Autoencoders

Autoencoers are neural network that are trained to learn how to map their input to their input. Internally, it has an hidden layer $\boldsymbol{h}$ that contains a lossy summary of the relevant feature for the task.
An autoencoder can be seen has a two parts network

- Encoder function: $\boldsymbol{h} = f(\boldsymbol{x})$
- Decoder function: $\tilde{\boldsymbol{x}} = g(\boldsymbol{h})$

The simplest autoencoder is a MLP:

$$\begin{aligned} \boldsymbol{h} &= \sigma_1 \left(W_{xh} x \right) \\ \tilde{x} &= \sigma_2 \left(W_{hx} \boldsymbol{h} \right) \end{aligned} \tag{1}$$

Autoencoders

- Encoder function: $\boldsymbol{h} = f(\boldsymbol{x})$
- Decoder function: $\tilde{\boldsymbol{x}} = g(\boldsymbol{h})$

The simplest autoencoder is a MLP:

$$\begin{aligned} \boldsymbol{h} &= \sigma_1 \left(W_{xh} x \right) \\ \tilde{x} &= \sigma_2 \left(W_{hx} \boldsymbol{h} \right) \end{aligned} \tag{2}$$