Training Variational Autoencoder Models

JPD

2025-06-29

## R Markdown

## Training variational autoencoder models

An ‘autoencoder’ is a type of neural network that consists of an encoder and a decoder network with an information bottleneck layer with latent variables (that is, , where denotes the number of features) in the middle.

It generates an embedding such that the information present in the original space is preserved in this lower-dimensional space. Specifically, the encoder network, defined as:

maps input to latent embedding . Similarly, the decoder network:

maps the embedding back to the input space. We optimize both networks by minimizing the squared -norm between the input and the reconstructed output as follows:

To train the autoencoder, we jointly optimized the encoder and decoder networks by minimizing the reconstruction error between the original input and its reconstruction. Formally, given an input sample , the encoder maps into a lower-dimensional latent representation, and the decoder attempts to reconstruct the original input from this latent embedding. The training objective is to minimize the expected squared Euclidean (L2) norm between the input and its reconstruction:

where and are the learnable parameters of the encoder and decoder, respectively. Here, represents the reconstructed input. The L2 loss term captures the total reconstruction error across all dimensions of the input. Explicitly, this corresponds to:

where and denote the components of the original and reconstructed input vectors, respectively. This loss encourages the model to faithfully capture the essential structure of the data in the latent space, enabling effective data compression and reconstruction.

Samples with 500 principal components (PCs) were used to construct the input matrix , where is the number of samples and is the number of features. This matrix was passed to an encoder , which outputs a mean vector and a variance vector :

A decoder reconstructs the input from the sampled latent vector . The model is trained to minimize the following loss:

The first term ensures accurate reconstruction, while the KL divergence regularizes the latent space by encouraging it to resemble a standard Gaussian distribution. After training, the learned latent variables were used for gene importance analysis using Integrated Gradients, followed by pathway enrichment.