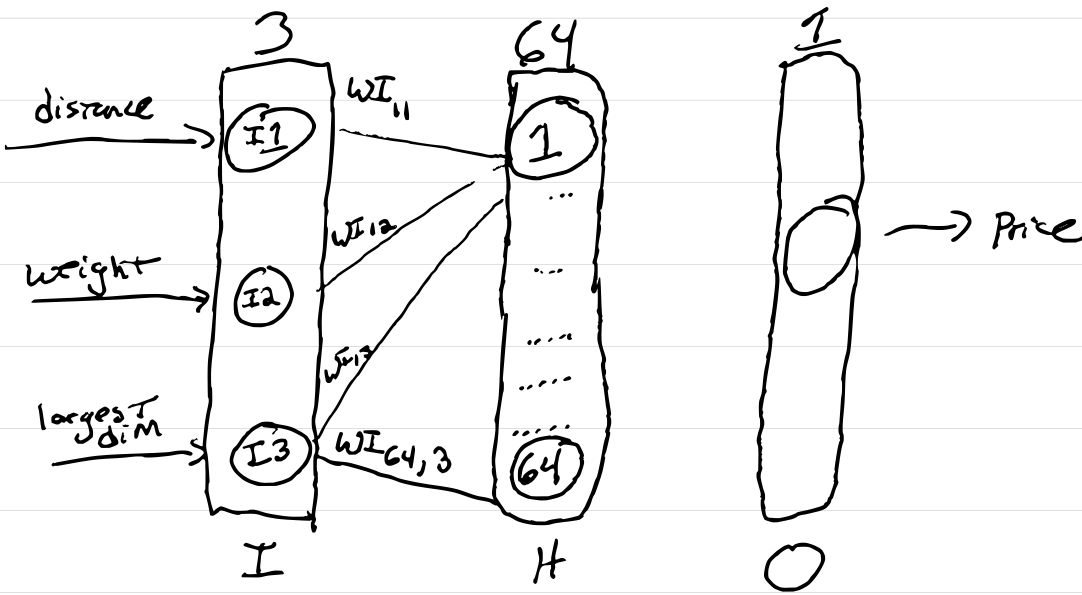
Simple prediction Network



distance, weight, largest dim Need to be standardized

output may also benefit from being scaled

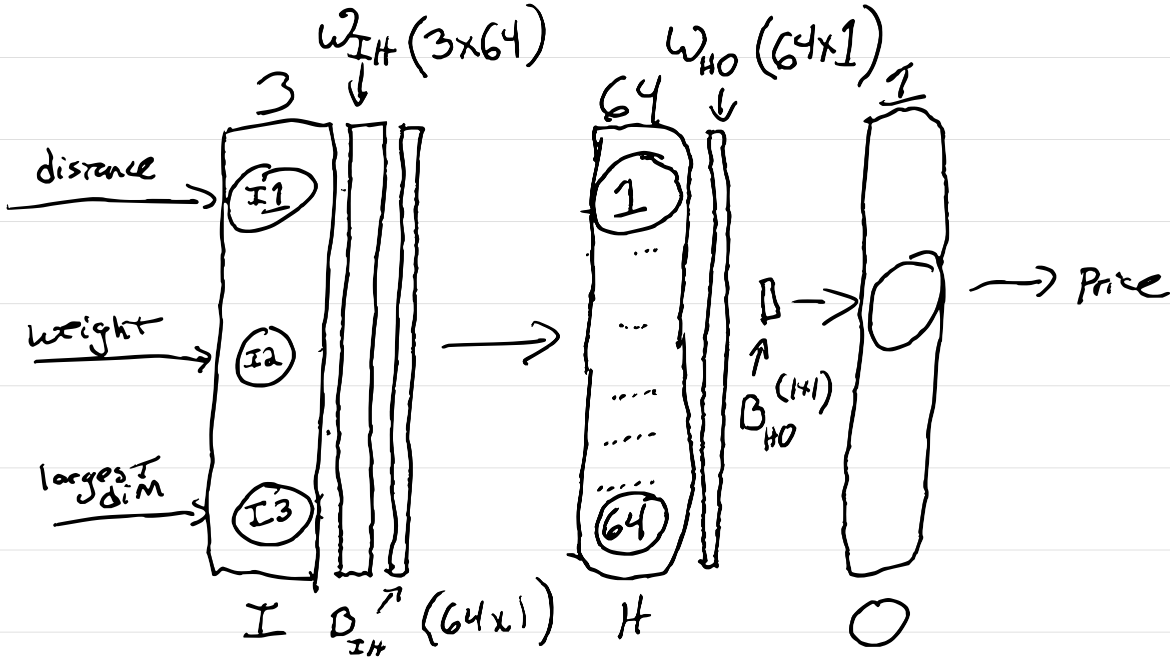
Matrix mult $(a, b) \cdot (b, c)$ results in (a, c)

Weight Matrix

Number of inputs

Number of elements

$$W = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \\ w_{41} & w_{42} & w_{43} \\ \vdots & \vdots & \vdots \\ w_{x1} & w_{x2} & w_{x3} \end{bmatrix}$$



Computing Gradients

Need partial derivatives for both the weights and the bias for backpropagation.

* We will use MSE as our loss function*

L = loss fn, BS = Batch Size, y_i = prediction, t = True value

I = output from input layer

$$\boxed{L = \frac{1}{BS} \cdot \sum_{i=1}^{BS} \left((y_i - t_i)^2 \right)}$$

Gradient for hidden to output

I = hidden layer output after activation

$$\frac{\partial L}{\partial W_{HO}} = \frac{1}{BS} \cdot \sum_{i=1}^{BS} \left(2 \cdot \frac{\partial}{\partial W_{HO}} (y_i - t_i) \right)$$

remember that $y_i = \text{ReLU}(W_{HO} \cdot I + b)$

$$= \frac{1}{BS} \cdot \sum_{i=1}^{BS} \left(2 \cdot (y_i - t_i) \cdot \begin{cases} W_{HO} \cdot I + b > 0 & 1 \\ W_{HO} \cdot I + b \leq 0 & 0 \end{cases} \cdot I \right)$$

Gradient for weights of input to hidden (W_{IH})

I = inputs to regression

X = output (post activation) from IH

$$\hat{y}_i = HO(IH(I))$$

$$IH(I) = \mathrm{ReLU}(W_{IH} \cdot I + B_{IH})$$

$$HO(X) = \mathrm{ReLU}(W_{HO} \cdot X + B_{HO})$$

$$\boxed{L = \frac{1}{BS} \cdot \sum_{i=1}^{BS} \left((\hat{y}_i - t_i)^2 \right)}$$

$$\frac{\partial L}{\partial W_{IH}} = \frac{\partial L}{\hat{y}_i} \frac{\partial HO}{\partial X} \frac{\partial IH}{\partial I}$$

$$= \frac{1}{BS} \cdot \sum_{i=1}^{BS} \left(2(\hat{y}_i - t_i) \cdot \frac{\partial HO}{\partial X} \cdot \frac{\partial IH}{\partial W_{IH}} \right)$$

$$= \frac{1}{BS} \cdot \sum_{i=1}^{BS} \left(2(\hat{y}_i - t_i) \cdot \left\{ \begin{array}{l|l} W_{HO} \cdot X + B_{HO} > 0 & 1 \\ W_{HO} \cdot X + B_{HO} \leq 0 & 0 \end{array} \right\} \cdot W_{HO} \cdot \frac{\partial IH}{\partial W_{IH}} \right)$$

$$= \frac{1}{BS} \cdot \sum_{i=1}^{BS} \left(2(\hat{y}_i - t_i) \cdot \left\{ \begin{array}{l|l} W_{HO} \cdot X + B_{HO} > 0 & 1 \\ W_{HO} \cdot X + B_{HO} \leq 0 & 0 \end{array} \right\} \cdot W_{HO} \cdot \left\{ \begin{array}{l|l} W_{IH} \cdot I + B_{IH} > 0 & 1 \\ W_{IH} \cdot I + B_{IH} \leq 0 & 0 \end{array} \right\} \cdot I \right)$$