- blog site: https://r2rt.com/recurrent-neural-networks-in-tensorflow-i.html
- edit date: 2017.06.15

[Input Sequence X]

At time step t, x_t in X has a 50% chance of being 1 (and a 50% chance of being 0).
E.g., X:[x_1, x_2, x_3, x_4, x_5...] might be [1, 0, 0, 1, 1, …].

[Output sequence (Y)]

General Situation: At time step t, y_t has a base 50% chance of being 1 (and a 50% base chance to be 0).

- RULE 1: The chance of y_t being 1 is increased by 50% (i.e., 50% to 100%) if X_{t-3} is 1,
- RULE2: The chance of y_t being 1 decreased by 25% (i.e., 50% to 25%) if X_{t-8} is 1.

E.g., If both Xt−3 and Xt−8 are 1, the chance of Yt being 1 is 50% + 50% - 25% = 75%.

[Training Result]

There will be three situations.
Scenario one: Learn nothing: the model only has the probability distribution of Y, ignore all the information between Y and X.
Scenario two: Learn RULE 1: model learn the information between y_t and x_{t-3}

Scenario three: Learn RULE1 & RULE2: model learn the information between y_t and (x_{t-3} , x_{t-8})

CROSS ENTROPY

Definition: suppose there are n events in space

- cross entropy : - **Σ** P(event)*log(Q(event))
- Real distribution: P
- Model distribution:Q
- P(event) : the probability of event happens in P

IDEA: we use cross entropy as our loss function, the less the better. Therefore, the model is trying make the distribution of Q fit Y. As a result, the least of cross entropy happens when P(Y)=Q(Y). The least: - **Σ** P(event)*log(P(event))

Scenario one: consider two events

REAL DISTRIBUTION

- Event 1: $P(y_t=1) = P(y_t=1 \mid x_{t-3}=1, x_{t-8}=1)*P(x_{t-3}=1,)*P(x_{t-8}=1)$

$$+ P(y_t=1 \mid x_{t-3}=1, x_{t-8}=0)*P(x_{t-3}=1)*P(x_{t-8}=0)$$

$+ P(y_t=1 \mid x_{t-3}=0, x_{t-8}=1)*P(x_{t-3}=0)*P(x_{t-8}=1)$

$+ P(y_t=1 \mid x_{t-3}=0, x_{t-8}=0)*P(x_{t-3}=0)*P(x_{t-8}=0)$

- $P(y_t=1) = 0.75 * 0.25 + 1 * 0.25 + 0.25 * 0.25 + 0.5 *0.25 = 0.625$

- Event 2: $P(y_t=0) = P(y_t=0 \mid x_{t-3}=1, x_{t-8}=1)*P(x_{t-3}=1)*P(x_{t-8}=1)$

  $+ P(y_t=0 \mid x_{t-3}=1, x_{t-8}=0)*P(x_{t-3}=1)*P(x_{t-8}=0)$

  $+ P(y_t=0 \mid x_{t-3}=0\ x_{t-8}=1)*P(x_{t-3}=0)*P(x_{t-8}=1)$

  $+ P(y_t=0 \mid x_{t-3}=0, x_{t-8}=0)*P(x_{t-3}=0)*P(x_{t-8}=0)$

  - $P(y_t=0) = 0.25 * 0.25 + 0 * 0.25 + 0.75 * 0.25 + 0.5 * 0.25 = 0.375$
- owing to the reason that the the distribution of Q tries to fit with Y

  cross entropy = -(0.625 * np.log(0.625) + 0.375 * np.log(0.375))

Scenario two: consider four events

REAL DISTRIBUTION

- Event 1: $P(y_t=1 \mid x_{t-3} = 1) = P(y_t=1 \mid x_{t-3}=1, x_{t-8}=1) * P(x_{t-8}=1) + P(y_t=1 \mid x_{t-3}=1, x_{t-8}=0) * P(x_{t-8} = 0) = 0.75 * 0.5 + 1 * 0.5$
- Event 2: $P(y_t=1 \mid x_{t-3} = 0) = P(y_t=1 \mid x_{t-3}=0, x_{t-8}=1) * P(x_{t-8}=1) + P(y_t=1 \mid x_{t-3}=0, x_{t-8}=0) * P(x_{t-8} = 0) = 0.25 * 0.5 + 0.5 * 0.5$
- Event 3: $P(y_t=0 \mid x_{t-3} = 1) = P(y_t=0 \mid x_{t-3}=1, x_{t-8}=1) * P(x_{t-8}=1) + P(y_t=0 \mid x_{t-3}=1, x_{t-8}=0) * P(x_{t-8} = 0) = 0.25 * 0.5 + 0 * 0.5$
- Event 4: $P(y_t=0 \mid x_{t-3} = 0) = P(y_t=0 \mid x_{t-3}=0, x_{t-8}=1) * P(x_{t-8}=1) + P(y_t=0 \mid x_{t-3}=0, x_{t-8}=0) * P(x_{t-8} = 0) = 0.75 * 0.5 + 0.5 * 0.5$
- cross entropy = -0.5 * (0.875 * np.log(0.875) + 0.125 * np.log(0.125) - 0.5 * (0.625 * np.log(0.625) + 0.375 * np.log(0.375)))
  - 0.5 stands for $P(x_{t-3} = 1)$ and $P(x_{t-3} = 0)$

Scenario three: consider eight events

REAL DISTRIBUTION

- Event 1: $P(y_t=1 \mid x_{t-3}=1, x_{t-8}=1) = 0.75$
- Event 2: $P(y_t=1 \mid x_{t-3}=1, x_{t-8}=0) = 1$
- Event 3: $P(y_t=1 \mid x_{t-3}=0, x_{t-8}=1) = 0.25$
- Event 4: $P(y_t=1 \mid x_{t-3}=0, x_{t-8}=0) = 0.5$
- Event 5: $P(y_t=0 \mid x_{t-3}=1, x_{t-8}=1) = 0.25$
- Event 6: $P(y_t=0 \mid x_{t-3}=1, x_{t-8}=0) = 0$
- Event 7: $P(y_t=0 \mid x_{t-3}=0, x_{t-8}=1) = 0.75$
- Event 8: $P(y_t=0 \mid x_{t-3}=0, x_{t-8}=0) = 0.5$
- cross entropy = -0.25 * (2 * 0.75 * np.log(0.75) + 2 * 0.25 * np.log(0.25) + 2 * 0.50 * np.log (0.50)) - 0.25 * (0 * np.log(0)) -0.25 * (1 * log(1))

- 0.25 stands for $P(x_{t-3} = 1,)*P(x_{t-8} = 1)$, $P(x_{t-3}= 1)*P(x_{t-8} = 0)$, $P(x_{t-3}= 0)*P(x_{t-8} = 1)$, $P(x_{t-3}= 0)*P(x_{t-8} = 0)$