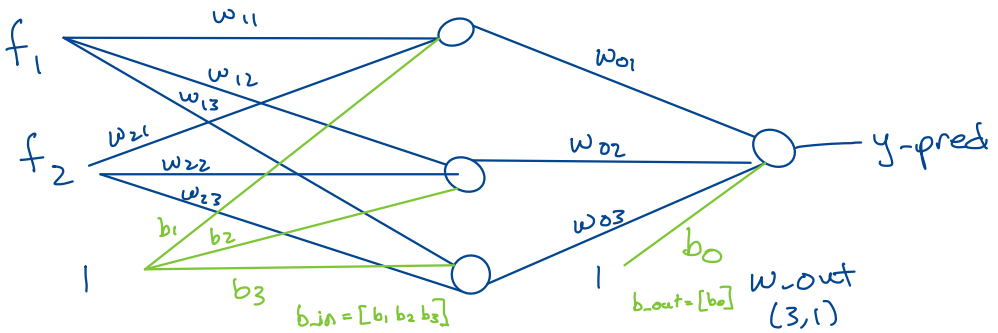
# FFN (updated)

1 sample, 2 features, hidden size = 3




$$h\_out^{*} = \begin{bmatrix} f_1 w_{11} + f_2 w_{21} + b_1 & f_1 w_{12} + f_2 w_{22} + b_2 & f_1 w_{13} + f_2 w_{23} + b_3 \end{bmatrix}$$

X shape: (1, 2)

w_in shape: (2, 3)

b_in shape: (3,)

if b_in is a scalar, then $b_1 == b_2 == b_3$. It's better to initialize b_in as a vector with dimension hidden_size, so that each node has its own bias.

$h\_out = X \cdot w + b$

$(1,2) \cdot (2,3) + (3,)$

$(1,3) + (3,) \Rightarrow (1,3)$

$$= \begin{bmatrix} f_1 & f_2 \end{bmatrix} \cdot \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} + \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix} =$$

w_out shape: (3, 1)

b_out shape: (num_classes,)

scalar for binary classif.

$$y\_pred^{*} = h\_out \cdot w\_out + b\_out$$

$(1,3) \cdot (3,1) + (1,)$

$(1,1)$

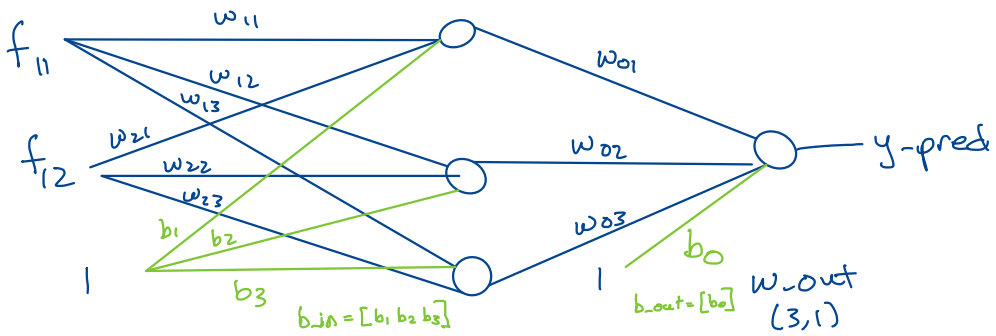
$$= \begin{bmatrix} h_1 & h_2 & h_3 \end{bmatrix} \cdot \begin{bmatrix} w_{01} \\ w_{02} \\ w_{03} \end{bmatrix} + b_0$$

$$= h_1 w_{01} + h_2 w_{02} + h_3 w_{03} + b_0$$

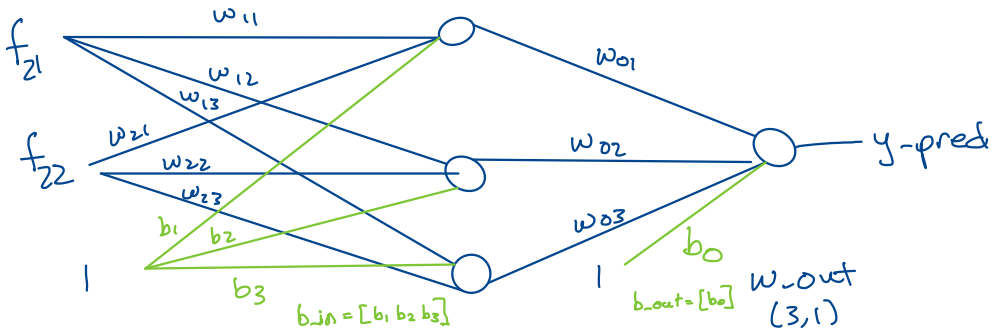
* apply the layer activation functions

2 samples, 2 features, hidden size = 3

sample1




sample2




$$h\_out^{*} = \begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix} \cdot \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} + [b_1\ b_2\ b_3]$$

added to each row, called 'broadcasting'

$$(2,2) \cdot (2,3) + (3,)$$

$$(2,3)$$

$$y\_pred^{*} = hout \cdot w\_out + b\_out$$

$$(2,3] \cdot (3,1) + (1,)$$

$$(2,1)$$

2 predictions

* apply the layer activation functions