Lecture Twenty - Five

1. AR & Nonlinear AR
2. RNN
3. GRU
4. LSTM

AR & NAR

$y_1 \cdots y_n$

$y_t, \quad x_t = (y_{t-1}, y_{t-2}, \dots, y_{t-p})$

$t = p+1, \dots, n$

AR: $\rightarrow \mu_t = \beta_0 + \beta^T x_t$

$$\boxed{\text{Loss}: \quad \sum (y_t - \mu_t)^2}$$

$$\boxed{y_t = \mu_t + \varepsilon_t \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)}$$

Qn: How to make AR nonlinear?

$\boxed{p=1}$ $\quad x_t = y_{t-1}$

$$\mu_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 \underbrace{(y_{t-1} - c_1)_+}_{} + \cdots + \beta_{k+1} \underbrace{(y_{t-1} - c_k)_+}_{}$$

$c_1 \cdots c_k$

$\beta_0, \beta_1, \beta_2 \cdots \beta_{k+1}$

$$y_{t-1} = y_{t-1} - c_0 + c_0 = \boxed{(y_{t-1} - c_0)_+} + c_0$$

$$\boxed{\mu_t = \beta_0 + \beta_1 (y_{t-1} - c_1)_+ + \cdots + \beta_k (y_{t-1} - c_k)_+}$$

input or covariate at time t

$p = 1$

$$x_t = y_{t-1}$$

$$s_t = \begin{pmatrix} x_t - c_1 \\ \vdots \\ x_t - c_k \end{pmatrix} \rightarrow k \times 1$$

$$\sigma(u) = \begin{pmatrix} \sigma(u_1) \\ \vdots \\ \sigma(u_k) \end{pmatrix} = \begin{pmatrix} (u_1)_+ \\ \vdots \\ (u_k)_+ \end{pmatrix}$$

Feature vector at time t.

$$r_t = \sigma(s_t)$$

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

NAR(1)

$\boxed{c_1 \ldots c_k}$ β_0, β

$\boxed{\text{NAR}(p)}$

$$x_t = (y_{t-1} \cdots y_{t-p})^T \rightarrow p \times 1$$

$$s_t = \begin{pmatrix} x_t - c_1 \\ \vdots \\ x_t - c_k \end{pmatrix} \qquad c_1 \cdots c_k \in \mathbb{R}^p$$

$$r_t = \sigma(s_t)$$

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

ADDITIVE NON-LINEAR Model

Check:

$$\left.\begin{aligned} \mu_t = & \Big[(\;)(\underbrace{x_t^{(1)}}_{y_{t-1}} - c_1)_+ + \cdots + (\;)(x_t^{(1)} - c_k)_+ \Big] \\ & + \Big[x_t^{(2)} \qquad \Big] \\ & + \Big[x_t^{(3)} \qquad \Big] \end{aligned}\right\} \text{Additive model}$$

μ_t $y_{t-1} \times y_{t-2}$

NAR(p):

$$x_t = (y_{t-1}, \dots, y_{t-p})^T \rightarrow p \times 1$$

$$\underset{k\times 1}{s_t} = \underset{k \times p}{W} x_t + \underset{k \times 1}{b}$$

$$r_t = \sigma(s_t)$$

$$\mu_t = \beta_0 + \beta^T \boxed{r_t}$$

parameters:

W, b, β_0, β

$W \rightarrow k \times p$

μ_t : depends on x_t because r_t depends on x_t

$$\boxed{r_t = \sigma(W x_t + b)}$$

Recurrent Neural Network (RNN)

$$x_t = (\underbrace{y_{t-1} \dots y_{t-p}}) ^T$$

$$s_t = W x_t + b$$

$$r_t = \sigma(s_t)$$

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

$\rightarrow$ NAR(p)

$$\boxed{y_t = f(y_{t-80}) + \varepsilon_t}$$

RNN:

$$x_t = (y_{t-1}, \dots, y_{t-p})^T$$

$$s_t = \boxed{W_r r_{t-1} +} W x_t + b$$

$$r_t = \sigma(s_t)$$

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

RNN

$$x_t = (y_{t-1} \cdots y_{t-p})^T$$
$$r_t = \sigma(W_r r_{t-1} + W x_t + b), \quad r_0 = 0$$
$$\mu_t = \beta_0 + \beta^T r_t$$

$t=1$

$$r_1 = \sigma(W x_1 + b)$$
$$r_2 = \sigma(W_r r_1 + W x_2 + b)$$
$$= \sigma(W_r \sigma(W x_1 + b) + W x_2 + b)$$

$\rightarrow$ depends on both x_2 & x_1

$$r_3 = \sigma(W_r r_2 + W x_3 + b)$$
$$= \sigma(W_r \sigma(W_r \sigma(W x_1 + b) + W x_2 + b) + W x_3 + b)$$

$\rightarrow$ depends on x_3, x_2, x_1

More generally,

r_t depends on $x_t, x_{t-1}, x_{t-2}, \ldots, x_1$

It can have stability issues.

$$\sigma(u) = \max(u, 0)$$
$$\sigma_{\tanh}(u) = \tanh(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}} \in (-1, 1)$$

$$\boxed{r_t = \tanh(W_r r_{t-1} + W x_t + b), \quad \mu_t = \beta_0 + \beta^T r_t}$$

We want $\underbrace{W_r \text{ to not be more than } 1.}$

(spectral radius of W_r should be < 1) largest modulus of any eigenvalue

But when W_r has spectral radius < 1,

[r_t might have weak dependence on x_u when u is much smaller than t.]

To verify this, calculate $\dfrac{\partial r_t \rightarrow k\times 1}{\partial x_u \rightarrow p\times 1}$ JACOBIAN

$$\frac{\partial r_3}{\partial x_3} = \frac{\partial}{\partial x_3}\sigma(\underbrace{W_r r_2 + W x_3 + b}_{s_3})$$

$$= \underbrace{\sigma'(s_3)\frac{\partial s_3}{\partial x_3}}_{\text{chain rule}} = \sigma'(s_3)\, W$$

diagonal matrix $\begin{bmatrix} \sigma'(s_3(1)) & & \\ & \ddots & \\ & & \sigma'(s_3(k)) \end{bmatrix}$

$$\frac{\partial r_3}{\partial x_2} = \frac{\partial}{\partial x_2}\sigma(\underbrace{W_r r_2 + W x_3 + b}_{s_3})$$

$$= \sigma'(s_3)\frac{\partial s_3}{\partial x_2}$$

$$= \sigma'(s_3)\, W_r \frac{\partial r_2}{\partial x_2} = \sigma'(s_3)\, W_r\, \sigma'(s_2)\, W$$

$$\frac{\partial r_3}{\partial x_1} = \sigma'(s_3) W_r \sigma'(s_2) W_r \sigma'(s_1) W$$

$$\boxed{\frac{\partial r_{100}}{\partial x_{60}}} = \sigma'(s_{100}) W_r \sigma'(s_{99}) W_r \sigma'(s_{98}) W_r \cdots W_r \sigma'(s_{60}) W$$

very small

$$\sigma(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}} \Rightarrow \sigma'(u) = 1 - \sigma^2(u) \in (0,1)$$

$\frac{\partial r_t}{\partial x_u}$ is probably very small when $u \ll t$.

RNNs may not have LONG RANGE DEPENDENCE in practice.

$$\underbrace{r_t = \underset{\tanh}{\sigma}(\boxed{W_r} r_{t-1} + W x_t + b)}, \quad r_0 = 0$$

Gated Recurrent Unit (GRU)

$$\boxed{\tilde{r}_t = \tanh(W_r r_{t-1} + W x_t + b)}$$

ⓐ $r_t = \tilde{r}_t \rightarrow$ RNN (either explosion or lack of long range dependence)

ⓑ $\boxed{r_t = r_{t-1}} \rightarrow$ will not use the current input x_t.

GRU:

$$r_t = z_t \odot r_{t-1} + (1 - z_t) \odot \tilde{r}_t, \quad t = 1, \dots, n$$

z_t: $k \times 1$ $\quad r_{t-1}$: $k \times 1$ $\quad \tilde{r}_t$: $k \times 1$ $\quad (1-z_t)$: $k \times 1$

$$\rightarrow z_t = \sigma(W_{rz} r_{t-1} + W_z x_t + b_z)$$

sigmoid $\quad \sigma(u) = \dfrac{e^u}{1+e^u} \in (0,1)$

GRU:

$$\tilde{r}_t = \tanh(W_r r_{t-1} + W x_t + b)$$

$$z_t = \sigma(W_{rz} r_{t-1} + W_z x_t + b_z)$$

$$r_t = z_t \odot r_{t-1} + (1 - z_t) \odot \tilde{r}_t$$

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

$\boxed{z_t : \text{UPDATE GATE}}$

GRU:

$$\tilde{r}_t = \tanh(W_r (r_{t-1} \odot g_t) + W x_t + b)$$

UPDATE GATE $\rightarrow$ $z_t = \sigma(W_{rz} r_{t-1} + W_z x_t + b_z)$

$$r_t = z_t \odot r_{t-1} + (1 - z_t) \odot \tilde{r}_t$$

RESET GATE $\rightarrow$ $g_t = \sigma(W_{rg} r_{t-1} + W_g x_t + b_g)$

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

$$h_t \longrightarrow \boxed{h_{t-1}} \;\&\; \boxed{x_t}$$

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

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

$t = 100$ $\quad z_t = 0.9999$

LSTM (Long Short Term Memory)

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

$h_{t-1} \longrightarrow h_t$: GRU

$h_{t-1} \longrightarrow h_t$ RNN

$$\begin{pmatrix} s_{t-1} \\ h_{t-1} \end{pmatrix} \longrightarrow \begin{pmatrix} s_t \\ h_t \end{pmatrix}$$

$$s_t = \underbrace{\text{gate}}_{\text{forget}} \odot h_{t-1} + \overset{\text{another}}{\underbrace{\text{gate}}_{\text{update}}} \odot s_{t-1}$$

$$h_t = \text{gate} \odot \sigma(s_t)$$