Mathematical techniques in data science

Lecture 8: Neural networks

Reminders

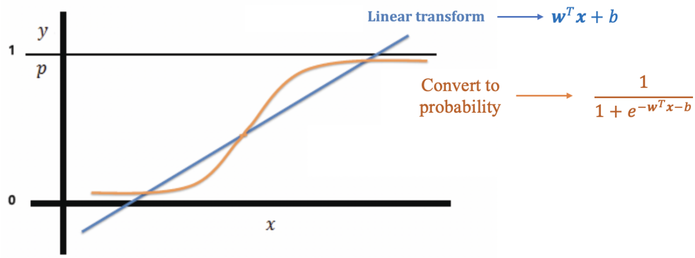
- Office hours:
  - MW 3pm-4pm, Ewing Hall 312
  - By appointments
- Homework 1: due 03/07
- Final project's description now available on the course's webpage
- Sign up for group projects by the end of Week 4

Logistic regression

- Data point $(\mathbf{x}, y)$ where
  - $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ is a vector with d features
  - y is the label (0 or 1)
- Logistic regression models $P[y = 1|X = \mathbf{x}]$

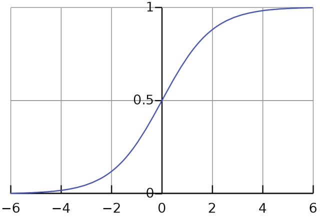
Logistic regression

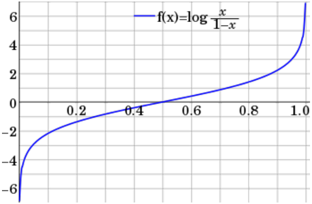
$$\boldsymbol{x} \xrightarrow{\text{Linear transform}} \boldsymbol{w}^T\boldsymbol{x} + b \xrightarrow{\text{Convert to probability}} \frac{1}{1 + e^{-\boldsymbol{w}^T\boldsymbol{x} - b}}$$



Logistic function and logit function

Transformation between $(-\infty, \infty)$ and $[0, 1]$





$$f(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

$$logit(p) = \log \frac{p}{1-p}$$

Logistic regression with more than 2 classes

- Suppose now the response can take any of $\{1, \ldots, K\}$ values
- We use the categorical distribution instead of the Bernoulli distribution

$$P[Y = k|X = \mathbf{x}] = p_k(\mathbf{x}), \quad \sum_{k=1}^{K} p_k(\mathbf{x}) = 1.$$

- Model

$$p_k(\mathbf{x}) = \frac{e^{w_k^T \mathbf{x}_k + b_k}}{\sum_{k=1}^{K} e^{w_k^T \mathbf{x}_k + b_k}}$$

Logistic regression: pros and cons

Pros:

- Simple algorithm
- Prediction is fast
- Easy to implement
- The forward map has a closed-form formula of the derivatives

$$\frac{\partial \ell}{\partial \beta_j}(\beta) = \sum_{i=1}^{n} \left[y_i x_{ij} - x_{ij} \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right].$$

Cons:

- Linear model

How to make logistic regression better?

We want a model that

- compute the derivatives (of the objective function, with respect to the parameters) easily
- can capture complex relationships

This is difficult because complex models often have high numbers of parameters and don't have closed-form derivatives, and computations of

$$\frac{\partial \ell}{\partial \beta_i}(x) \approx \frac{\ell(x + \epsilon_i) - \ell(x)}{\epsilon_i}$$

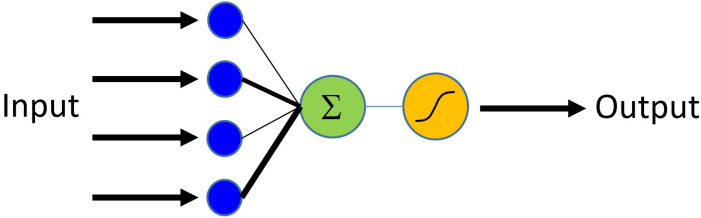
are large (and unstable)

Next lecture

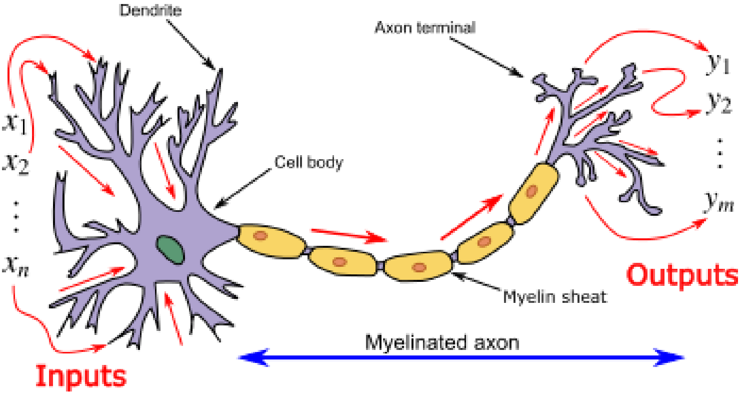
- Automatic differentiation and back-propagation
- Ideas:
  - Organizing informations using graphs (networks)
  - Chain rule

$$(f \circ g)'(x) = f'(g(x))g'(x)$$

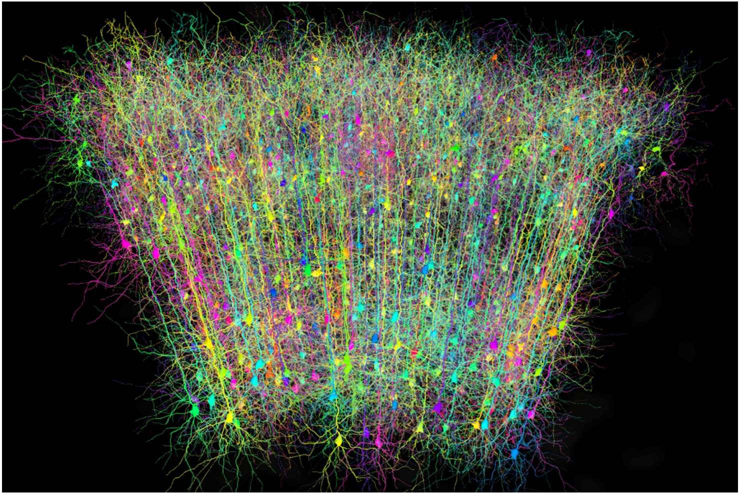
Logistic neuron



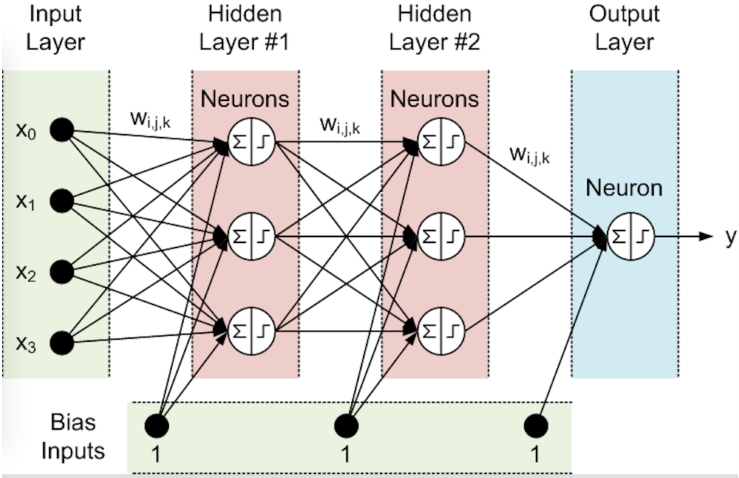
Why neuron?


Dendrite
Axon terminal
Cell body
Myelin sheat
Myelinated axon
x_1
x_2
$\vdots$
x_n
Inputs
y_1
y_2
$\vdots$
y_m
Outputs

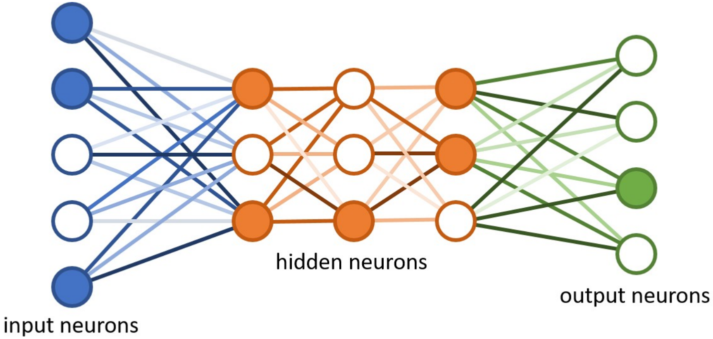
Neural circuit



Feed-forward neural networks



Feed-forward neural networks



Feed-forward neural networks