

  - Note the multi-dimensional integral due to $\boldsymbol{x}$

Why exponential?

- Why $g_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp(f_{\boldsymbol{\theta}}(\boldsymbol{x}))$ and not $g_{\boldsymbol{\theta}}(\boldsymbol{x}) = f_{\boldsymbol{\theta}}^2(\boldsymbol{x})$?

- Couples well with maximum likelihood and natural logarithms
- Many existing distributions are exponential-based
- They arise often in statistical physics → Good inspiration

Advantages & disadvantages

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})}\exp(f_{\boldsymbol{\theta}}(\boldsymbol{x}))$$

- Very flexible in defining our energy function
- Sampling from $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ can be very hard
  - The CDF introduces another integral
- Evaluating and optimizing likelihood can be hard $\Rightarrow$ Learning is hard
  - Must be able to compute the partition function
- In vanilla case no latent variables $\Rightarrow$ no representation learning
  - Latent variables can be added though

Ratio of likelihoods

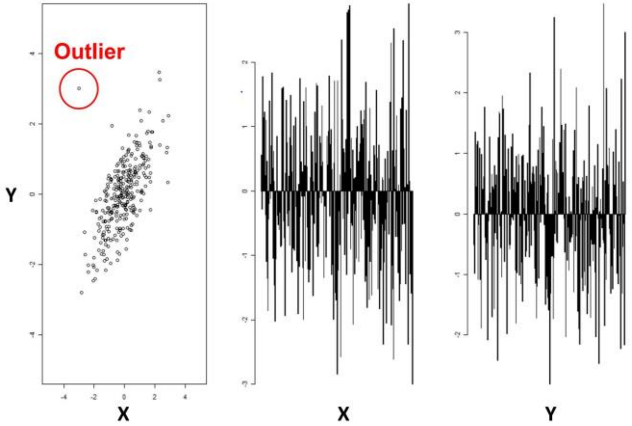
- The partition function is often very hard to compute analytically
- But if we have pairs of inputs

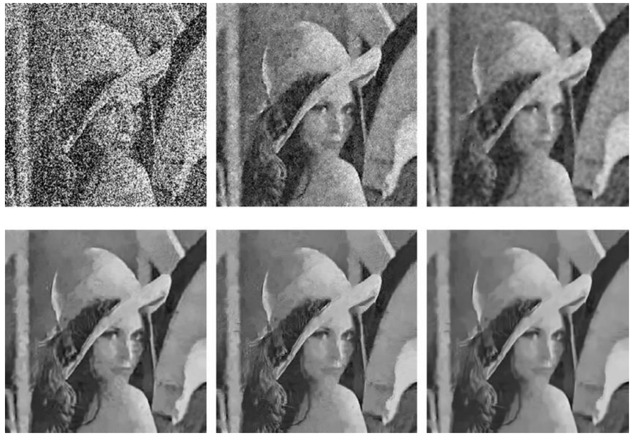
$$\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x_a})}{p_{\boldsymbol{\theta}}(\boldsymbol{x_b})} = \exp(f_{\boldsymbol{\theta}}(\boldsymbol{x}_a) - f_{\boldsymbol{\theta}}(\boldsymbol{x}_b))$$

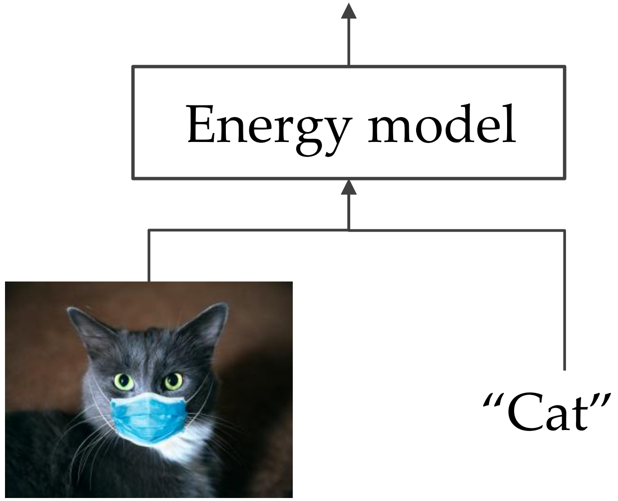
- No partition function anymore

Applications

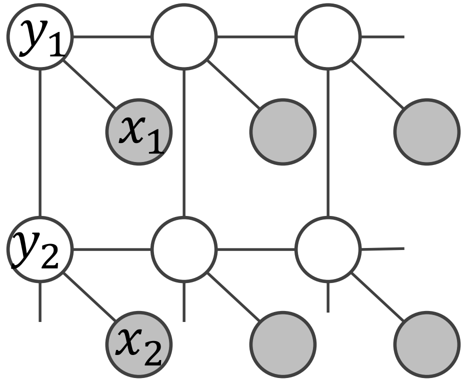
- Given trained model
  - Anomaly detection
  - Denoising & restoration
  - Classification







Examples of energy models



- Ising model

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{x}) = \frac{1}{Z} \exp\Big(\sum_i \psi_i\,(x_i, y_i) + \sum_{i,j \in E} \psi_{ij}\,(y_i, y_j)\Big)$$

- Product of experts (similar to AND)

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\omega})} q_\theta(\boldsymbol{x}) r_\varphi(\boldsymbol{x}) s_\omega(\boldsymbol{x})$$

- Hopfield networks
- Boltzmann machines
- Deep belief networks