Neural Network: architecture

A Neural Network (Sequential architecture) is composed of

- sequence of Layers
    - layer $\ll$ transforms its input $\backslash\mathbf{y}_{(\ll-1)}$ to output $\backslash\mathbf{y}_{\backslash\mathrm{llp}}$
        - through a transformation: operation parameterized by weights $\backslash\mathbf{W}_{\backslash\mathrm{llp}}$

In [6]:

```
fig_tf_seq
```

Out[6]:

- initial layer 0 is Input layer:
  - outputs the network's inputs $\backslash\mathbf{x}$

$$\backslash\mathbf{y}_{(0)} = \backslash\mathbf{x}$$

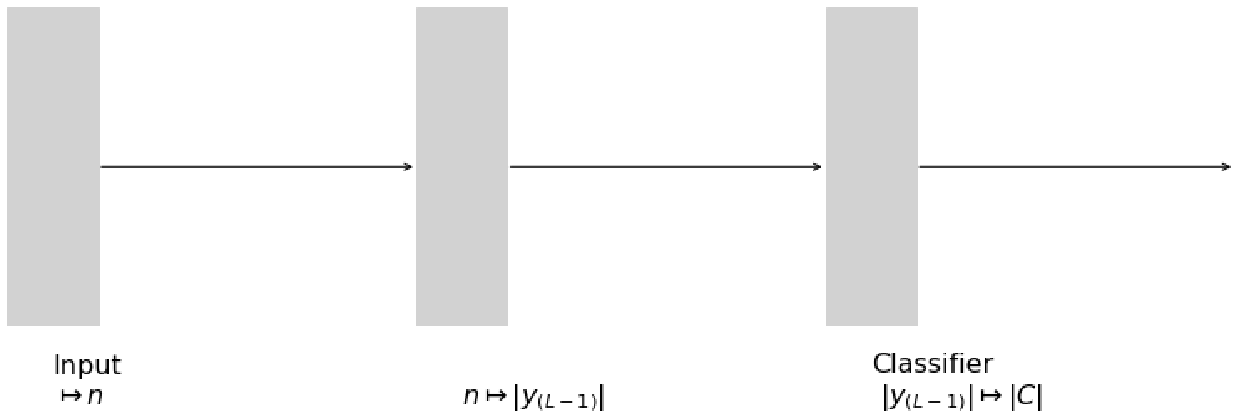
- final layer L transforms its input $\backslash\mathbf{y}_{(L-1)}$ to prediction $\backslash\hat{\mathbf{y}}$

$$\backslash\hat{\mathbf{y}} = \backslash\mathbf{y}_{(L)}$$

  - transformation of layer L usually: Regression or Classification

Head layer: Classification or Regression

In [7]:
```
fig_tf_test
```
Out[7]:



In the above diagram

- the central box represents a sequence of 1 or more layers
- a sub-network
- just for brevity

Non-head layers (central box)

- transform raw features into synthetic features
- for consumption by Head layer

Head Layer (last box)

- Linear Regression or Classification

Network computes an *unknown* function

- from input of length n (the raw input)
- to a vector of length $|\backslash\mathbf{y}_{(L-1)}|$

whose purpose is

- to make the final layer L (e.g., the Classifier)
- create predictions (outputs)
- that have a very-low loss $\backslash\mathrm{loss}$
    - a good "approximation" of the function described by the training data

Dense (Fully Connected) Layer

Dense layer ≪ with $n_{\backslash\mathbf{llp}}$ units and sigmoid activation

$y_{(l-1)}$

Layer l

$$y_{(l),1} = \sigma\Big(y_{(l-1)} \cdot W_{(l),1}\Big)$$
$$y_{(l),2} = \sigma\Big(y_{(l-1)} \cdot W_{(l),2}\Big)$$
$$y_{(l),n_{(l)}} = \sigma\Big(y_{(l-1)} \cdot W_{(l),n_{(l)}}\Big)$$

$y_{(l)}$

Output $\backslash\mathbf{y}_{(\ll-1)}$ of layer $\ll -1$

- matched against $n_{\backslash\mathrm{llp}}$ patterns
$$\backslash\mathbf{W}_{(\ll,1)}, \ldots, \backslash\mathbf{W}_{(\ll,n_{\backslash\mathrm{llp}})}$$
- passed through a sigmoid activation function

$n_{\backslash\mathrm{llp}}$ logistic regressions

- Does $\backslash\mathbf{y}_{(\ll-1)}$ match each of the $n_{\backslash\mathrm{llp}}$ patterns

Special cases

Usually for Head layer.

- Sigmoid activation with $n_{\text{\llp}} = 1$
  - Binary logistic regression
- Softmax activation
  - Multinomial logistic regression
- Linear (or no) activation, $n_{\text{\llp}} = 1$
  - Linear Regression

Importance of training dataset

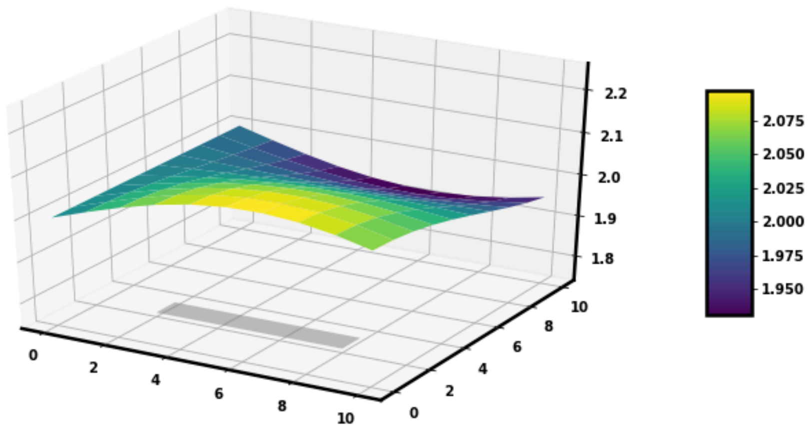
The Neural network thus computes a *function* from $\backslash\mathbf{x}$ to $\hat{\backslash\mathbf{y}}$.

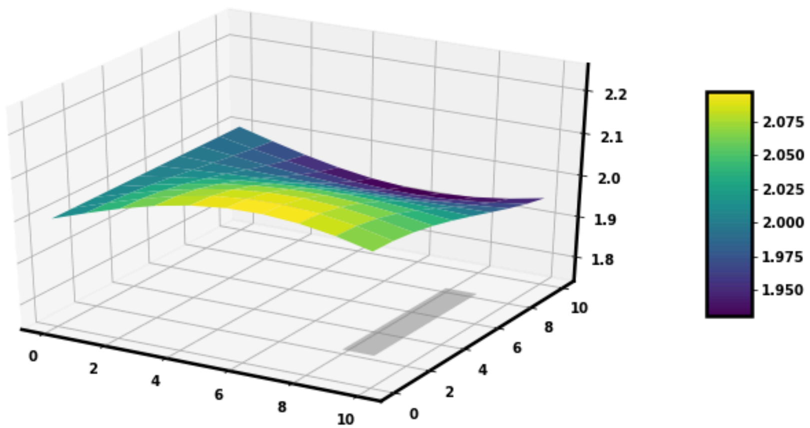
The function *mimics* the training data

$$\langle \backslash\mathbf{X}, \backslash\mathbf{y} \rangle = [\backslash\mathbf{x}^{\backslash\mathrm{ip}}, \backslash\mathbf{y}^{\backslash\mathrm{ip}} | 1 \leq i \leq m]$$

where each *example* $\langle \backslash\mathbf{x}^{\backslash\mathrm{ip}}, \backslash\mathbf{y}^{\backslash\mathrm{ip}} \rangle$

- describes the mapping of the function on input $\backslash\mathbf{x}^{\backslash\mathrm{ip}}$ to output $\backslash\mathbf{y}^{\backslash\mathrm{ip}}$
- e.g., from input features $\backslash\mathbf{x}^{\backslash\mathrm{ip}}$ to
    - continuous value $\backslash\mathbf{y}^{\backslash\mathrm{ip}}$
    - discrete class $\backslash\mathbf{y}^{\backslash\mathrm{ip}}$
        - really: output is a probability vector over finite set C of discrete classes


0
2
4
6
8
10
0
2
4
6
8
10
1.8
1.9
2.0
2.1
2.2
2.075
2.050
2.025
2.000
1.975
1.950


0
2
4
6
8
10
0
2
4
6
8
10
1.8
1.9
2.0
2.1
2.2
2.075
2.050
2.025
2.000
1.975
1.950

The Neural Network is trained ("learns") to mimic the training data

- by solving for the weights $\backslash\mathbf{W}_{\backslash\mathrm{llp}}$ of each layer $1 \le \ll \le L$
- that minimize a loss function

$$\backslash\mathbf{loss} = \sum_{i-1}^{m} \backslash\mathbf{loss}^{\backslash\mathrm{ip}}$$

- where $\backslash\mathbf{loss}^{\backslash\mathrm{ip}}$ if a function of
    - how much prediction $\backslash\hat{\mathbf{y}}^{\backslash\mathrm{ip}}$ deviates from true target/label $\backslash\mathbf{y}^{\backslash\mathrm{ip}}$

The minimization procedure is usually a variant of *Gradient Descent*

The challenge is that the operation of each layer $1 \leq \ll < L$

- is usually not *interpretable*
- we can describe *how* it transforms $\backslash\mathbf{y}_{(\ll-1)}$ to $\backslash\mathbf{y}_{\backslash\mathrm{llp}}$
- but not *why* it is performing the transformation
    - objective (describe) rather than subjection (why)

So the sub-network in the diagram

In [9]:
```
print("Done")
```

```
Done
```