1. Entropy, Cross-Entropy, Kullback-Leibler (KL)-divergence

(a) Entropy is a measure of expected surprise. For a given discrete Random variable \( Y \), we know that from Information Theory that a measure the surprise of observing that \( Y \) takes the value \( k \) by computing:

\[
\log \frac{1}{p(Y = k)} = - \log[p(Y = k)]
\]

As given:
- if \( p(Y = k) \to 0 \), the surprise of observing \( k \) approaches \( \infty \)
- if \( p(Y = k) \to 1 \), the surprise of observing \( k \) approaches 0

The Entropy of the distribution of \( Y \) is then the expected surprise given by:

\[
H(Y) = E_Y \left[ - \log(p(Y = k)) \right] = - \sum_k [p(Y = k) \log[p(Y = k)]]
\]

On the other hand, Cross-entropy is a measure building upon entropy, generally calculating the difference between two probability distributions \( p \) and \( q \). It is given by:

\[
H(p, q) = E_p \left[ \frac{1}{\log(q(x))} \right]
\]

\[
= \sum_x \left[ p(x) \log\left( \frac{1}{q(x)} \right) \right]
\]

Relative Entropy also known as KL Divergence measures how much one distribution diverges from another. For two discrete probability distributions, \( p \) and \( q \), it is defined as:

\[
D_{KL}(p||q) = \sum_x \left[ p(x) \log\left( \frac{p(x)}{q(x)} \right) \right]
\]

Let’s define the following probability distributions given by:

\[
p(x) = \begin{cases} 
1 & \text{with probability 0.5} \\
-1 & \text{with probability 0.5}
\end{cases}
\]

\[
q(x) = \begin{cases} 
1 & \text{with probability 0.1} \\
-1 & \text{with probability 0.9}
\end{cases}
\]
Show that KL-divergence is not symmetric and hence does not satisfy some intuitive attributes of distances.

(b) Re-write $D_{KL}(p||q)$ in term of the Entropy $H(p)$ and the cross entropy $H(p, q)$.

(c) Show that KL - divergence is always non-negative using Jensen’s Inequality which states: $E[ \log X] \leq \log E[X]$ and the fact that $\log$ is a concave function.

(d) Knowing that the equality in Jensen’s inequality can only hold if $X$ is a constant random variable, please state when is $D_{KL}(q||p) = 0$.

2. Simple Latent Variable Models

Formally, a latent variable model $p$ is a probability distribution over observed variables $x$ and latent variables $z$ (variables that are not directly observed but inferred), $p_\theta(x, z)$. Because we know $z$ is unobserved, using learning methods learned in class (like supervised learning methods) is unsuitable. Indeed, our learning problem of maximizing the log-likelihood of the data turns from:

$$\theta \leftarrow \arg \max_\theta \frac{1}{N} \sum_{i=1}^{N} \log[p_\theta(x_i)]$$

to:

$$\theta \leftarrow \arg \max_\theta \frac{1}{N} \sum_{i=1}^{N} \log[\int p_\theta(x_i \mid z)p(z)dz]$$

where $p(x)$ has become $\int p_\theta(x_i \mid z)p(z)dz$.

(a) State whether or not we could directly maximize the likelihood above and why?

(b) We define the proxy likelihood given by:

$$\mathcal{L}(x_i, \theta, \phi) = E_{z \sim q(z|x_i)} \left[ \log[p_\theta(x_i \mid z)] \right] - D_{KL}\left[q(z \mid x_i) || p(z)\right]$$

Please show that $\mathcal{L}(x_i, \theta, \phi)$ is always a lower bound to the true log likelihood for $x_i$.

Hint: You can show that something is a lower bound by showing that adding a non-negative term to it gives the original quantity — remember, the KL divergence is always non-negative.

(c) To optimize the Variational Lower Bound derived in the previous problem, which distribution do we sample $z$ from?

(d) To be able to take a derivative through a sampling operation, we need to show how sampling can be done as a deterministic and continuous function of functions of parameters as well as an external independent source of randomness. Otherwise, it is hard to understand how things would change a little bit if the parameters changed a little bit. Such explicit representations of sampling are called "the reparameterization trick" in machine-learning communities. Assume we have a normal distribution for $x$ with both means and variance parameterized by parameters $\theta$ and we would like to solve for:
\[
\min_{\theta} \mathbb{E}_q[x^2]
\]

Assuming that \( \epsilon \) is an independent standard Normal \( \mathcal{N}(0, 1) \) random variable, write \( x \) as a function of \( \epsilon \) and use that to compute the gradient of the objective function above.

(e) Describe step-by-step what happens during a forward pass during VAE training

(f) Describe what the encoder and decoder of the VAE are doing to capture and encode this information into a latent representation of space \( z \).

(g) Once the VAE is trained, how do we use it to generate a new fresh sample from the learned approximation of the data-generating distribution?

Contributors:

- Jerome Quenum.
- Anant Sahai.
- Past CS282 Staff.