Lecture Twenty-Six

Time Series: $\underbrace{y_1, \ldots, y_n}$ $\{y_t\}$, $t = 1 \ldots n$

↓

Convert it into regression:

$$\left.\begin{array}{l} \longrightarrow (x_t, y_t), \quad t = 1, \ldots, n \\ x_t = (y_{t-1}, y_{t-2}, \ldots, y_{t-p}) \end{array}\right\}$$

① Linear AR: $\mu_t = \beta_0 + \beta^T x_t$ AR(p)

Loss: $\sum (y_t - \mu_t)^2$

② Nonlinear AR:

$x_t : p \times 1$

$$\boxed{\begin{array}{l} r_t = \sigma(\underbrace{W x_t} + b) \\ \mu_t = \beta_0 + \beta^T r_t \end{array}}$$

$k \times 1$

Single $\boxed{\text{Hidden Layer}}$ Neural Network

r_t: Hidden layer output or Hidden State

$$\sigma\begin{pmatrix} u_1 \\ \vdots \\ u_k \end{pmatrix} = \begin{pmatrix} \sigma(u_1) \\ \sigma(u_2) \\ \vdots \\ \sigma(u_k) \end{pmatrix}$$

$\sigma(u_i) = \max(u_i, 0)$
$= (u_i)_+$

e.g $p = 1$

$$r_t = \begin{pmatrix} (x_t - c_1)_+ \\ \vdots \\ (x_t - c_k)_+ \end{pmatrix}$$

③ Recurrent Neural Networks (RNN)

$$r_t = \tanh(W_r r_{t-1} + W x_t + b), \quad t = 1, \ldots$$

$$\mu_t = \beta_0 + \beta^T r_t$$

$$r_0 = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

$W : k \times p$

$W_r : k \times k$

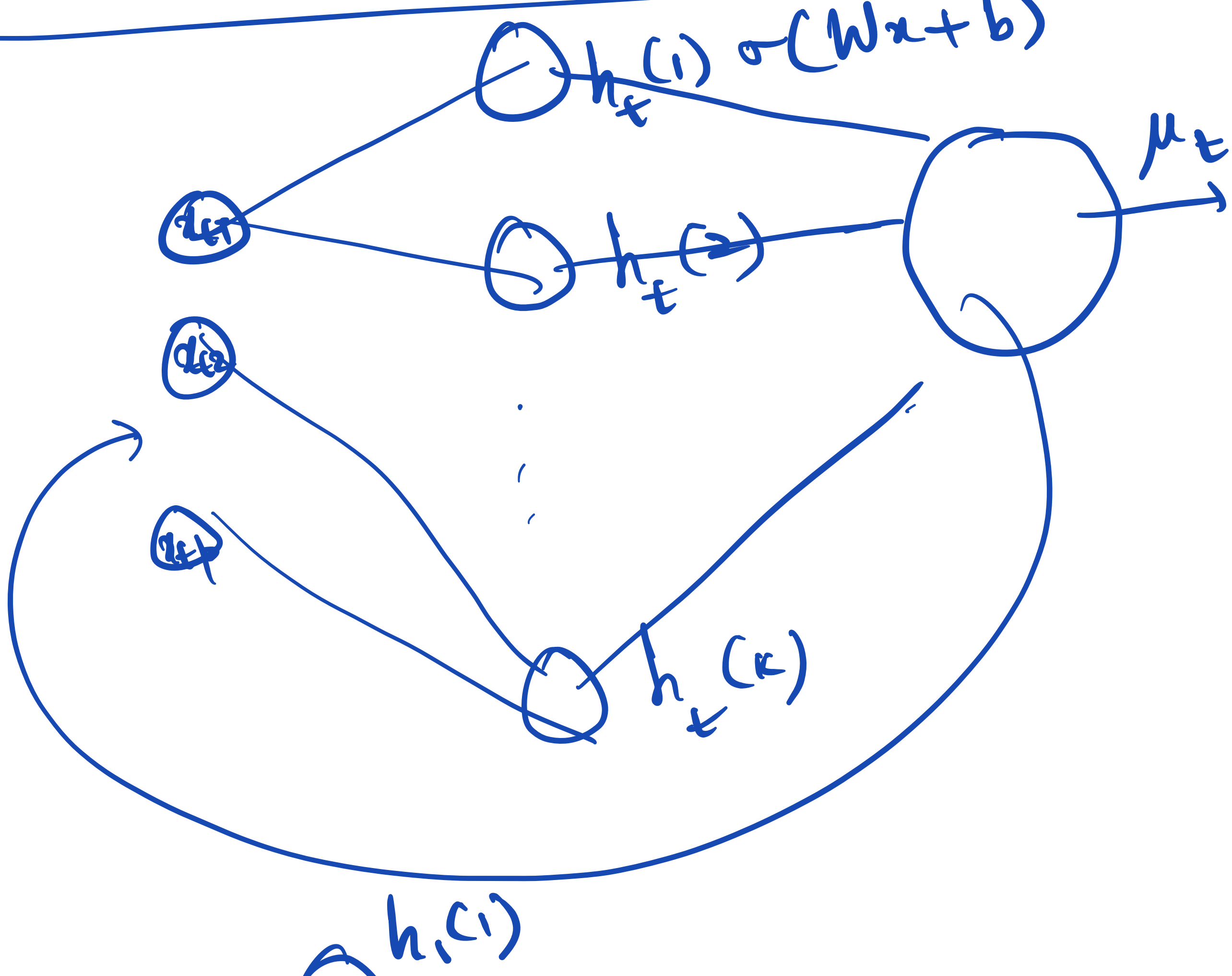
$b : k \times 1$

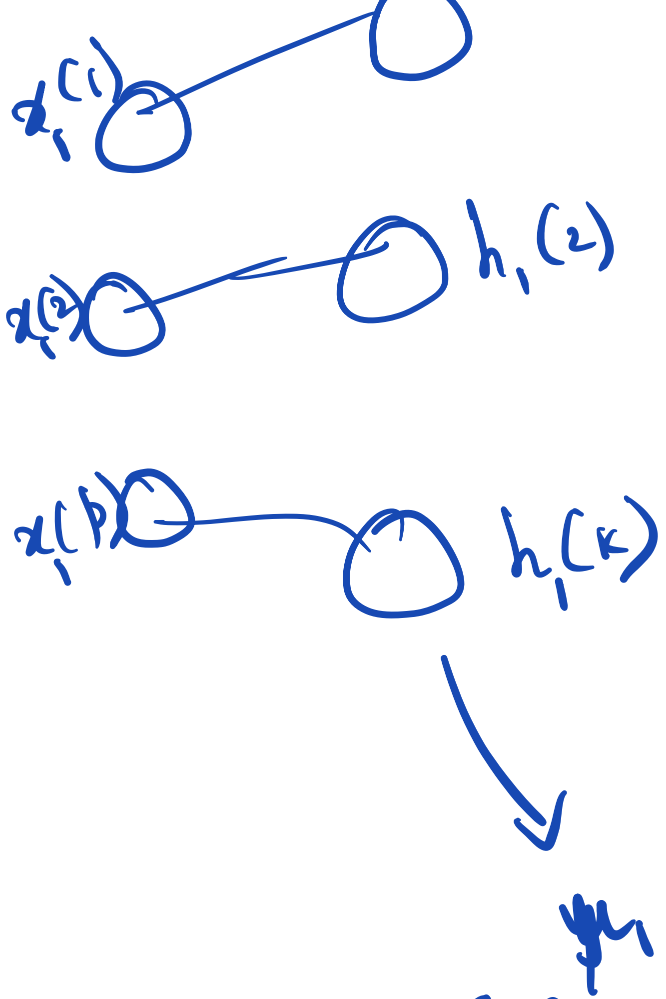
r_t : Hidden State

More commonly people use h_t instead of r_t.

$$h_t = \tanh(\boxed{W_h} h_{t-1} + W x_t + b) \quad t = 1, 2, \ldots$$

$$h_0 = 0$$





If all eigenvalues of W_h < 1 in modulus, then RNN equations are stable but h_t may not have dependence on x_u for $u << t$.

④ GRU (Gated Recurrent Unit)

$$\tilde{h}_t = \tanh(W_h h_{t-1} + W x_t + b)$$

$$h_t = (1-z_t) \odot \tilde{h}_t + z_t \odot h_{t-1}$$

$$\boxed{h_t = z_t \odot h_{t-1} + (1-z_t) \odot \tilde{h}_t}$$

$$z_t = \sigma_{\text{sigmoid}}(W_{hz} h_{t-1} + W_z x_t + b_z)$$

$$\mu_t = \beta_0 + \beta^T h_t$$

$$\text{Loss} = \min_{\substack{W_h, W, b \\ W_{hz}, W_z, b \\ \beta_0, \beta}} \sum (y_t - \mu_t)^2$$

GRU:

$$\tilde{h}_t = \tanh(W_h(h_{t-1} \odot g_t) + W x_t + b)$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t$$

z_t : UPDATE GATE

$$z_t = \sigma_{\text{sigmoid}}(W_{zh} h_{t-1} + W_z x_t + b_z)$$

$$g_t = \text{RESET GATE}$$
$$= \sigma_{\text{sigmoid}}(W_{gh} h_{t-1} + W_g x_t + b_g)$$

⑤ LSTM (Long Short Term Memory)

The hidden state $\{h_t\}$ has two purposes

→ ① Immediate prediction of y_t

② Saving information relevant for future predictions of y_t

$$\mu_t = \beta_0 + \beta^T h_t$$

$$(y_t - \mu_t)^2$$

y_{t+1}

μ_{t+1}

h_t $\quad$ h_{t+1}

h_t : hidden state
(immediate prediction of 'y_t')
current context, 'Short Term Memory'

c_t : cell state ('Long Term Memory')

$$\begin{pmatrix} c_t \\ h_t \end{pmatrix}$$

$$h_{t-1}, \quad x_t \;\rightarrow\; h_t$$

$$(c_{t-1}, h_{t-1}) \quad x_t \;\rightarrow\; (c_t, h_t)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

$$\tilde{c}_t = \tanh(W_{ch}h_{t-1} + W_c x_t + b_c)$$

$$f_t = \sigma_{sigmoid}(W_{fh}h_{t-1} + W_f x_t + b_f)$$

$$i_t = \sigma_{sigmoid}(W_{ih}h_{t-1} + W_i x_t + b_i)$$

$$h_t = o_t \odot \tanh(c_t)$$

$$o_t = \sigma_{sigmoid}(W_{oh}h_{t-1} + W_o x_t + b_o)$$

$\longrightarrow$ LSTM Unit