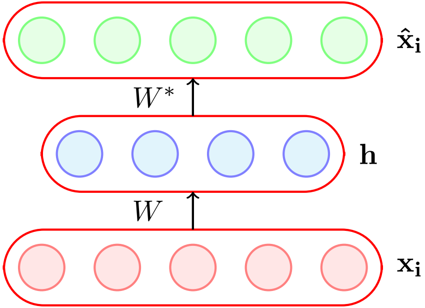
CS7015 (Deep Learning) : Lecture 7

Autoencoders and relation to PCA, Regularization in autoencoders, Denoising autoencoders, Sparse autoencoders, Contractive autoencoders

Mitesh M. Khapra

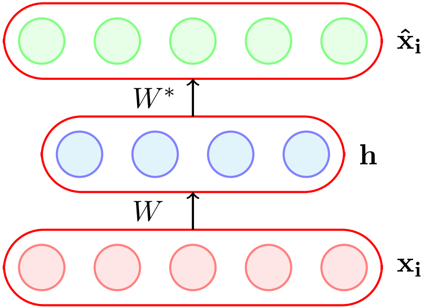
Department of Computer Science and Engineering
Indian Institute of Technology Madras

Module 7.1: Introduction to Autoencoders



$$\mathbf{h} = g(W\mathbf{x_i} + \mathbf{b})$$
$$\hat{\mathbf{x}}_\mathbf{i} = f(W^*\mathbf{h} + \mathbf{c})$$

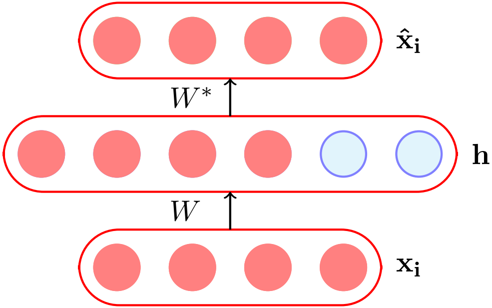
- An autoencoder is a special type of feed forward neural network which does the following
- <u>Encodes</u> its input $\mathbf{x_i}$ into a hidden representation $\mathbf{h}$
- <u>Decodes</u> the input again from this hidden representation
- The model is trained to minimize a certain loss function which will ensure that $\hat{\mathbf{x}}_\mathbf{i}$ is close to $\mathbf{x_i}$ (we will see some such loss functions soon)



$$\mathbf{h} = g(W\mathbf{x}_\mathbf{i} + \mathbf{b})$$
$$\hat{\mathbf{x}}_\mathbf{i} = f(W^*\mathbf{h} + \mathbf{c})$$

An autoencoder where $\dim(\mathbf{h}) < \dim(\mathbf{x}_\mathbf{i})$ is called an under complete autoencoder

- Let us consider the case where $\dim(\mathbf{h}) < \dim(\mathbf{x}_\mathbf{i})$
- If we are still able to reconstruct $\hat{\mathbf{x}}_\mathbf{i}$ perfectly from $\mathbf{h}$, then what does it say about $\mathbf{h}$?
- $\mathbf{h}$ is a loss-free encoding of $\mathbf{x}_\mathbf{i}$. It captures all the important characteristics of $\mathbf{x}_\mathbf{i}$
- Do you see an analogy with PCA?



$$\mathbf{h} = g(W\mathbf{x}_\mathbf{i} + \mathbf{b})$$
$$\hat{\mathbf{x}}_\mathbf{i} = f(W^*\mathbf{h} + \mathbf{c})$$

An autoencoder where $\dim(\mathbf{h}) \geq \dim(\mathbf{x}_\mathbf{i})$ is called an over complete autoencoder

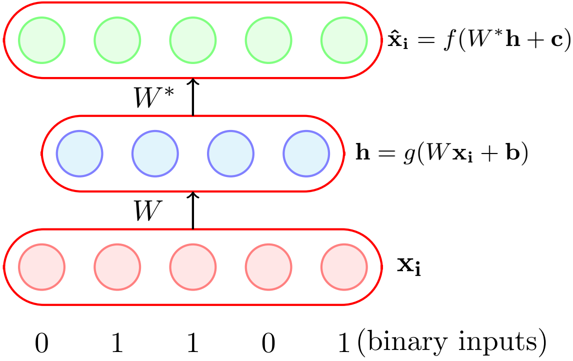
- Let us consider the case when $\dim(\mathbf{h}) \geq \dim(\mathbf{x}_\mathbf{i})$
- In such a case the autoencoder could learn a trivial encoding by simply copying $\mathbf{x}_\mathbf{i}$ into $\mathbf{h}$ and then copying $\mathbf{h}$ into $\hat{\mathbf{x}}_\mathbf{i}$
- Such an identity encoding is useless in practice as it does not really tell us anything about the important characteristics of the data

The Road Ahead

- Choice of $f(\mathbf{x_i})$ and $g(\mathbf{x_i})$
- Choice of loss function

The Road Ahead

- Choice of $f(\mathbf{x_i})$ and $g(\mathbf{x_i})$
- Choice of loss function

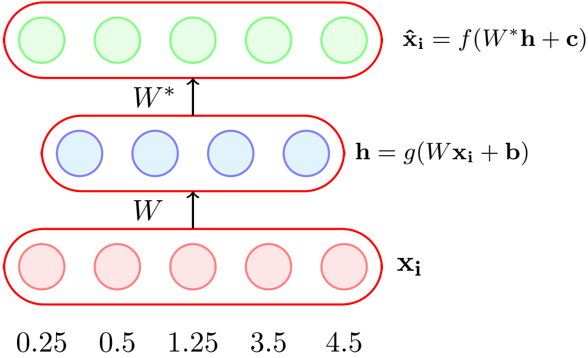


g is typically chosen as the sigmoid function

- Suppose all our inputs are binary (each $x_{ij} \in \{0, 1\}$)
- Which of the following functions would be most apt for the decoder?

$$\hat{\mathbf{x}}_\mathbf{i} = \tanh(W^*\mathbf{h} + \mathbf{c})$$
$$\hat{\mathbf{x}}_\mathbf{i} = W^*\mathbf{h} + \mathbf{c}$$
$$\hat{\mathbf{x}}_\mathbf{i} = logistic(W^*\mathbf{h} + \mathbf{c})$$

- Logistic as it naturally restricts all outputs to be between 0 and 1



(real valued inputs)

Again, g is typically chosen as the sigmoid function

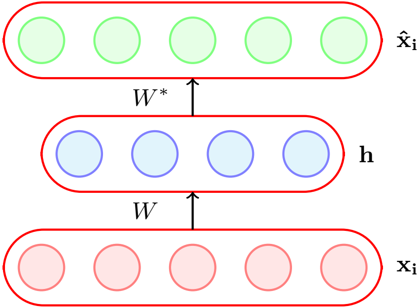
- Suppose all our inputs are real (each $x_{ij} \in \mathbb{R}$)
- Which of the following functions would be most apt for the decoder?

$$\hat{\mathbf{x}}_\mathbf{i} = \tanh(W^*\mathbf{h} + \mathbf{c})$$
$$\hat{\mathbf{x}}_\mathbf{i} = W^*\mathbf{h} + \mathbf{c}$$
$$\hat{\mathbf{x}}_\mathbf{i} = \text{logistic}(W^*\mathbf{h} + \mathbf{c})$$

- What will logistic and tanh do?
- They will restrict the reconstructed $\hat{\mathbf{x}}_\mathbf{i}$ to lie between [0,1] or [-1,1] whereas we want $\hat{\mathbf{x}}_\mathbf{i} \in \mathbb{R}^n$

The Road Ahead

- Choice of $f(\mathbf{x_i})$ and $g(\mathbf{x_i})$
- Choice of loss function



$$\mathbf{h} = g(W\mathbf{x}_\mathbf{i} + \mathbf{b})$$
$$\hat{\mathbf{x}}_\mathbf{i} = f(W^*\mathbf{h} + \mathbf{c})$$

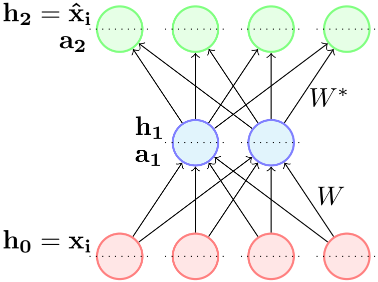
- Consider the case when the inputs are real valued
- The objective of the autoencoder is to reconstruct $\hat{\mathbf{x}}_\mathbf{i}$ to be as close to $\mathbf{x}_\mathbf{i}$ as possible
- This can be formalized using the following objective function:

$$\min_{W,W^*,\mathbf{c},\mathbf{b}} \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}(\hat{x}_{ij} - x_{ij})^2$$

$$i.e., \min_{W,W^*,\mathbf{c},\mathbf{b}} \frac{1}{m}\sum_{i=1}^{m}(\hat{\mathbf{x}}_\mathbf{i} - \mathbf{x}_\mathbf{i})^T(\hat{\mathbf{x}}_\mathbf{i} - \mathbf{x}_\mathbf{i})$$

- We can then train the autoencoder just like a regular feedforward network using backpropagation
- All we need is a formula for $\frac{\partial \mathscr{L}(\theta)}{\partial W^*}$ and $\frac{\partial \mathscr{L}(\theta)}{\partial W}$ which we will see now

$$\mathscr{L}(\theta) = (\hat{\mathbf{x}}_\mathbf{i} - \mathbf{x}_\mathbf{i})^T(\hat{\mathbf{x}}_\mathbf{i} - \mathbf{x}_\mathbf{i})$$



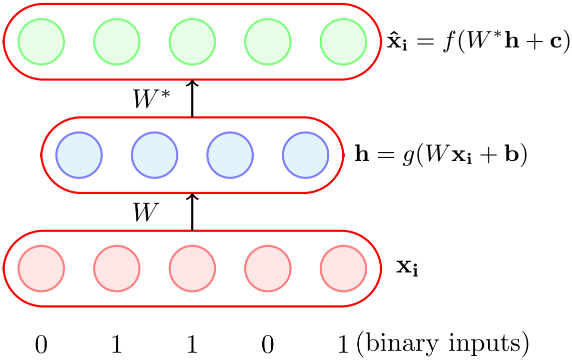
- Note that the loss function is shown for only one training example.

- $$\frac{\partial \mathscr{L}(\theta)}{\partial W^*} = \frac{\partial \mathscr{L}(\theta)}{\partial \mathbf{h_2}} \boxed{\frac{\partial \mathbf{h_2}}{\partial \mathbf{a_2}} \frac{\partial \mathbf{a_2}}{\partial W^*}}$$

- $$\frac{\partial \mathscr{L}(\theta)}{\partial W} = \frac{\partial \mathscr{L}(\theta)}{\partial \mathbf{h_2}} \boxed{\frac{\partial \mathbf{h_2}}{\partial \mathbf{a_2}} \frac{\partial \mathbf{a_2}}{\partial \mathbf{h_1}} \frac{\partial \mathbf{h_1}}{\partial \mathbf{a_1}} \frac{\partial \mathbf{a_1}}{\partial W}}$$

- We have already seen how to calculate the expression in the boxes when we learnt backpropagation

$$\begin{aligned}
\frac{\partial \mathscr{L}(\theta)}{\partial \mathbf{h_2}} &= \frac{\partial \mathscr{L}(\theta)}{\partial \hat{\mathbf{x}}_\mathbf{i}} \\
&= \nabla_{\hat{\mathbf{x}}_\mathbf{i}} \{(\hat{\mathbf{x}}_\mathbf{i} - \mathbf{x}_\mathbf{i})^T(\hat{\mathbf{x}}_\mathbf{i} - \mathbf{x}_\mathbf{i})\} \\
&= 2(\hat{\mathbf{x}}_\mathbf{i} - \mathbf{x}_\mathbf{i})
\end{aligned}$$



What value of $\hat{x}_{ij}$ will minimize this function?

- If $x_{ij} = 1$?
- If $x_{ij} = 0$?

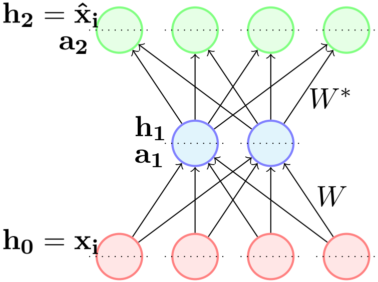
Indeed the above function will be minimized when $\hat{x}_{ij} = x_{ij}$!

- Consider the case when the inputs are binary
- We use a sigmoid decoder which will produce outputs between 0 and 1, and can be interpreted as probabilities.
- For a single n-dimensional i^{th} input we can use the following loss function

$$\min\{-\sum_{j=1}^{n}(x_{ij}\log\hat{x}_{ij} + (1 - x_{ij})\log(1 - \hat{x}_{ij}))\}$$

- Again we need is a formula for $\frac{\partial\mathscr{L}(\theta)}{\partial W^*}$ and $\frac{\partial\mathscr{L}(\theta)}{\partial W}$ to use backpropagation

$$\mathscr{L}(\theta) = -\sum_{j=1}^{n}(x_{ij}\log\hat{x}_{ij} + (1-x_{ij})\log(1-\hat{x}_{ij}))$$



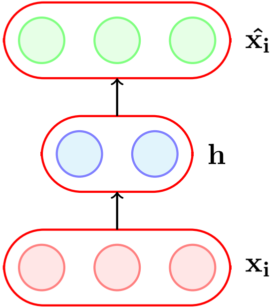
$$\frac{\partial \mathscr{L}(\theta)}{\partial \mathbf{h_2}} = \begin{pmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{21}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{22}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{2n}} \end{pmatrix}$$

- $$\frac{\partial \mathscr{L}(\theta)}{\partial W^*} = \frac{\partial \mathscr{L}(\theta)}{\partial \mathbf{h_2}}\frac{\partial \mathbf{h_2}}{\partial \mathbf{a_2}}\boxed{\frac{\partial \mathbf{a_2}}{\partial W^*}}$$
- $$\frac{\partial \mathscr{L}(\theta)}{\partial W} = \frac{\partial \mathscr{L}(\theta)}{\partial \mathbf{h_2}}\frac{\partial \mathbf{h_2}}{\partial \mathbf{a_2}}\boxed{\frac{\partial \mathbf{a_2}}{\partial \mathbf{h_1}}\frac{\partial \mathbf{h_1}}{\partial \mathbf{a_1}}\frac{\partial \mathbf{a_1}}{\partial W}}$$
- We have already seen how to calculate the expressions in the square boxes when we learnt BP
- The first two terms on RHS can be computed as:

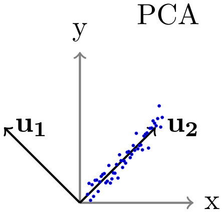
$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{2j}} = -\frac{x_{ij}}{\hat{x}_{ij}} + \frac{1-x_{ij}}{1-\hat{x}_{ij}}$$

$$\frac{\partial h_{2j}}{\partial a_{2j}} = \sigma(a_{2j})(1-\sigma(a_{2j}))$$

Module 7.2: Link between PCA and Autoencoders



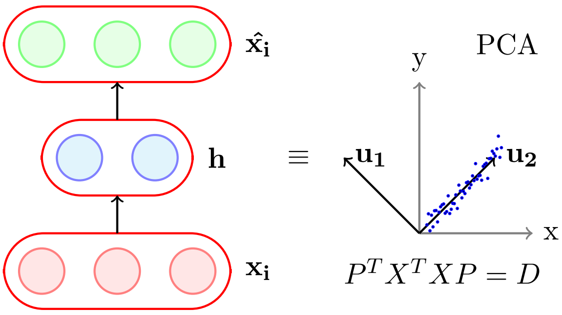
$\equiv$



$$P^T X^T X P = D$$

- We will now see that the encoder part of an autoencoder is equivalent to PCA if we
  - use a linear encoder
  - use a linear decoder
  - use squared error loss function
  - normalize the inputs to

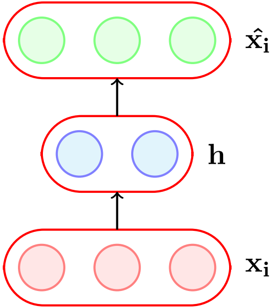
$$\hat{x}_{ij} = \frac{1}{\sqrt{m}}\left(x_{ij} - \frac{1}{m}\sum_{k=1}^{m} x_{kj}\right)$$



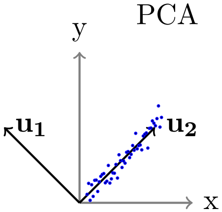
- First let us consider the implication of normalizing the inputs to

$$\hat{x}_{ij} = \frac{1}{\sqrt{m}}\left(x_{ij} - \frac{1}{m}\sum_{k=1}^{m} x_{kj}\right)$$

- The operation in the bracket ensures that the data now has 0 mean along each dimension j (we are subtracting the mean)
- Let X' be this zero mean data matrix then what the above normalization gives us is $X = \frac{1}{\sqrt{m}}X'$
- Now $(X)^TX = \frac{1}{m}(X')^TX'$ is the covariance matrix (recall that covariance matrix plays an important role in PCA)



$\equiv$



$$P^T X^T X P = D$$

- First we will show that if we use linear decoder and a squared error loss function then
- The optimal solution to the following objective function

$$\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \hat{x}_{ij})^2$$

is obtained when we use a linear encoder.

$$\min_{\theta} \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \hat{x}_{ij})^2 \tag{1}$$

- This is equivalent to

$$\min_{W^*H} (\|X - HW^*\|_F)^2 \qquad\qquad \|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$$

(just writing the expression (1) in matrix form and using the definition of $||A||_F$) (we are ignoring the biases)

- From SVD we know that optimal solution to the above problem is given by

$$HW^* = U_{.,\leq k} \Sigma_{k,k} V_{.,\leq k}^T$$

- By matching variables one possible solution is

$$H = U_{.,\leq k} \Sigma_{k,k}$$
$$W^* = V_{.,\leq k}^T$$

We will now show that H is a linear encoding and find an expression for the encoder weights W

$$
\begin{aligned}
H &= U_{.,\leq k}\Sigma_{k,k} \\
&= (XX^T)(XX^T)^{-1}U_{.,\leq K}\Sigma_{k,k} && \textit{(pre-multiplying } (XX^T)(XX^T)^{-1} = I\textit{)} \\
&= (XV\Sigma^T U^T)(U\Sigma V^T V\Sigma^T U^T)^{-1}U_{.,\leq k}\Sigma_{k,k} && \textit{(using } X = U\Sigma V^T\textit{)} \\
&= XV\Sigma^T U^T(U\Sigma\Sigma^T U^T)^{-1}U_{.,\leq k}\Sigma_{k,k} && (V^TV = I) \\
&= XV\Sigma^T U^T U(\Sigma\Sigma^T)^{-1}U^T U_{.,\leq k}\Sigma_{k,k} && ((ABC)^{-1} = C^{-1}B^{-1}A^{-1}) \\
&= XV\Sigma^T(\Sigma\Sigma^T)^{-1}U^T U_{.,\leq k}\Sigma_{k,k} && (U^TU = I) \\
&= XV\Sigma^T\Sigma^{T^{-1}}\Sigma^{-1}U^T U_{.,\leq k}\Sigma_{k,k} && ((AB)^{-1} = B^{-1}A^{-1}) \\
&= XV\Sigma^{-1}I_{.,\leq k}\Sigma_{k,k} && (U^TU_{.,\leq k} = I_{.,\leq k}) \\
&= XVI_{.,\leq k} && (\Sigma^{-1}I_{.,\leq k} = \Sigma^{-1}_{k,k}) \\
H &= XV_{.,\leq k}
\end{aligned}
$$

Thus H is a linear transformation of X and $W = V_{.,\leq k}$

- We have encoder $W = V_{.,\leq k}$
- From SVD, we know that V is the matrix of eigen vectors of X^TX
- From PCA, we know that P is the matrix of the eigen vectors of the covariance matrix
- We saw earlier that, if entries of X are normalized by

$$\hat{x}_{ij} = \frac{1}{\sqrt{m}}\left(x_{ij} - \frac{1}{m}\sum_{k=1}^{m} x_{kj}\right)$$

then X^TX is indeed the covariance matrix

- Thus, the encoder matrix for linear autoencoder(W) and the projection matrix(P) for PCA could indeed be the same. Hence proved

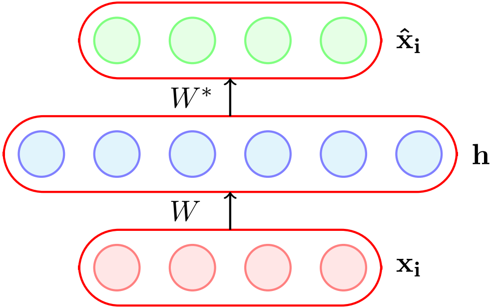
Remember

The encoder of a linear autoencoder is equivalent to PCA if we

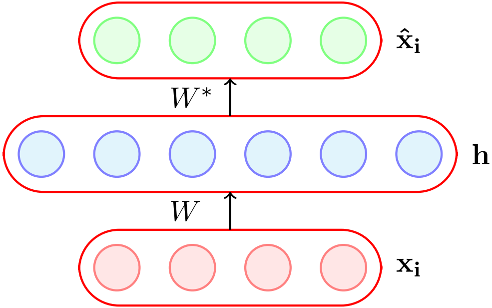
- use a linear encoder
- use a linear decoder
- use a squared error loss function
- and normalize the inputs to

$$\hat{x}_{ij} = \frac{1}{\sqrt{m}}\left(x_{ij} - \frac{1}{m}\sum_{k=1}^{m} x_{kj}\right)$$

Module 7.3: Regularization in autoencoders (Motivation)



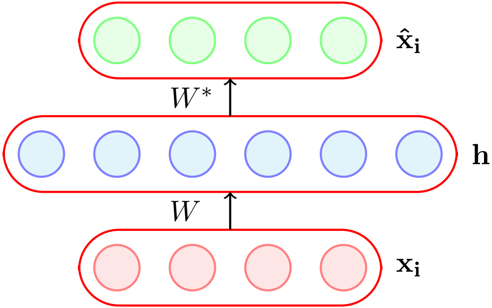
- While poor generalization could happen even in undercomplete autoencoders it is an even more serious problem for overcomplete auto encoders
- Here, (as stated earlier) the model can simply learn to copy $\mathbf{x_i}$ to $\mathbf{h}$ and then $\mathbf{h}$ to $\hat{\mathbf{x}}_\mathbf{i}$
- To avoid poor generalization, we need to introduce regularization



- The simplest solution is to add a L_2-regularization term to the objective function

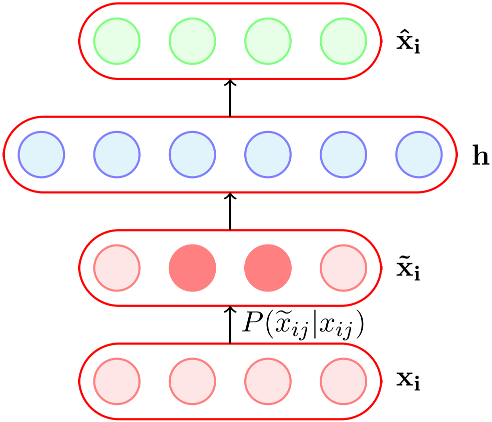
$$\min_{\theta, w, w^*, \mathbf{b}, \mathbf{c}} \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} (\hat{x}_{ij} - x_{ij})^2 + \lambda \|\theta\|^2$$

- This is very easy to implement and just adds a term λW to the gradient $\frac{\partial \mathscr{L}(\theta)}{\partial W}$ (and similarly for other parameters)



- Another trick is to tie the weights of the encoder and decoder i.e., $W^* = W^T$
- This effectively reduces the capacity of Autoencoder and acts as a regularizer

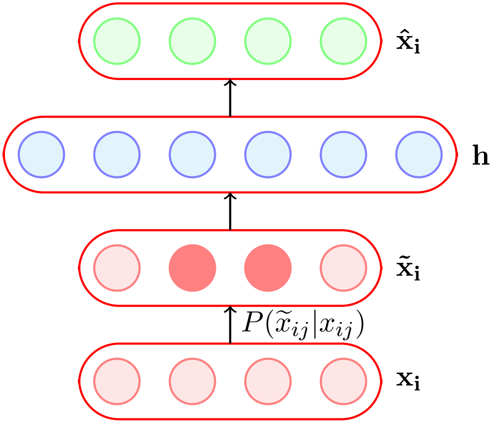
Module 7.4: Denoising Autoencoders



- A denoising encoder simply corrupts the input data using a probabilistic process ($P(\widetilde{x}_{ij}|x_{ij})$) before feeding it to the network
- A simple $P(\widetilde{x}_{ij}|x_{ij})$ used in practice is the following

$$P(\widetilde{x}_{ij} = 0|x_{ij}) = q$$
$$P(\widetilde{x}_{ij} = x_{ij}|x_{ij}) = 1 - q$$

- In other words, with probability q the input is flipped to 0 and with probability $(1 - q)$ it is retained as it is



For example, it will have to learn to reconstruct a corrupted x_{ij} correctly by relying on its interactions with other elements of $\mathbf{x}_i$

- How does this help ?
- This helps because the objective is still to reconstruct the original (un-corrupted) $\mathbf{x}_i$

$$\arg\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} (\hat{x}_{ij} - x_{ij})^2$$

- It no longer makes sense for the model to copy the corrupted $\widetilde{\mathbf{x}}_i$ into $h(\widetilde{\mathbf{x}}_i)$ and then into $\hat{\mathbf{x}}_i$ (the objective function will not be minimized by doing so)
- Instead the model will now have to capture the characteristics of the data correctly.

We will now see a practical application in which AEs are used and then compare Denoising Autoencoders with regular autoencoders

Task: Hand-written digit recognition

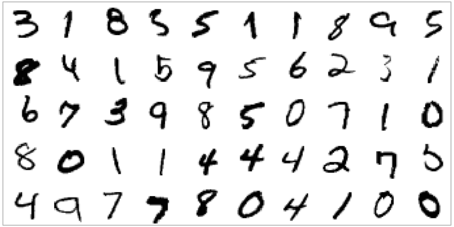
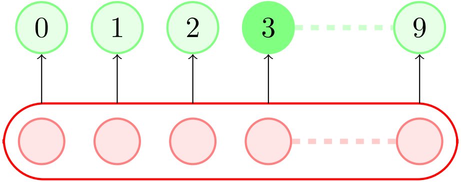


Figure: MNIST Data



$$|\mathbf{x}_i| = 784 = 28 \times 28$$

28*28

Figure: Basic approach(we use raw data as input features)

Task: Hand-written digit recognition

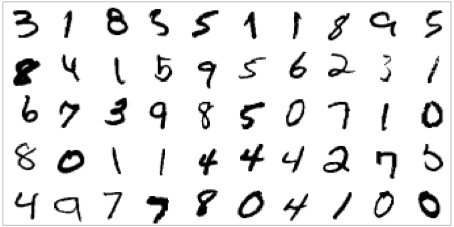


Figure: MNIST Data

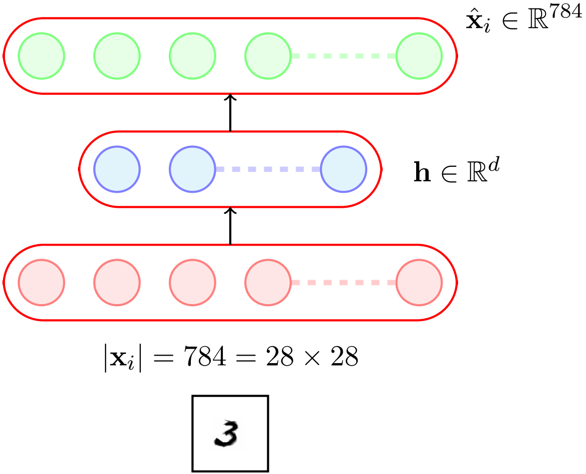


Figure: AE approach (first learn important characteristics of data)

Task: Hand-written digit recognition

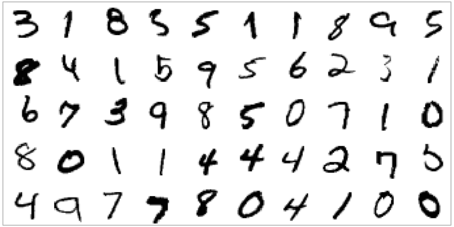


Figure: MNIST Data

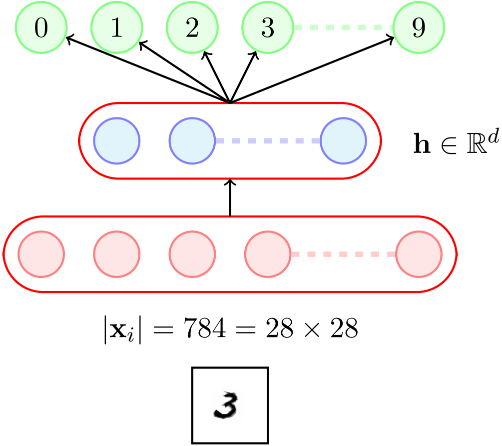
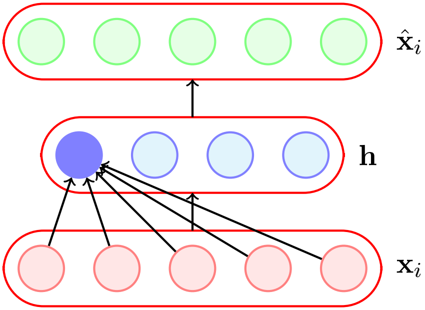


Figure: AE approach (and then train a classifier on top of this hidden representation)

We will now see a way of visualizing AEs and use this visualization to compare different AEs



$$\max_{\mathbf{x}_i} \quad \{W_1^T \mathbf{x}_i\}$$

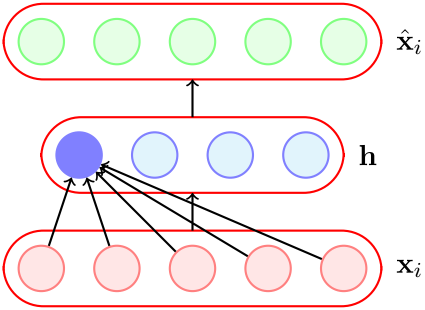
$$s.t. \quad ||\mathbf{x}_i||^2 = \mathbf{x}_i^T \mathbf{x}_i = 1$$

$$\text{Solution:} \quad \mathbf{x}_i = \frac{W_1}{\sqrt{W_1^T W_1}}$$

- We can think of each neuron as a filter which will fire (or get maximally) activated for a certain input configuration $\mathbf{x}_i$
- For example,

$$\mathbf{h}_1 = \sigma(W_1^T \mathbf{x}_i) \ [\textit{ignoring bias } b]$$

  Where W_1 is the trained vector of weights connecting the input to the first hidden neuron
- What values of $\mathbf{x}_i$ will cause $\mathbf{h}_1$ to be maximum (or maximally activated)
- Suppose we assume that our inputs are normalized so that $\|\mathbf{x}_i\| = 1$



$$\max_{\mathbf{x}_i} \quad \{W_1^T \mathbf{x}_i\}$$

$$s.t. \quad ||\mathbf{x}_i||^2 = \mathbf{x}_i^T \mathbf{x}_i = 1$$

$$\text{Solution:} \quad \mathbf{x}_i = \frac{W_1}{\sqrt{W_1^T W_1}}$$

- Thus the inputs

$$\mathbf{x}_i = \frac{W_1}{\sqrt{W_1^T W_1}}, \frac{W_2}{\sqrt{W_2^T W_2}}, \ldots \frac{W_n}{\sqrt{W_n^T W_n}}$$

  will respectively cause hidden neurons 1 to n to maximally fire

- Let us plot these images ($\mathbf{x}_i$'s) which maximally activate the first k neurons of the hidden representations learned by a vanilla autoencoder and different denoising autoencoders
- These $\mathbf{x}_i$'s are computed by the above formula using the weights ($W_1, W_2 \ldots W_k$) learned by the respective autoencoders

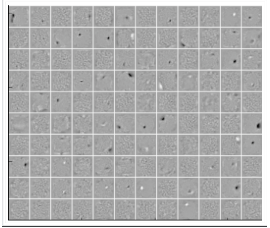


Figure: Vanilla AE (No noise)

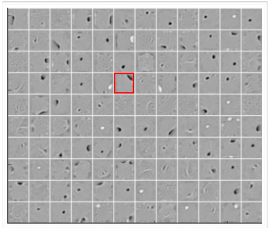


Figure: 25% Denoising AE (q=0.25)

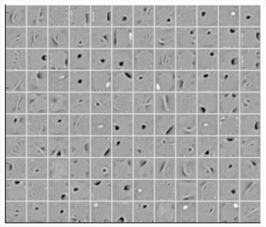
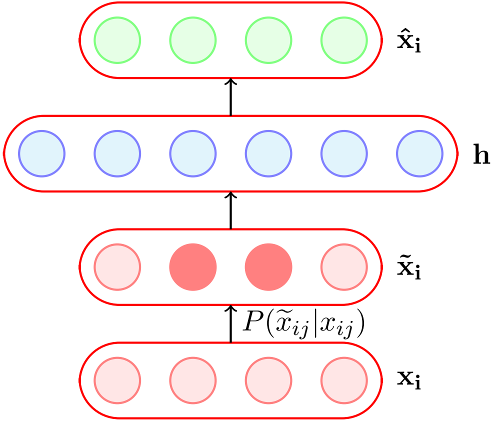


Figure: 50% Denoising AE (q=0.5)

- The vanilla AE does not learn many meaningful patterns
- The hidden neurons of the denoising AEs seem to act like pen-stroke detectors (for example, in the highlighted neuron the black region is a stroke that you would expect in a '0' or a '2' or a '3' or a '8' or a '9')
- As the noise increases the filters become more wide because the neuron has to rely on more adjacent pixels to feel confident about a stroke



- We saw one form of $P(\widetilde{x}_{ij}|x_{ij})$ which flips a fraction q of the inputs to zero
- Another way of corrupting the inputs is to add a Gaussian noise to the input

$$\widetilde{x}_{ij} = x_{ij} + \mathcal{N}(0, 1)$$

- We will now use such a denoising AE on a different dataset and see their performance

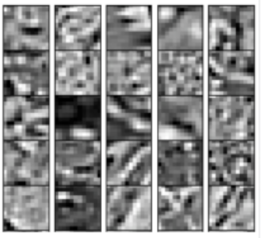


Figure: Data

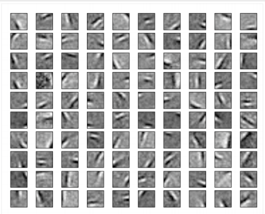


Figure: AE filters

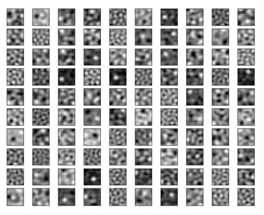
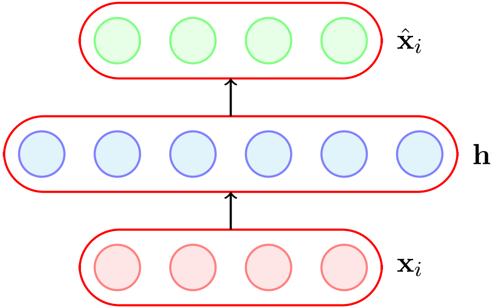


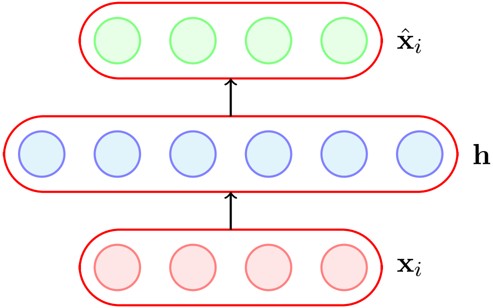
Figure: Weight decay filters

- The hidden neurons essentially behave like edge detectors
- PCA does not give such edge detectors

Module 7.5: Sparse Autoencoders



- A hidden neuron with sigmoid activation will have values between 0 and 1
- We say that the neuron is activated when its output is close to 1 and not activated when its output is close to 0.
- A sparse autoencoder tries to ensure the neuron is inactive most of the times.



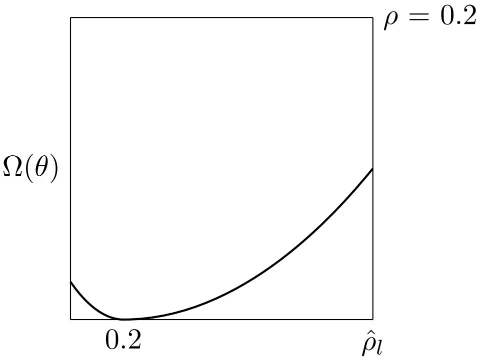
The average value of the activation of a neuron l is given by

$$\hat{\rho}_l = \frac{1}{m}\sum_{i=1}^{m} h(\mathbf{x}_i)_l$$

- If the neuron l is sparse (i.e. mostly inactive) then $\hat{\rho}_l \to 0$
- A sparse autoencoder uses a sparsity parameter ρ (typically very close to 0, say, 0.005) and tries to enforce the constraint $\hat{\rho}_l = \rho$
- One way of ensuring this is to add the following term to the objective function

$$\Omega(\theta) = \sum_{l=1}^{k} \rho \log \frac{\rho}{\hat{\rho}_l} + (1-\rho)\log \frac{1-\rho}{1-\hat{\rho}_l}$$

- When will this term reach its minimum value and what is the minimum value? Let us plot it and check.



- The function will reach its minimum value(s) when $\hat{\rho}_l = \rho$.

$$\Omega(\theta) = \sum_{l=1}^{k} \rho log \frac{\rho}{\hat{\rho}_l} + (1-\rho) log \frac{1-\rho}{1-\hat{\rho}_l}$$

Can be re-written as

$$\Omega(\theta) = \sum_{l=1}^{k} \rho log \rho - \rho log \hat{\rho}_l + (1-\rho) log(1-\rho) - (1-\rho) log(1-\hat{\rho}_l)$$

By Chain rule:

$$\frac{\partial \Omega(\theta)}{\partial W} = \frac{\partial \Omega(\theta)}{\partial \hat{\rho}} . \frac{\partial \hat{\rho}}{\partial W}$$

$$\frac{\partial \Omega(\theta)}{\partial \hat{\rho}} = \left[\frac{\partial \Omega(\theta)}{\partial \hat{\rho}_1}, \frac{\partial \Omega(\theta)}{\partial \hat{\rho}_2}, \ldots \frac{\partial \Omega(\theta)}{\partial \hat{\rho}_k} \right]^T$$

For each neuron $l \in 1 \ldots k$ in hidden layer, we have

$$\frac{\partial \Omega(\theta)}{\partial \hat{\rho}_l} = -\frac{\rho}{\hat{\rho}_l} + \frac{(1-\rho)}{1-\hat{\rho}_l}$$

and $\quad \dfrac{\partial \hat{\rho}_l}{\partial W} = \mathbf{x}_i (g'(W^T \mathbf{x}_i + \mathbf{b}))^T$ (see next slide)

- Now,

$$\hat{\mathscr{L}}(\theta) = \mathscr{L}(\theta) + \Omega(\theta)$$

- $\mathscr{L}(\theta)$ is the squared error loss or cross entropy loss and $\Omega(\theta)$ is the sparsity constraint.
- We already know how to calculate $\frac{\partial \mathscr{L}(\theta)}{\partial W}$
- Let us see how to calculate $\frac{\partial \Omega(\theta)}{\partial W}$.
- Finally,

$$\frac{\partial \hat{\mathscr{L}}(\theta)}{\partial W} = \frac{\partial \mathscr{L}(\theta)}{\partial W} + \frac{\partial \Omega(\theta)}{\partial W}$$

(and we know how to calculate both terms on R.H.S)

Derivation

$$\frac{\partial \hat{\rho}}{\partial W} = \begin{bmatrix} \frac{\partial \hat{\rho}_1}{\partial W} & \frac{\partial \hat{\rho}_2}{\partial W} \dots \frac{\partial \hat{\rho}_k}{\partial W} \end{bmatrix}$$

For each element in the above equation we can calculate $\frac{\partial \hat{\rho}_l}{\partial W}$ (which is the partial derivative of a scalar w.r.t. a matrix = matrix). For a single element of a matrix W_{jl}:-

$$\begin{aligned}
\frac{\partial \hat{\rho}_l}{\partial W_{jl}} &= \frac{\partial \left[\frac{1}{m} \sum_{i=1}^{m} g\left(W_{:,l}^T \mathbf{x_i} + b_l\right) \right]}{\partial W_{jl}} \\
&= \frac{1}{m} \sum_{i=1}^{m} \frac{\partial \left[g\left(W_{:,l}^T \mathbf{x_i} + b_l\right) \right]}{\partial W_{jl}} \\
&= \frac{1}{m} \sum_{i=1}^{m} g'\left(W_{:,l}^T \mathbf{x_i} + b_l\right) x_{ij}
\end{aligned}$$

So in matrix notation we can write it as :

$$\frac{\partial \hat{\rho}_l}{\partial W} = \mathbf{x}_i (g'(W^T \mathbf{x}_i + \mathbf{b}))^T$$

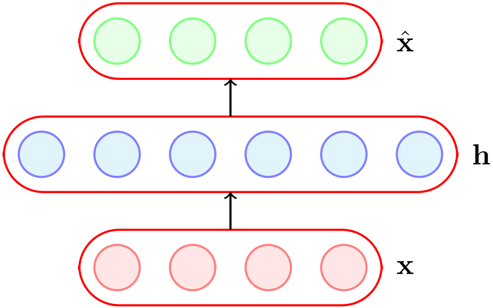
Module 7.6: Contractive Autoencoders

- A contractive autoencoder also tries to prevent an overcomplete autoencoder from learning the identity function.
- It does so by adding the following regularization term to the loss function

$$\Omega(\theta) = \|J_{\mathbf{x}}(\mathbf{h})\|_F^2$$

where $J_{\mathbf{x}}(\mathbf{h})$ is the Jacobian of the encoder.

- Let us see what it looks like.



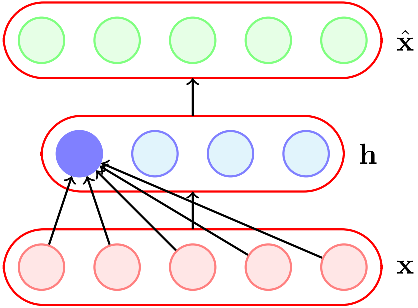
- If the input has n dimensions and the hidden layer has k dimensions then
- In other words, the (l, j) entry of the Jacobian captures the variation in the output of the l^{th} neuron with a small variation in the j^{th} input.

$$J_{\mathbf{x}}(\mathbf{h}) = \begin{bmatrix} \frac{\partial h_1}{\partial x_1} & \dots & \dots & \dots & \frac{\partial h_1}{\partial x_n} \\ \frac{\partial h_2}{\partial x_1} & \dots & \dots & \dots & \frac{\partial h_2}{\partial x_n} \\ \vdots & & \ddots & & \vdots \\ \frac{\partial h_k}{\partial x_1} & \dots & \dots & \dots & \frac{\partial h_k}{\partial x_n} \end{bmatrix}$$

$$\|J_{\mathbf{x}}(\mathbf{h})\|_F^2 = \sum_{j=1}^{n} \sum_{l=1}^{k} \left(\frac{\partial h_l}{\partial x_j} \right)^2$$

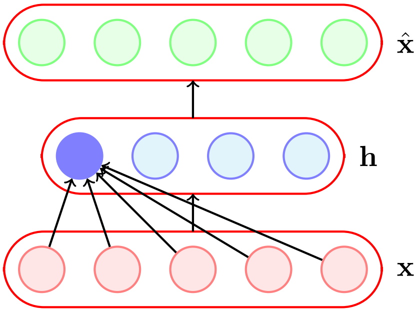
- What is the intuition behind this ?
- Consider $\frac{\partial h_1}{\partial x_1}$, what does it mean if $\frac{\partial h_1}{\partial x_1} = 0$
- It means that this neuron is not very sensitive to variations in the input x_1.
- But doesn't this contradict our other goal of minimizing $\mathcal{L}(\theta)$ which requires $\mathbf{h}$ to capture variations in the input.

$$\|J_{\mathbf{x}}(\mathbf{h})\|_F^2 = \sum_{j=1}^{n}\sum_{l=1}^{k}\left(\frac{\partial h_l}{\partial x_j}\right)^2$$

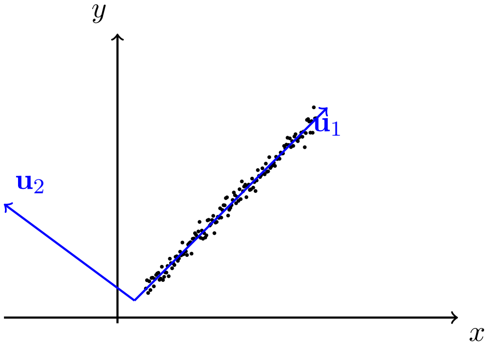


- Indeed it does and that's the idea
- By putting these two contradicting objectives against each other we ensure that **h** is sensitive to only very important variations as observed in the training data.
- $\mathcal{L}(\theta)$ - capture important variations in data
- $\Omega(\theta)$ - do not capture variations in data
- Tradeoff - capture only very important variations in the data

$$\|J_{\mathbf{x}}(\mathbf{h})\|_F^2 = \sum_{j=1}^{n}\sum_{l=1}^{k}\left(\frac{\partial h_l}{\partial x_j}\right)^2$$

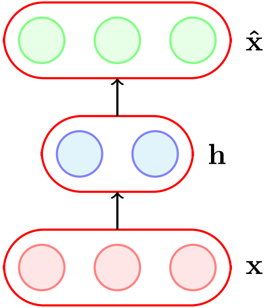


Let us try to understand this with the help of an illustration.

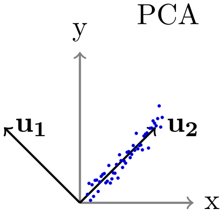


- Consider the variations in the data along directions $\mathbf{u}_1$ and $\mathbf{u}_2$
- It makes sense to maximize a neuron to be sensitive to variations along $\mathbf{u}_1$
- At the same time it makes sense to inhibit a neuron from being sensitive to variations along $\mathbf{u}_2$ (as there seems to be small noise and unimportant for reconstruction)
- By doing so we can balance between the contradicting goals of good reconstruction and low sensitivity.
- What does this remind you of ?

Module 7.7 : Summary

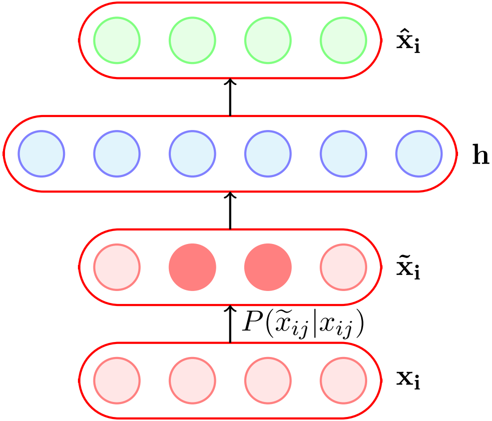


$\equiv$



$$P^T X^T X P = D$$

$$\min_{\theta} \|X - \underbrace{HW^*}_{\substack{U\Sigma V^T \\ \text{(SVD)}}}\|_F^2$$



Regularization

$$\Omega(\theta) = \lambda\|\theta\|^2 \quad \boxed{\text{Weight decaying}}$$

$$\Omega(\theta) = \sum_{l=1}^{k} \rho \log \frac{\rho}{\hat{\rho}_l} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_l} \quad \boxed{\text{Sparse}}$$

$$\Omega(\theta) = \sum_{j=1}^{n} \sum_{l=1}^{k} \left(\frac{\partial h_l}{\partial x_j}\right)^2 \quad \boxed{\text{Contractive}}$$