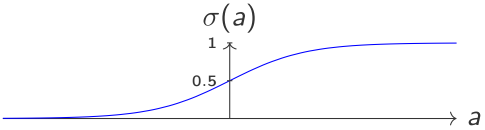
Activation Functions

- Recall that a perceptron has a non-differentiable activation function, i.e., step function.
  - Zero-derivative everywhere except at 0 where it is non-differentiable.
- Prevents gradient descent.
- Can we use a smooth activation function that behaves similar to a step function?
- Perceptron with a smooth activation function is called a *neuron*.
- Neural networks are also called multilayer perceptrons (MLP) even though they do not contain any perceptron.

Logistic Sigmoid Function

- For $a \in \mathbb{R}$, the *logistic sigmoid* function is given by $\sigma(a) = \frac{1}{1+e^{-a}}$
- *Sigmoid* means S-shaped.
- Maps $-\infty \leq a \leq \infty$ to the range $0 \leq \sigma \leq 1$. Also called *squashing* function.
- **Can be treated as a probability value.**
- Symmetry $\sigma(-a) = 1 - \sigma(a)$. **Prove it.**
- Easy derivative $\sigma' = \sigma(1 - \sigma)$. **Prove it.**



Activation Functions

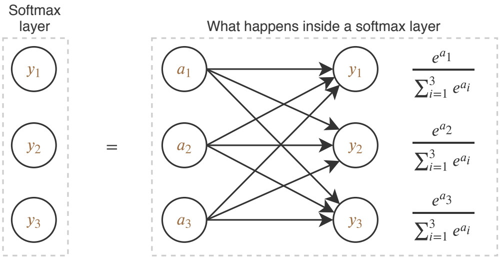
Regression

- Univariate: use 1 output neuron with identity activation function $y(a) = a$.
- Multivariate: use K output neurons with identity activation functions $y(a_k) = a_k$.

Classification

- Binary: use 1 output neuron with logistic sigmoid $y(a) = \sigma(a)$.
- Multiclass: use K output neurons with *softmax* activation function.

Softmax Activation Function



- For real numbers $a_1, \ldots, a_K$, the *softmax* function is given by

$$y(a_k; a_1, a_2, \ldots, a_K) = \frac{e^{a_k}}{\sum_{i=1}^{K} e^{a_i}}$$

- Output of k-th neuron depends on activations of *all neurons in the same layer*.

Softmax Activation Function

- Softmax is ≈ 1 when $a_k >> a_j \; \forall j \neq k$ and ≈ 0 otherwise.
- Provides a smooth (differentiable) approximation to finding the *index of the maximum element.*
  - Compute softmax for $1, 10, 100$.
  - Does not work everytime.
    - Compute softmax for $1, 2, 3$. Solution: multiply by 100.
    - Compute softmax for $1, 10, 1000$. Solution: subtract maximum before computing softmax.
- Also called the *normalized exponential* function.
- Since $0 \leq y_k \leq 1$ and $\sum_{k=1}^{K} y_k = 1$, *softmax outputs can be treated as probability values.*
- Show that $\frac{\partial y_k}{\partial a_j} = y_k(\delta_{jk} - y_j)$ where $\delta_{jk} = 1$ if $j = k$ and 0 otherwise.