Course number: 80240743

Deep Learning

Xiaolin Hu (胡晓林) & Jun Zhu (朱军)
Dept. of Computer Science and Technology
Tsinghua University

Topic 3: Multi-layer Perceptron

Xiaolin Hu
Dept. of Computer Science and Technology
Tsinghua University

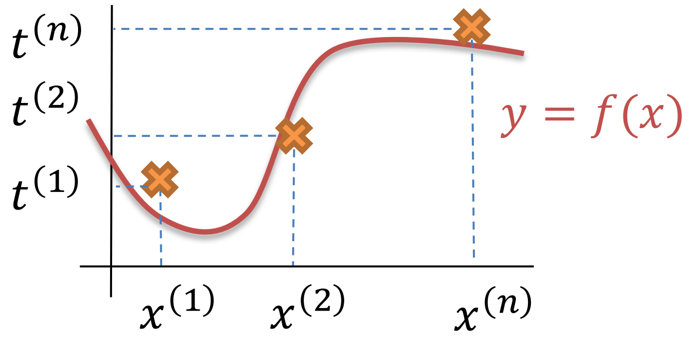
Outline

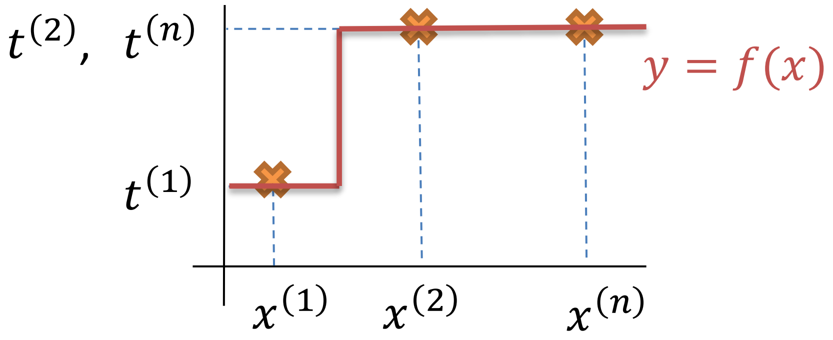
- Regression and classification
- Multi-layer perceptron
- Backpropagation
- Optimization techniques
- Coding considerations

Regression and classification

Given a set of data points $\boldsymbol{x}^{(n)} \in R^m$ and the corresponding labels $t^{(n)} \in \Omega$: $\{(\boldsymbol{x}^{(1)}, t^{(1)}), (\boldsymbol{x}^{(2)}, t^{(2)}), \dots, (\boldsymbol{x}^{(N)}, t^{(N)})\}$, for a new data point $\boldsymbol{x}$, predict its label

- The goal is to find a mapping
  $$f: R^m \to \Omega$$
- If Ω is a continuous set, this is called regression
- If Ω is a discrete set, this is called classification



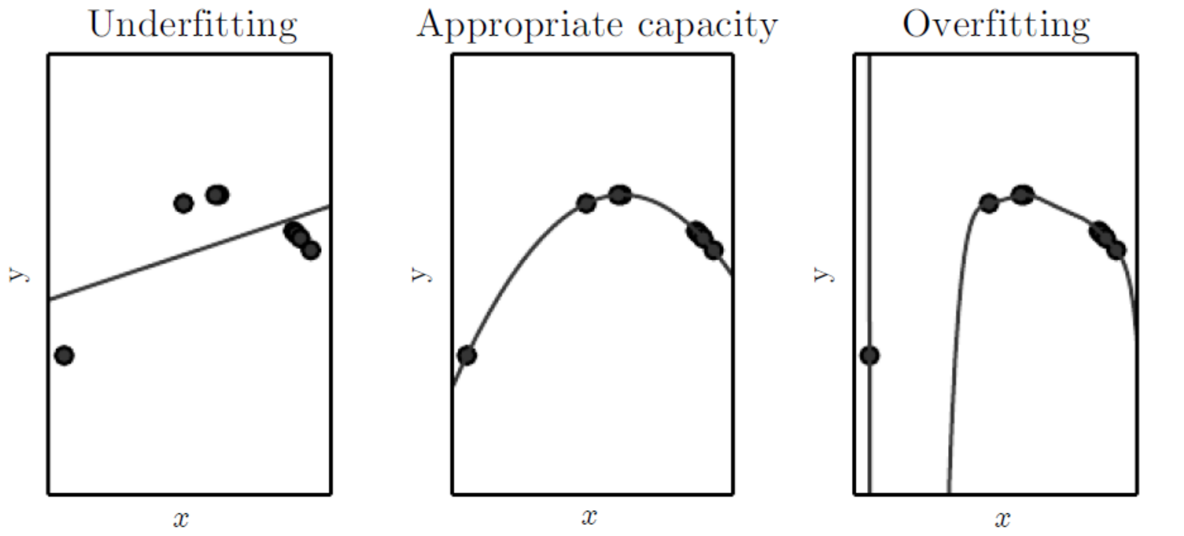


Recall: polynomial regression

- Consider a regression problem in which the input x and output y are both scalars. Find a function $f: \mathbb{R} \to \mathbb{R}$ to fit the data
  - $f(x) = b + wx$
  - $f(x) = b + w_1 x + w_2 x^2$
  - $f(x) = b + \sum_{i=1}^{9} w_i x^i$

MSE training:

$$\min_w \frac{1}{N} \sum_{n=1}^{N} \left\| f\left(x^{(n)}\right) - y^{(n)} \right\|_2^2$$



Linear regression

- $f(\boldsymbol{x})$ is linear

$$f(x) = \boldsymbol{w}^\top \boldsymbol{x} + \overset{\text{bias}}{\boxed{b}}$$

where $\boldsymbol{w} \in R^m, b \in R$.

- Choose the cost function as the mean squared error (MSE)

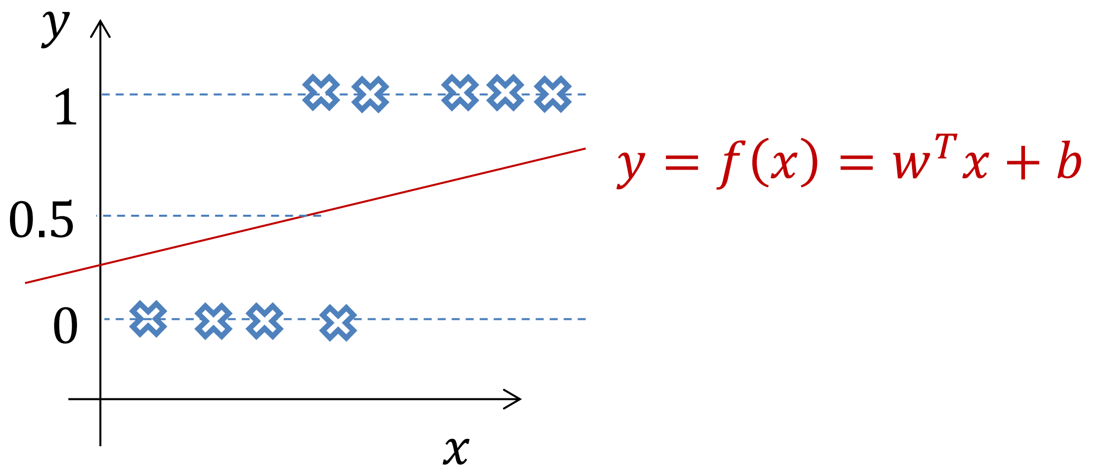
$$E = \frac{1}{2}\sum_{n=1}^{N}\left(f\left(\boldsymbol{x}^{(n)}\right) - t^{(n)}\right)^2 = \frac{1}{2}\sum_{n=1}^{N}\left(\boldsymbol{w}^\top \boldsymbol{x}^{(n)} + b - t^{(n)}\right)^2$$

- Find optimal $\boldsymbol{w}^*$ and $\boldsymbol{b}^*$ by minimizing the cost function

$$\left.\begin{aligned} \nabla_{\boldsymbol{w}} E &= \sum_{n=1}^{N}\left(\boldsymbol{w}^\top \boldsymbol{x}^{(n)} + b - t^{(n)}\right)\boldsymbol{x}^{(n)} = 0 \\ \nabla_b E &= \sum_{n=1}^{N}\left(\boldsymbol{w}^\top \boldsymbol{x}^{(n)} + b - t^{(n)}\right) = 0 \end{aligned}\right\} \boldsymbol{w}^*, b^*$$

Do classification using regression

- Suppose $t \in \{0,1\}$. Consider the 1D case



- Regression
  - Prediction $y = f(x)$ which is continuous
- Classification
  - Prediction $y = \begin{cases} 1, \text{if } f(x) \geq 0.5 \\ 0, \text{if } f(x) < 0.5 \end{cases}$

Representation of class labels

- For classification, given $\{(\boldsymbol{x}^{(1)}, t^{(1)}), \dots, (\boldsymbol{x}^{(N)}, t^{(N)})\}$, the goal is to find a mapping from $\boldsymbol{x}^{(n)}$ to $t^{(n)}$

$$f: R^m \to \Omega$$

where Ω is a discrete set

- $t^{(n)}$ can be a (discrete) scalar or vector

Suppose there are 5 classes in total

Seldom used

Scalar representation

$t^{(1)} = 1$

$t^{(3)} = 3$

Vector representation

$\boldsymbol{t}^{(1)} = (1, 0, 0, 0, 0)^{\mathsf{T}}$

$\boldsymbol{t}^{(3)} = (0, 0, 1\ 0, 0)^{\mathsf{T}}$

Usually used

- 1-of-K representation
- Property: $t_k{}^{(n)} \in \{0,1\}$; $\sum_k t_k^{(n)} = 1$

Linear regression for vectors

- Using the 1-of-K representation for class labels, one can do classification using linear regression
- $f_k(\boldsymbol{x})$ is linear for $k = 1, \dots, K$: $f_k(\boldsymbol{x}) = \boldsymbol{w}_k^\top \boldsymbol{x} + b_k$, where $\boldsymbol{w}_{\boldsymbol{k}} \in R^m, b_k \in R$.
- Choose the cost function as the mean squared error (MSE)

$$E = \frac{1}{2N} \sum_{n=1}^{N} \sum_{k=1}^{K} \left(f_k\left(\boldsymbol{x}^{(n)}\right) - t_k^{(n)} \right)^2$$

- Find optimal $\boldsymbol{w}_k^*$ and $\boldsymbol{b}_k^*$ by solving the linear system

$$\left.\begin{aligned} \nabla_{\boldsymbol{w}_{\boldsymbol{k}}} E &= \frac{1}{N} \sum_{n=1}^{N} \left(\boldsymbol{w}_k^\top \boldsymbol{x}^{(n)} + b_k - t_k^{(n)} \right) \boldsymbol{x}^{(n)} = 0 \\ \nabla_{b_k} E &= \frac{1}{N} \sum_{n=1}^{N} \left(\boldsymbol{w}_k^\top \boldsymbol{x}_k^{(n)} + b_k - t_k^{(n)} \right) = 0 \end{aligned}\right\} \boldsymbol{w}_k^*, b_k^*$$

Nonlinear regression for vectors

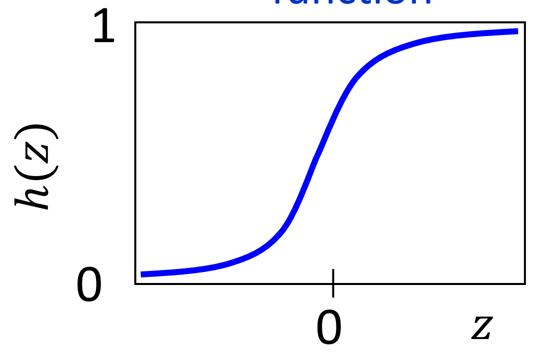
- $f_k(\boldsymbol{x})$ is nonlinear for $k = 1, \ldots, K$

  $$f_k(\boldsymbol{x}) = h\left(\boldsymbol{w}_k^\top \boldsymbol{x} + b_k\right)$$

  where $\boldsymbol{w_k} \in R^m, b_k \in R$, and

  $$h(z) = \frac{1}{1 + \exp(-z)}$$

logistic sigmoid function



- Choose the cost function as the MSE

  $$E = \frac{1}{2N} \sum_{n=1}^{N} \sum_{k=1}^{K} \left(f_k\left(\boldsymbol{x}^{(n)}\right) - t_k^{(n)}\right)^2$$

Local sensitivity or local gradient

- Denote $u_k = \boldsymbol{w}_k^\top \boldsymbol{x} + b_k$ and $\boxed{\delta_k = \frac{\partial E}{\partial u_k},}$ then $\frac{\partial E}{\partial \boldsymbol{w}_k} = \delta_k \frac{\partial u_k}{\partial \boldsymbol{w}_k}$ and $\frac{\partial E}{\partial b_k} = \delta_k \frac{\partial u_k}{\partial b_k}$

$$\delta_k = (f(u_k) - t_k) f'(u_k)$$

Vector-matrix form

- Define

$$\boldsymbol{W} = \begin{pmatrix} w_{11} & \cdots & w_{1m} \\ \vdots & \vdots & \vdots \\ w_{K1} & \cdots & w_{Km} \end{pmatrix} \quad \frac{\partial E}{\partial \boldsymbol{W}} = \begin{pmatrix} \partial E/\partial w_{11} & \cdots & \partial E/\partial w_{1m} \\ \vdots & \vdots & \vdots \\ \partial E/\partial w_{K1} & \cdots & \partial E/\partial w_{Km} \end{pmatrix}$$

N: the number of inputs; M: the number of outputs

- Output: $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$, where $\boldsymbol{f}, \boldsymbol{h}, \boldsymbol{b} \in R^K$, $\boldsymbol{x} \in R^m$
- Error function: $E = \frac{1}{2N} \sum_{n=1}^{N} \left|\left|\boldsymbol{f}\left(\boldsymbol{x}^{(n)}\right) - \boldsymbol{t}^{(n)}\right|\right|_2^2$
- Gradient

Elementwise product

$$\nabla_{\boldsymbol{W}} E = \frac{1}{N} \sum_{n=1}^{N} \left(\boldsymbol{f}\left(\boldsymbol{x}^{(n)}\right) - \boldsymbol{t}^{(n)}\right) \odot \boldsymbol{f}'\left(\boldsymbol{x}^{(n)}\right) \boldsymbol{x}^{(n)}$$

$$\nabla_{\boldsymbol{b}} E = \frac{1}{N} \sum_{n=1}^{N} \left(\boldsymbol{f}\left(\boldsymbol{x}^{(n)}\right) - \boldsymbol{t}^{(n)}\right) \odot \boldsymbol{f}'\left(\boldsymbol{x}^{(n)}\right)$$

Normal distribution assumption

- Assume the label follows a normal distribution with mean $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$:

$$p(\underline{\mathbf{t}}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{t}; \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{I}) \propto \exp\left(-\left|\left|\boldsymbol{t} - \boldsymbol{f}(\boldsymbol{x})\right|\right|_2^2\right)$$

(Hereafter we don't distinguish between plain and italic type faces)

- Given a dataset $\left\{\left(\boldsymbol{x}^{(1)}, \boldsymbol{t}^{(1)}\right), \ldots, \left(\boldsymbol{x}^{(N)}, \boldsymbol{t}^{(N)}\right)\right\}$. View $\boldsymbol{t}$ and $\boldsymbol{x}$ as random variables. The conditional data likelihood function (independence assumption)

$$P\left(\boldsymbol{t}^{(1)}, \ldots, \boldsymbol{t}^{(N)} \middle| \boldsymbol{X}\right) = \Pi_{n=1}^{N} P(\boldsymbol{t}^{(n)}|\boldsymbol{x}^{(n)})$$

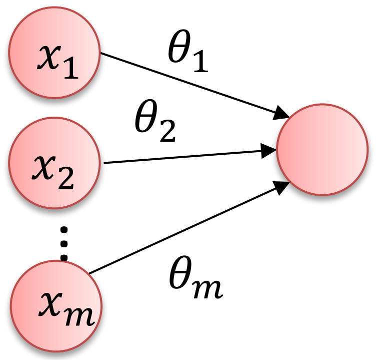
- $\max \log P\left(\boldsymbol{t}^{(1)}, \ldots, \boldsymbol{t}^{(N)} \middle| \boldsymbol{X}\right)$ is equivalent to $\min E = \frac{1}{2N} \sum_{n=1}^{N} \left|\left|\boldsymbol{f}\left(\boldsymbol{x}^{(n)}\right) - \boldsymbol{t}^{(n)}\right|\right|_2^2$

Motivation for other cost functions for classification

- For regression ($\boldsymbol{t}$ is continuous), the normal distribution assumption is natural
  - How to predict the label of a test sample $\boldsymbol{x}$ after obtaining $\boldsymbol{w}^*$ and $\boldsymbol{b}^*$?
- For classification ($\boldsymbol{t}$ is one-hot vector), it is not so natural (why?)
  - How to predict the label of a test sample $\boldsymbol{x}$ after obtaining $\boldsymbol{w}^*$ and $\boldsymbol{b}^*$?
- We have more suitable assumptions on the data distribution for classification
  - Bernoulli distribution
  - Multinoulli or categorical distribution

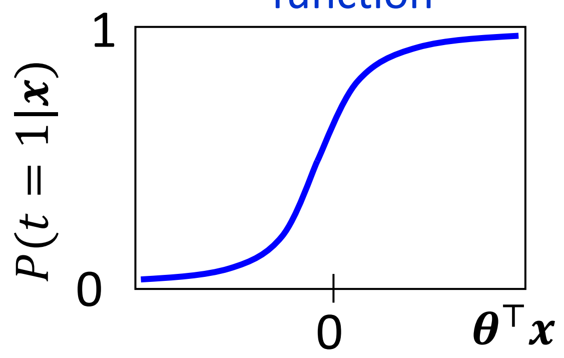
Logistic regression

- For 2-class problems, one 0-1 unit is enough for representing a label



$t^{(1)}$	$t^{(2)}$	$t^{(3)}$	…
1	0	1	…

logistic sigmoid function



- We try to learn a conditional probability (we've absorbed b in θ)

$$P(t=1|\boldsymbol{x}) = \frac{1}{1+\exp(-\theta^\top \boldsymbol{x})} \triangleq h(\boldsymbol{x})$$

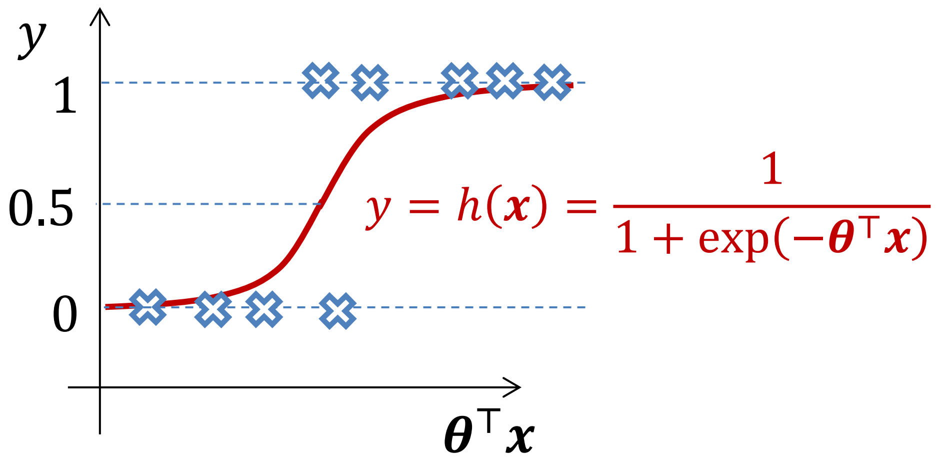
$$P(t=0|\boldsymbol{x}) = 1 - P(t=1|\boldsymbol{x}) = 1 - h(\boldsymbol{x})$$

where $\boldsymbol{x}$ is input and t is label

$P(t=1|\boldsymbol{x})$ is a Bernoulli distribution

Logistic regression

- Our goal is to search for a value of $\boldsymbol{\theta}$ so that the probability $P(t = 1|\boldsymbol{x}) = h(\boldsymbol{x})$ is
  - large when $\boldsymbol{x}$ belongs to class 1 and
  - small when $\boldsymbol{x}$ belongs to class 0 (so that $P(t = 0|\boldsymbol{x})$ is large)
- We're essentially "regress" a discrete function ($x \rightarrow t$) using another function h which is continuous



Regression

Prediction $y = h(\boldsymbol{x})$

Classification

Prediction $y = \begin{cases} 1, \text{if } h(\boldsymbol{x}) \geq 0.5 \\ 0, \text{if } h(\boldsymbol{x}) < 0.5 \end{cases}$

Or equivalently

$$y = \begin{cases} 1, \text{if } \boldsymbol{\theta}^\top \boldsymbol{x} \geq 0 \\ 0, \text{if } \boldsymbol{\theta}^\top \boldsymbol{x} < 0 \end{cases}$$

Maximum conditional data likelihood

- *Recall* the maximum conditional likelihood estimation:
  - 1. write down the conditional likelihood function
  - 2. take log and maximize
- Given a dataset $\left\{\left(\boldsymbol{x}^{(1)}, t^{(1)}\right), \ldots, \left(\boldsymbol{x}^{(N)}, t^{(N)}\right)\right\}$ where $t^{(n)} \in \{0,1\}$
- View t as a Bernoulli variable and $P(t = 1|\boldsymbol{x}) = h(\boldsymbol{x}; \boldsymbol{\theta})$. The conditional likelihood function

$$P(t^{(1)}, \ldots, t^{(N)}|X; \boldsymbol{\theta}) = \Pi_{n=1}^{N} h(\boldsymbol{x}^{(n)})^{t^{(n)}} (1 - h(\boldsymbol{x}^{(n)}))^{1-t^{(n)}}$$

- Maximizing the likelihood is equivalent to minimizing

$$\begin{aligned} E(\boldsymbol{\theta}) &= -\frac{1}{N} \ln P(t^{(1)}, \ldots, t^{(N)}) \\ &= -\frac{1}{N} \sum_{n=1}^{N} \left(t^{(n)} \ln h(\boldsymbol{x}^{(n)}) + (1 - t^{(n)}) \ln(1 - h(\boldsymbol{x}^{(n)})) \right) \end{aligned}$$

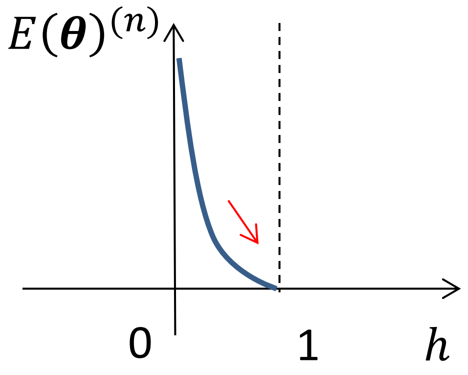
Cross-entropy error function

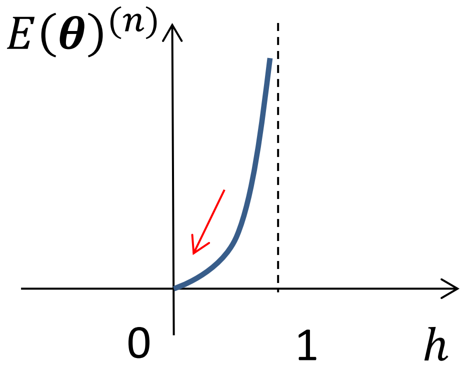
- For a set of training examples with binary labels $\{(\boldsymbol{x}^{(n)}, t^{(n)}) : n = 1, \dots, N\}$ define the *cross-entropy* error function

$$E(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{n=1}^{N}\left(t^{(n)}\ln(h(\boldsymbol{x}^{(n)})) + (1-t^{(n)})\ln(1-h(\boldsymbol{x}^{(n)}))\right)$$

$$E(\boldsymbol{\theta})^{(n)} = -t^{(n)}\ln(h(\boldsymbol{x}^{(n)})) - (1-t^{(n)})\ln(1-h(\boldsymbol{x}^{(n)}))$$

- If $t^{(n)} = 1$, then
  $E(\boldsymbol{\theta})^{(n)} = -\ln(h)$
- If $t^{(n)} =0$, then
  $E(\boldsymbol{\theta})^{(n)} = -\ln(1-h)$





Training and testing

$$E(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{n=1}^{N}\left(t^{(n)}\ln h(\boldsymbol{x}^{(n)}) + (1-t^{(n)})\ln(1-h(\boldsymbol{x}^{(n)}))\right)$$

- Calculate the gradient (exercise)

$$h(z) = \frac{1}{1+\exp(-z)}$$

$$\frac{\partial h}{\partial z} = h(1-h)$$

- Some regularization term can be incorporated into the cost function

$$J(\boldsymbol{\theta}) = E(\boldsymbol{\theta}) + \lambda\left|\left|\boldsymbol{\theta}\right|\right|^2/2$$

- Training: learn $\boldsymbol{\theta}$ to minimize the cost function

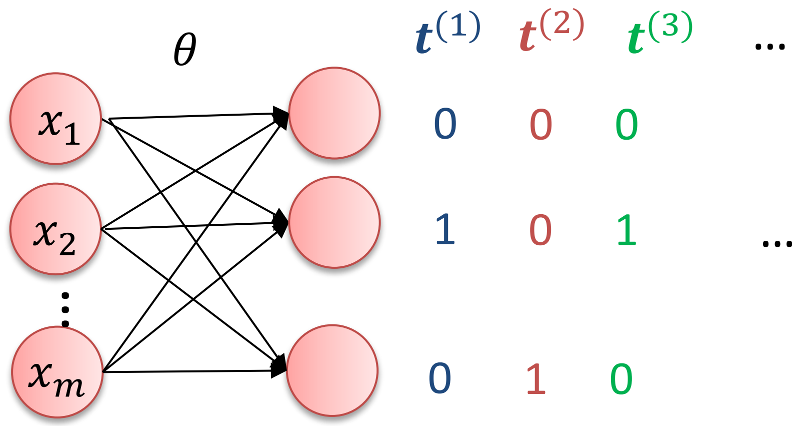
$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha\nabla J(\boldsymbol{\theta})$$

  where α is the learning rate

- Testing: for a new input $\boldsymbol{x}$, if $P(t=1|\boldsymbol{x}) > P(t=0|\boldsymbol{x})$ then we predict the input as class 1, and 0 otherwise

More than two classes

- For K-class problems ($K > 2$), we try to learn a Multinoulli distribution $P(t = k|\boldsymbol{x})$ where $k = 1, \dots, K$
- With one-hot representation $\boldsymbol{t} = (0, \dots, 1, \dots, 0)^\top$, then $P(\boldsymbol{t}|\boldsymbol{x}) = \Pi_{k=1}^{K} P(t_k = 1)^{t_k}$



θ
x_1
x_2
x_m
$\boldsymbol{t}^{(1)}$ $\boldsymbol{t}^{(2)}$ $\boldsymbol{t}^{(3)}$ …
0 0 0
1 0 1 …
0 1 0

- Let $P(\boldsymbol{t}|\boldsymbol{x})$ take the following form

$$\boldsymbol{h}(\boldsymbol{x}) \triangleq \begin{bmatrix} P(t_1 = 1|\boldsymbol{x};\boldsymbol{\theta}) \\ P(t_2 = 1|\boldsymbol{x};\boldsymbol{\theta}) \\ \vdots \\ P(t_K = 1|\boldsymbol{x};\boldsymbol{\theta}) \end{bmatrix} = \frac{1}{\sum_{j=1}^{K} \exp(\boldsymbol{\theta}^{(j)\top}\boldsymbol{x})} \begin{bmatrix} \exp(\boldsymbol{\theta}^{(1)\top}\boldsymbol{x}) \\ \exp(\boldsymbol{\theta}^{(2)\top}\boldsymbol{x}) \\ \vdots \\ \exp(\boldsymbol{\theta}^{(K)\top}\boldsymbol{x}) \end{bmatrix}$$

More than two classes

Then $$h_k(\boldsymbol{x}) = P(t_k = 1|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}^{(k)\top}\boldsymbol{x})}{\sum_{j=1}^{K}\exp(\boldsymbol{\theta}^{(j)\top}\boldsymbol{x})}$$

- Given a test input $\boldsymbol{x}$, estimate $P(t_k = 1|\boldsymbol{x})$ for each value of $k = 1, \dots, K$.
- Goal: search for a value of $\boldsymbol{\theta}$ so that the probability $P(t_k = 1|\boldsymbol{x})$ is
  - large when $\boldsymbol{x}$ belongs to the k-th class and
  - small when $\boldsymbol{x}$ belongs to other classes

where $$\boldsymbol{\theta} = \begin{bmatrix} | & | & | & | \\ \boldsymbol{\theta}^{(1)} & \boldsymbol{\theta}^{(2)} & \cdots & \boldsymbol{\theta}^{(K)} \\ | & | & | & | \end{bmatrix}^\top.$$

- Since $h_k(\boldsymbol{x})$ is a (continuous) probability, we need to transform it into discrete values for classification ←How?

Softmax function

$$h_k(\boldsymbol{x}) = P(t_k = 1|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}^{(k)\top}\boldsymbol{x})}{\sum_{j=1}^{K}\exp(\boldsymbol{\theta}^{(j)\top}\boldsymbol{x})}$$

- The following function is called *softmax* function

$$\psi(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)} = \frac{\exp(z_i)}{\exp(z_i) + \sum_{j\neq i}\exp(z_j)} \in (0,1)$$

- If $z_i > z_j$ for all $j \neq i$
  - Then $\psi(z_i) > \psi(z_j)$ for all $j \neq i$ but it is smaller than 1
- If $z_i \gg z_j$ for all $j \neq i$,
  - then $\psi(z_i) \rightarrow 1$ and $\psi(z_j) \rightarrow 0$ for $j \neq i$.

Maximum conditional likelihood

- Since
$$P(\boldsymbol{t}|\boldsymbol{p}) = \Pi_{k=1}^{K} P(t_k = 1)^{t_k}$$
- Given a dataset $\{(\boldsymbol{x}^{(1)}, \boldsymbol{t}^{(1)}), \dots, (\boldsymbol{x}^{(N)}, \boldsymbol{t}^{(N)})\}$. The conditional likelihood function (independence assumption):
$$P(\boldsymbol{t}^{(1)}, \dots, \boldsymbol{t}^{(N)} | \boldsymbol{X}) = \Pi_{n=1}^{N} \Pi_{k=1}^{K} P(\boldsymbol{t}_k^{(n)} = 1 | \boldsymbol{x}^{(n)})^{t_k^{(n)}}$$
- Maximizing this likelihood function is equivalent to minimizing

$$E(\boldsymbol{\theta}) = -\frac{1}{N} \ln P(\boldsymbol{t}^{(1)}, \dots, \boldsymbol{t}^{(N)})$$
$$= -\frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} t_k^{(n)} \ln \frac{\exp(\boldsymbol{\theta}^{(k)\top} \boldsymbol{x}^{(n)})}{\sum_{j=1}^{K} \exp(\boldsymbol{\theta}^{(j)\top} \boldsymbol{x}^{(n)})}$$

Take log and negative, then minimize

Cross-entropy error function

$$
\begin{aligned}
E(\boldsymbol{\theta}) &= -\frac{1}{N}\ln P(\boldsymbol{t}^{(1)},\dots,\boldsymbol{t}^{(N)}) \\
&= -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} t_k^{(n)} \ln \frac{\exp(\boldsymbol{\theta}^{(k)\top}\boldsymbol{x}^{(n)})}{\sum_{j=1}^{K}\exp(\boldsymbol{\theta}^{(j)\top}\boldsymbol{x}^{(n)})}
\end{aligned}
$$

- This function is called cross-entropy error function
- To derive its gradient, let's first decompose it into a multiple functions

$$
E(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} E^{(n)}(\boldsymbol{\theta}), \quad E^{(n)}(\boldsymbol{\theta}) = -\sum_{i=1}^{K} t_i^{(n)} \ln h_i^{(n)}
$$

where $h_i^{(n)} = P(t_i^{(n)} = 1|\boldsymbol{x}^{(n)}) = \dfrac{\exp(u_i^{(n)})}{\sum_{j=1}^{K}\exp(u_j^{(n)})}, \quad u_k^{(n)} = \boldsymbol{\theta}^{(k)\top}\boldsymbol{x}^{(n)}$

Recap: Derivative of two-step composition

Suppose we have:

- Independent input variables $x_1, x_2, \dots, x_n$
- Dependent intermediate variables, $u_1, u_2, \dots, u_m$, each of which is a function of $x_1, x_2, \dots, x_n$
- Dependent output variables $w_1, w_2, \dots, w_p$, each of which is a function of $u_1, u_2, \dots, u_m$

Then for any $i \in \{1,2,\dots,p\}$ and $j \in \{1,2,\dots,n\}$ we have

$$\frac{\partial w_i}{\partial x_j} = \sum_{k=1}^{m} \frac{\partial w_i}{\partial u_k} \frac{\partial u_k}{\partial x_j}$$

Sum over the intermediate variables

From wikipedia

In this slide x, w, u denote general variables and they should not be confused with the notations in other slides!

Calculate the gradient

$$E(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} E^{(n)}(\boldsymbol{\theta}), \quad E^{(n)}(\boldsymbol{\theta}) = -\sum_{i=1}^{K} t_i^{(n)} \ln h_i^{(n)}$$

where $h_i^{(n)} = P(t_i^{(n)} = 1|\boldsymbol{x}^{(n)}) = \frac{\exp(u_i^{(n)})}{\sum_{j=1}^{K}\exp(u_j^{(n)})}, \quad u_k^{(n)} = \boldsymbol{\theta}^{(k)\top}\boldsymbol{x}^{(n)}$

$$\frac{\partial E^{(n)}}{\partial \boldsymbol{\theta}^{(k)}} = \frac{\partial E^{(n)}}{\partial u_k^{(n)}} \frac{\partial u_k^{(n)}}{\partial \boldsymbol{\theta}^{(k)}} = \sum_{i=1}^{K} \frac{\partial E^{(n)}}{\partial h_i^{(n)}} \frac{\partial h_i^{(n)}}{\partial u_k^{(n)}} \frac{\partial u_k^{(n)}}{\partial \boldsymbol{\theta}^{(k)}}$$

Local sensitivity or local gradient

$$\frac{\partial E^{(n)}}{\partial h_i^{(n)}} = -t_i^{(n)} \frac{1}{h_i^{(n)}}$$

?

$$\frac{\partial u_k^{(n)}}{\partial \boldsymbol{\theta}^{(k)}} = \boldsymbol{x}^{(n)}$$

Calculate the gradient

$$E(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} E^{(n)}(\boldsymbol{\theta}), \quad E^{(n)}(\boldsymbol{\theta}) = -\sum_{i=1}^{K} t_i^{(n)} \ln h_i^{(n)}$$

where $h_i^{(n)} = P(t_i^{(n)} = 1 | \boldsymbol{x}^{(n)}) = \frac{\exp(u_i^{(n)})}{\sum_{j=1}^{K} \exp(u_j^{(n)})}, \quad u_k^{(n)} = \boldsymbol{\theta}^{(k)\top} \boldsymbol{x}^{(n)}$

If $k \neq i$, u_k appears only in the denominator

$$\frac{\partial h_i^{(n)}}{\partial u_k^{(n)}} =$$

If $k = i$, u_k appears in both numerator and denominator

$$\frac{\partial h_i^{(n)}}{\partial u_k^{(n)}} =$$

Therefore $\frac{\partial h_i^{(n)}}{\partial u_k^{(n)}} = h_i^{(n)}(\Delta_{i,k} - h_k^{(n)})$ where $\Delta_{i,k} = \begin{cases} 1, & \text{if } i = k \\ 0, & \text{else.} \end{cases}$

Calculate the gradient

$$E(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} E^{(n)}(\boldsymbol{\theta}), \quad E^{(n)}(\boldsymbol{\theta}) = -\sum_{i=1}^{K} t_i^{(n)} \ln h_i^{(n)}$$

where $h_i^{(n)} = P(t_i^{(n)} = 1|\boldsymbol{x}^{(n)}) = \frac{\exp(u_i^{(n)})}{\sum_{j=1}^{K}\exp(u_j^{(n)})}, \quad u_k^{(n)} = \boldsymbol{\theta}^{(k)\top}\boldsymbol{x}^{(n)}$

$$\begin{aligned}
\frac{\partial E^{(n)}}{\partial \boldsymbol{\theta}^{(k)}} &= \sum_{i=1}^{K} \frac{\partial E^{(n)}}{\partial h_i^{(n)}} \frac{\partial h_i^{(n)}}{\partial u_k^{(n)}} \frac{\partial u_k^{(n)}}{\partial \boldsymbol{\theta}^{(k)}} \\
&= \sum_{i=1}^{K} \left(-t_i^{(n)} \frac{1}{h_i^{(n)}}\right) \left(h_i^{(n)}(\Delta_{i,k} - h_k^{(n)})\right) \left(\boldsymbol{x}^{(n)}\right) \\
&= -\left(\sum_{i=1}^{K} t_i^{(n)}\Delta_{i,k} - \underbrace{\sum_{i=1}^{K} t_i^{(n)}}_{=1} h_k^{(n)}\right)\boldsymbol{x}^{(n)} \\
&= -\left(t_k^{(n)} - h_k^{(n)}\right)\boldsymbol{x}^{(n)}
\end{aligned}$$

Calculate the gradient

$$E(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} E^{(n)}(\boldsymbol{\theta}), \quad E^{(n)}(\boldsymbol{\theta}) = -\sum_{i=1}^{K} t_i^{(n)} \ln h_i^{(n)}$$

where $h_i^{(n)} = P(t_i^{(n)} = 1 | \boldsymbol{x}^{(n)}) = \frac{\exp(u_i^{(n)})}{\sum_{j=1}^{K} \exp(u_j^{(n)})}, \quad u_k^{(n)} = \boldsymbol{\theta}^{(k)\top}\boldsymbol{x}^{(n)}$

$$\frac{\partial E^{(n)}}{\partial \theta^{(k)}} = \delta_k^{(n)} \boldsymbol{x}^{(n)},$$ where $\delta_k^{(n)} \triangleq \frac{\partial E^{(n)}}{\partial u_k^{(n)}} = -\left(t_k^{(n)} - h_k^{(n)}\right)$

is the local sensitivity.

The overall gradient

$$\nabla_{\boldsymbol{\theta}^{(k)}} E(\boldsymbol{\theta}) = \sum_{n=1}^{N} \frac{\partial E^{(n)}}{\partial \boldsymbol{\theta}^{(k)}} = -\sum_{n=1}^{N} \left(t_k^{(n)} - h_k^{(n)}\right) \boldsymbol{x}^{(n)}$$

$$= -\sum_{n=1}^{N} \left(t_k^{(n)} - P(t_k^{(n)} = 1 | \boldsymbol{x}^{(n)})\right) \boldsymbol{x}^{(n)}$$

Training and testing

- Calculate the gradient of the cross-entropy error function

$$\nabla_{\boldsymbol{\theta}^{(k)}} E(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{n=1}^{N} \left(t_k^{(n)} - h_k^{(n)} \right) \boldsymbol{x}^{(n)}$$

- As before, some regularization term can be incorporated into the cost function

$$J(\boldsymbol{\theta}) = E(\boldsymbol{\theta}) + \lambda \left|\left|\boldsymbol{\theta}\right|\right|^2 / 2$$

- Training: minimize the cost function with gradient $\nabla J(\boldsymbol{\theta})$

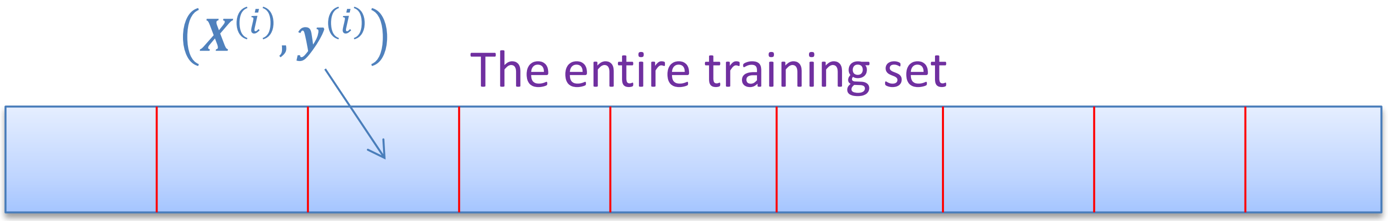
$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla J(\boldsymbol{\theta})$$

  where α is the learning rate

- Testing: find the maximum $P(t_k = 1 | \boldsymbol{x})$ among k for a new input $\boldsymbol{x}$

$$P(t_k = 1 | \boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}^{(k)\top} \boldsymbol{x})}{\sum_{j=1}^{K} \exp(\boldsymbol{\theta}^{(j)\top} \boldsymbol{x})}$$

Recall: Stochastic gradient decent



- Minimizing the cost function over the entire training set is computationally expensive
- We often decompose the training set into smaller subsets or minibatches and optimize the cost function defined over individual minibatches $(\boldsymbol{X}^{(i)}, \boldsymbol{y}^{(i)})$ and take the average

$$J(\boldsymbol{\theta}) = \frac{1}{N'} \sum_{i=1}^{N'} L(\boldsymbol{X}^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{\theta})$$

$$\boldsymbol{g} = \frac{1}{N'} \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{N'} L(\boldsymbol{X}^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{\theta})$$

$$\boldsymbol{\theta} = \boldsymbol{\theta} - \eta \boldsymbol{g}$$

- A total of N' minibatches
- The batchsize ranges from 1 to a few hundreds

Introducing bias

- So far we have assumed

$$h_k(\boldsymbol{x}) = P(t_k = 1|\boldsymbol{x}) = \frac{\exp(u_k^{(n)})}{\sum_{j=1}^{K} \exp(u_j^{(n)})} \qquad u_k^{(n)} = \boldsymbol{\theta}^{(k)\top} \boldsymbol{x}^{(n)}$$

- Sometimes a bias is introduced into $u_k^{(n)}$ and the parameters become $\{\boldsymbol{W}, \boldsymbol{b}\}$

$$u_k^{(n)} = \boldsymbol{w}^{(k)\top} \boldsymbol{x}^{(n)} + b^{(k)}$$

- It's easy to show that

$$\nabla_{\boldsymbol{w}^{(k)}} E(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{n=1}^{N} \left(t_k^{(n)} - h_k(\boldsymbol{x}^{(n)}) \right) \boldsymbol{x}^{(n)}$$

$$\nabla_{b^{(k)}} E(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{n=1}^{N} \left(t_k^{(n)} - h_k(\boldsymbol{x}^{(n)}) \right)$$

- Regularization is often applied on $\boldsymbol{W}$ only

$$J(\boldsymbol{W}, \boldsymbol{b}) = E(\boldsymbol{W}, \boldsymbol{b}) + \lambda \left|\left|\boldsymbol{W}\right|\right|^2 / 2$$

Softmax is over-parameterized

- The hypothesis

$$h_k(\boldsymbol{x}) = P(t_k = 1|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}^{(k)\top}\boldsymbol{x})}{\sum_{j=1}^{K}\exp(\boldsymbol{\theta}^{(j)\top}\boldsymbol{x})} = \frac{\exp((\boldsymbol{\theta}^{(k)} - \boldsymbol{\phi})^\top\boldsymbol{x})}{\sum_{j=1}^{K}\exp((\boldsymbol{\theta}^{(j)} - \boldsymbol{\phi})^\top\boldsymbol{x})}$$

  Then the new parameters $\widehat{\boldsymbol{\theta}}^{(k)} \equiv \boldsymbol{\theta}^{(k)} - \boldsymbol{\phi}$ will result in the same prediction

- Minimizing the cross-entropy function has infinite number of solutions since

$$E(\boldsymbol{\theta}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_k^{(n)} \ln \frac{\exp(\boldsymbol{\theta}^{(k)\top}\boldsymbol{x}^{(n)})}{\sum_{j=1}^{K}\exp(\boldsymbol{\theta}^{(j)\top}\boldsymbol{x}^{(n)})} = E(\boldsymbol{\theta} - \boldsymbol{\Phi})$$

  where $\boldsymbol{\Phi} = (\boldsymbol{\phi}, ..., \boldsymbol{\phi})$

Relationship between softmax regression and logistic regression

Let $K = 2$ in softmax

Sigmoid function

- The hypotheses

$$h_1(\boldsymbol{x}) = P(t_1 = 1|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}^{(1)\top}\boldsymbol{x})}{\exp(\boldsymbol{\theta}^{(1)\top}\boldsymbol{x}) + \exp(\boldsymbol{\theta}^{(2)\top}\boldsymbol{x})} = \sigma(\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)})$$

$$h_2(\boldsymbol{x}) = P(t_2 = 1|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}^{(2)\top}\boldsymbol{x})}{\exp(\boldsymbol{\theta}^{(1)\top}\boldsymbol{x}) + \exp(\boldsymbol{\theta}^{(2)\top}\boldsymbol{x})} = 1 - \sigma(\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)})$$

The same as in the two-unit version of the logistic regression if we define a new variable $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}$.

- The error function for each sample

$$E^{(n)}(\boldsymbol{\theta}) = -t_1^{(n)} \ln h_1^{(n)} - t_2^{(n)} \ln h_2^{(n)} = -t_1^{(n)} \ln h_1^{(n)} - (1 - t_1^{(n)}) \ln(1 - h_1^{(n)})$$

The same as in the logistic regression

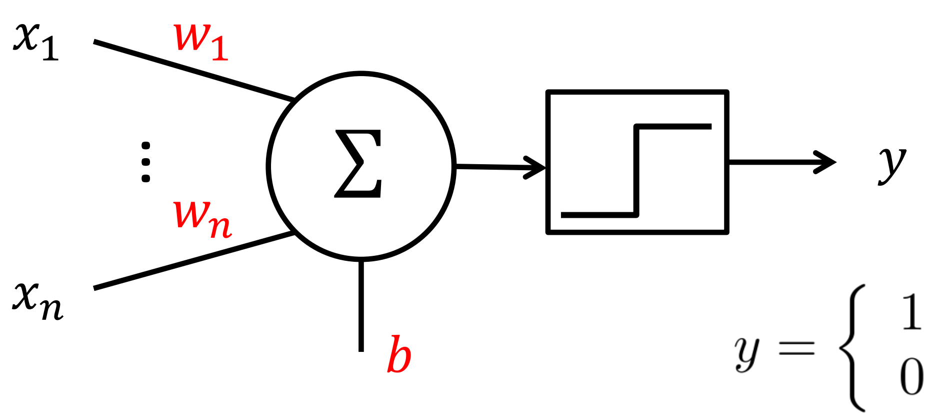
Summary so far

- Nonlinear regression (linear regression as a special case)
  - Output: $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{\theta}^\top \boldsymbol{x})$, where $\boldsymbol{h}$ could be any act function
  - MSE error: $E = \frac{1}{N}\sum_{n=1}^{N} E^{(n)}, E^{(n)} = \frac{1}{2}\left|\left|\boldsymbol{h}\left(\boldsymbol{x}^{(n)} - \boldsymbol{t}^{(n)}\right)\right|\right|_2^2$
  - Gradient: $\nabla_{\boldsymbol{\theta}} E = \frac{1}{N}\sum_{n=1}^{N}\left(\boldsymbol{f}\left(\boldsymbol{x}^{(n)}\right) - \boldsymbol{t}^{(n)}\right) \odot \boldsymbol{f}'\left(\boldsymbol{x}^{(n)}\right)\left(\boldsymbol{x}^{(n)}\right)^\top$
- Softmax regression (logistic regression as a special case)
  - Output: $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{\theta}^\top \boldsymbol{x})$, where $\boldsymbol{h}$ is the softmax function
  - Cross-entropy error: $E = \frac{1}{N}\sum_{n=1}^{N} E^{(n)}, E^{(n)} = -\left(\boldsymbol{t}^{(n)}\right)^\top \ln \boldsymbol{h}^{(n)}$
  - Gradient: $\nabla_{\boldsymbol{\theta}} E = \frac{1}{N}\sum_{n=1}^{N}\left(\boldsymbol{f}\left(\boldsymbol{x}^{(n)}\right) - \boldsymbol{t}^{(n)}\right)\left(\boldsymbol{x}^{(n)}\right)^\top$

Outline

- Regression and classification
- Multi-layer perceptron
- Backpropagation
- Optimization techniques
- Coding considerations

Recap: Perceptron



$$y = \begin{cases} 1 & \boldsymbol{w}^\top \boldsymbol{x} + b > 0 \\ 0 & \text{else} \end{cases}$$

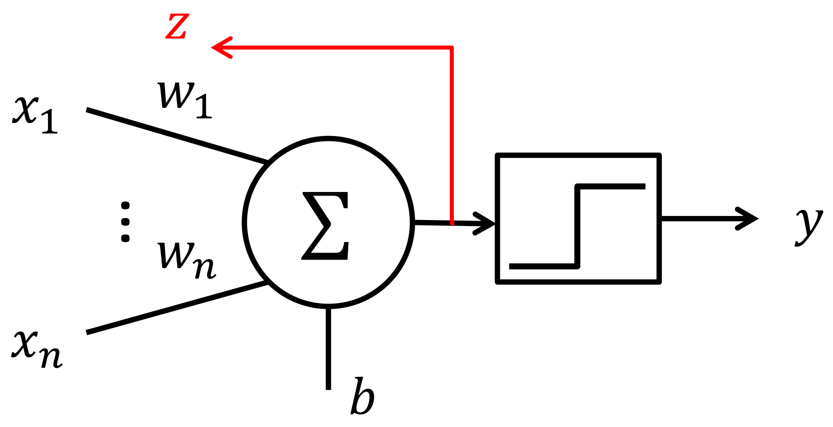
- For each data points $\boldsymbol{x}^{(j)} \in R^m$ and the corresponding labels $t^{(j)}$
  - Calculate the actual output $y^{(j)}$
  - Update the weights: $\boldsymbol{w}^{\text{new}} = \boldsymbol{w}^{old} + \eta\left(t^{(j)} - y^{(j)}\right)\boldsymbol{x}^{(j)}$;

$$b^{\text{new}} = b^{old} + \eta\left(t^{(j)} - y^{(j)}\right)$$

where $\eta > 0$ is the learning rate

- The decision boundary is a hyperplane

Recap: ADALINE



$$y = \begin{cases} 1 & \boldsymbol{w}^\top \boldsymbol{x} + b > 0.5 \\ 0 & \text{else} \end{cases}$$

Or

$$y = \begin{cases} 1 & \boldsymbol{w}^\top \boldsymbol{x} + b > 0 \\ -1 & \text{else} \end{cases}$$

- Same architecture as Perceptron; different training algorithm
  - $z = \boldsymbol{w}^\top \boldsymbol{x} + b$ instead of y is used to adjust the weights and bias
- Minimize MSE $E = \frac{1}{N}\sum_j \left(t^{(j)} - z^{(j)}\right)^2$. The learning algorithm:

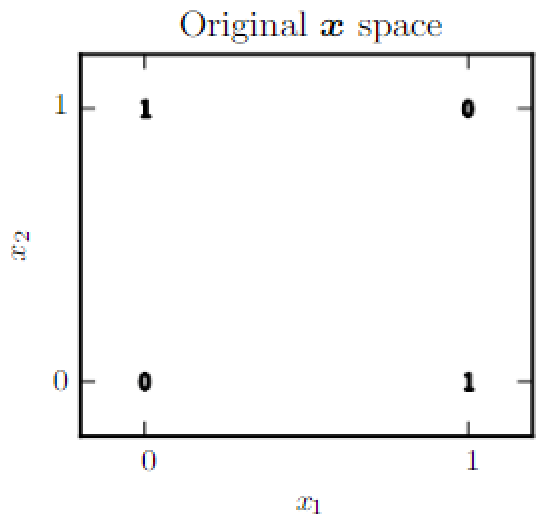
$$\boldsymbol{w}^{\text{new}} = \boldsymbol{w}^{old} + \eta\left(t^{(j)} - z^{(j)}\right)\boldsymbol{x}^{(j)}$$
$$b^{\text{new}} = b^{old} + \eta\left(t^{(j)} - z^{(j)}\right)$$

  where $\eta > 0$ is the learning rate
- The decision boundary is a hyperplane

Solve XOR problem using ADALINE

- Boolean function:

x_1	x_2	t
0	0	0
0	1	1
1	0	1
1	1	0

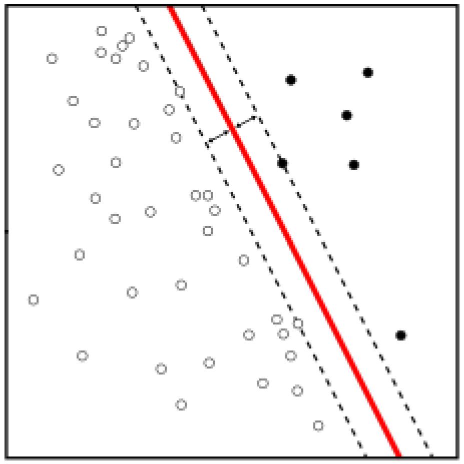


- Error function $E = \frac{1}{4}\sum_{j=1}^{4}\left(t^{(j)} - z^{(j)}\right)^2$ where $z^{(j)} = \boldsymbol{w}^\top \boldsymbol{x}^{(j)} + b$
- Let $\nabla_{\boldsymbol{w}} E = 0, \nabla_b E = 0$, then

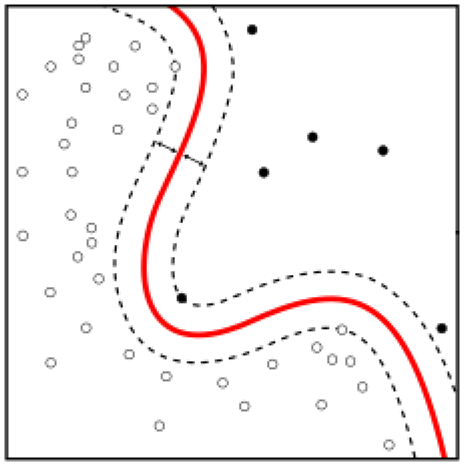
$$2w_1 + 2w_2 + 4b = 2$$
$$2w_1 + w_2 + 2b = 1 \Rightarrow \boldsymbol{w}^* = ?, b = ?$$
$$w_1 + 2w_2 + 2b = 1$$

Limitation

- Both Perceptron and ADALINE can only solve linearly separable classification problems



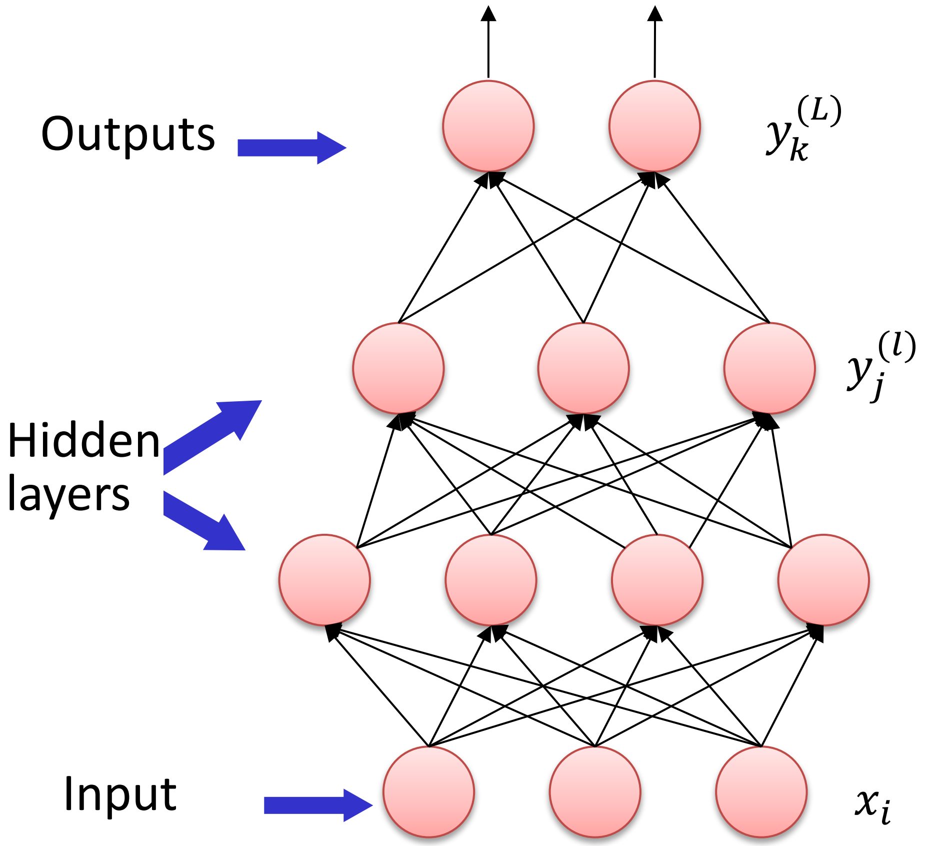
linearly separable



linearly non-separable

- This result discouraged the NN research in 1960s-1970s (1st winter)
- If the problem is linearly non-separable, what should we do?

Multi-layer Perceptron (MLP)



- There are a total of L layers except the input
- Connections:
  - Full connections between layers
  - No feedback connections between layers
  - No lateral connections in the same layer
- Every neuron receives input from previous layer and fire according to an activation function

Activation functions

- Logistic function

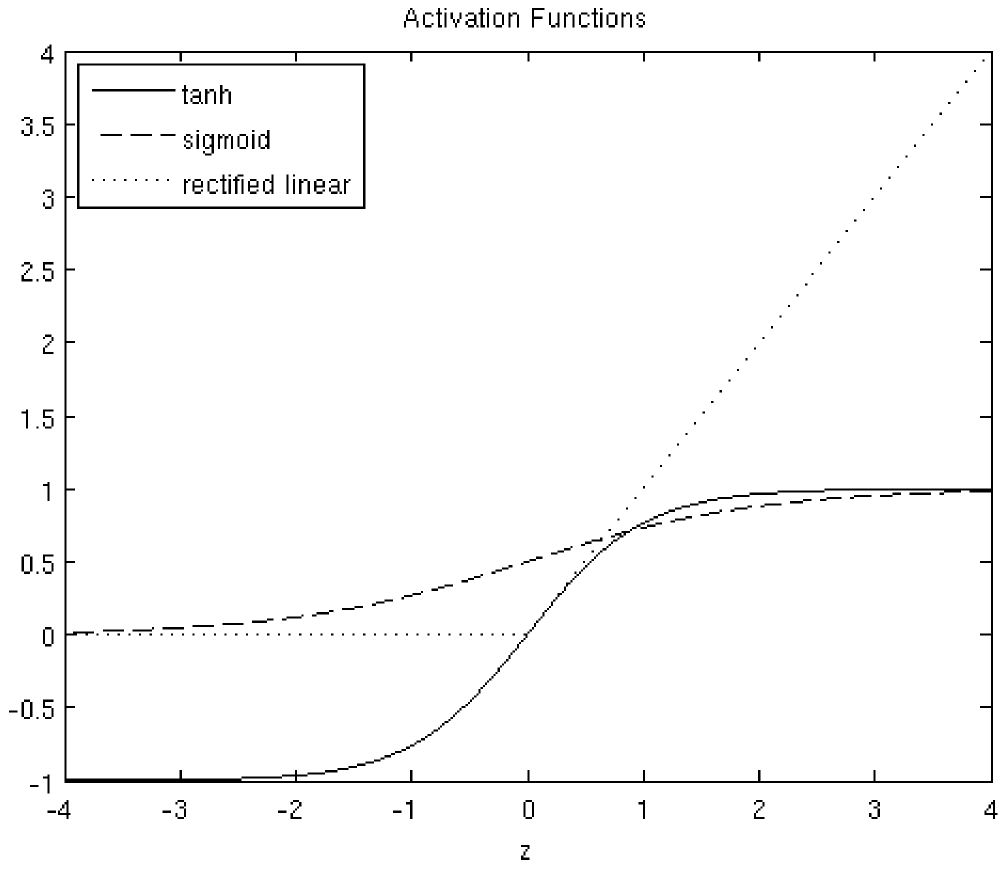
$$f(z) = \frac{1}{1+\exp(-z)}$$

- Hyperbolic tangent function

$$f(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

- Rectified linear activation function (ReLU)

$$f(z) = \max(0, z)$$



Activation functions

- Logistic function

$$f(z) = \frac{1}{1+\exp(-z)} \xrightarrow{\text{gradient}} f'(z) = f(z)(1 - f(z))$$

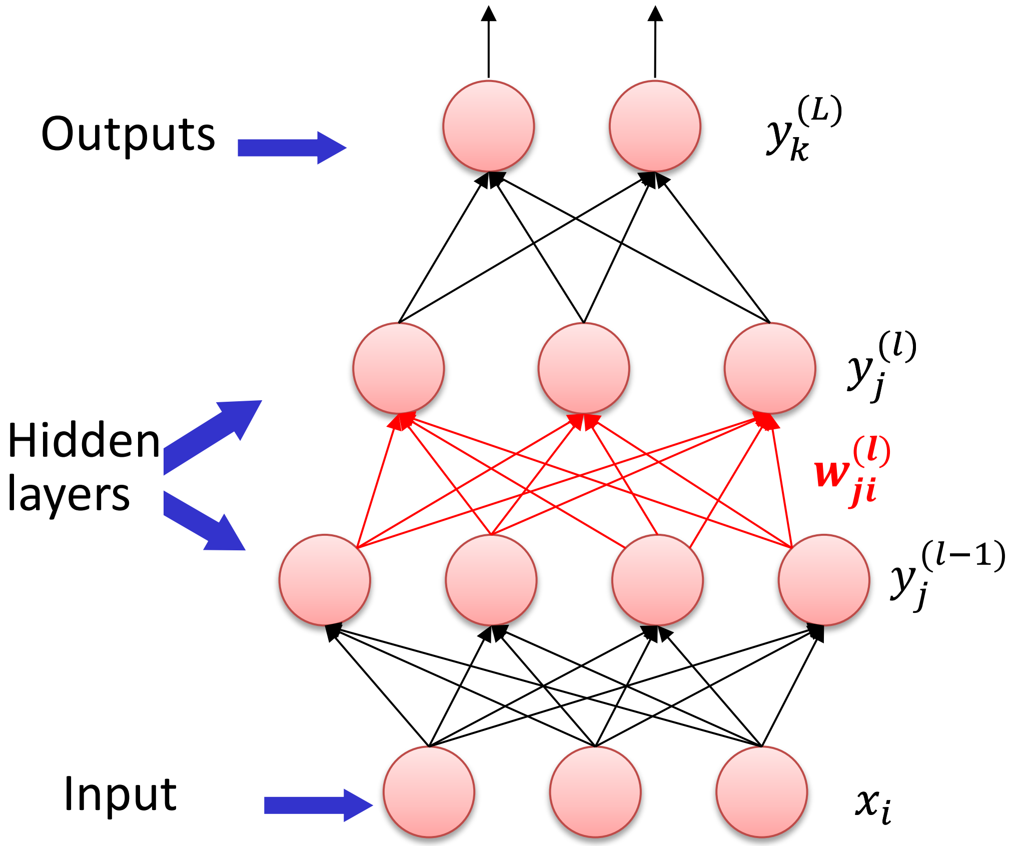
- Hyperbolic tangent function

$$f(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \xrightarrow{\text{gradient}} f'(z) = 1 - f(z)^2$$

- Rectified linear activation function (ReLU)

$$f(z) = \max(0, z) \xrightarrow{\text{gradient}} f'(z) = \begin{cases} 1, & \text{if } z \geq 0, \\ 0, & \text{else} \end{cases}$$

Forward pass

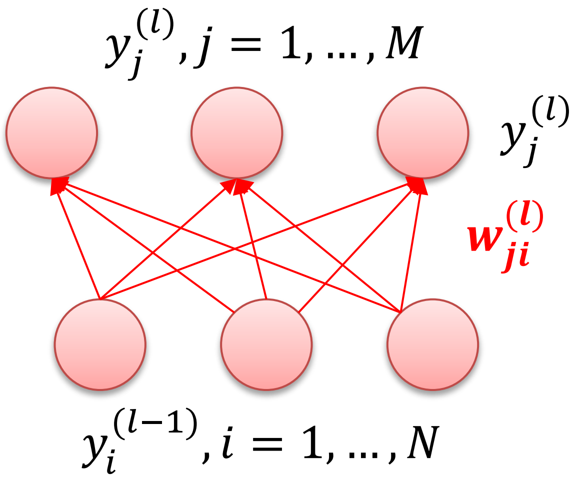


For clarity we don't separate the linear transformation and activation function

- For $l = 1, \dots, L-1$ calculate the input to neuron j in the l-th layer
  $u_j^{(l)} = \sum_i w_{ji}^{(l)} y_i^{(l-1)} + b_j^{(l)}$
  and its output
  $$y_j^{(l)} = f(u_j^{(l)})$$
  where $f(\cdot)$ is activation function
  - Note $y^{(0)} = x$
- $l = L$ corresponds to the classification layer

Forward pass in the vector-matrix form

- If the previous layer has N neurons and the current layer has M neurons, define the weight matrix and bias vector as


$y_j^{(l)}, j = 1, \dots, M$
$y_j^{(l)}$
$\boldsymbol{w}_{ji}^{(l)}$
$y_i^{(l-1)}, i = 1, \dots, N$

$$\boldsymbol{W}^{(l)} = \begin{pmatrix} w_{11} & \cdots & w_{1N} \\ \vdots & \vdots & \vdots \\ w_{M1} & \cdots & w_{MN} \end{pmatrix} \boldsymbol{b}^{(l)} = \begin{pmatrix} b_1 \\ \vdots \\ b_M \end{pmatrix}$$

- Then for $l = 1, \dots, L-1$

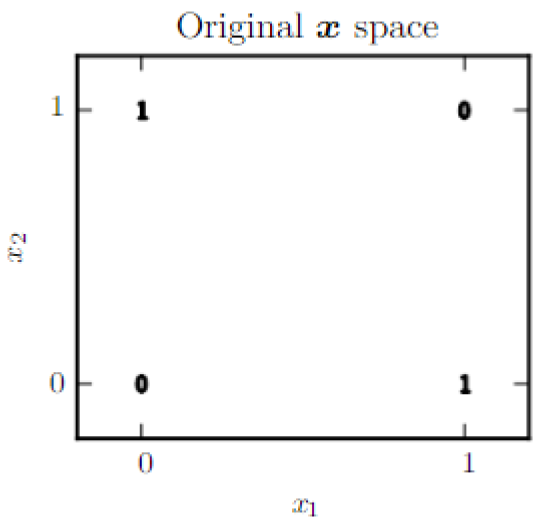
$$\boldsymbol{u}^{(l)} = \boldsymbol{W}^{(l)}\boldsymbol{y}^{(l-1)} + \boldsymbol{b}^{(l)} \text{ and } \boldsymbol{y}^{(l)} = \boldsymbol{f}(\boldsymbol{u}^{(l)})$$

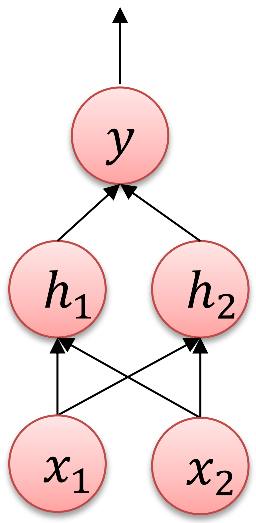
where $\boldsymbol{W}^{(l-1)} \in R^{M\times N}, \boldsymbol{b}^{(l)} \in R^M, u^{(l)}, \boldsymbol{y}^{(l)} \in R^M, \boldsymbol{y}^{(l-1)} \in R^N$

XOR problem revisited

- Boolean function:

x_1	x_2	t
0	0	0
0	1	1
1	0	1
1	1	0





$$y = \mathrm{ReLU}(\boldsymbol{w}^\top \boldsymbol{x} + b)$$

$$\boldsymbol{w} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}, b = 0$$

$$\boldsymbol{h} = \mathrm{ReLU}(\boldsymbol{V}\boldsymbol{x} + \boldsymbol{c})$$

$$\boldsymbol{V} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \boldsymbol{c} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

What are $\boldsymbol{h}^{(n)}$?

What are $y^{(n)}$?

Outline

- Regression and classification
- Multi-layer perceptron
- Backpropagation
- Optimization techniques
- Coding considerations

Error functions for BP

- Error function
$$E = \frac{1}{N}\sum_{n=1}^{N} E^{(n)}$$

where $E^{(n)}$ is the error function for each input sample n

- Squared error, or Euclidean loss

$$E^{(n)} = \frac{1}{2}\sum_{k=1}^{K}(t_k - y_k^{(L)})^2,\ y_k^{(L)} = \frac{1}{1+\exp(-\boldsymbol{w}_k^{(L)\top}\boldsymbol{y}^{(L-1)} - b_k^{(L)})}$$

Is ReLU applicable?

- Cross-entropy error

$$E^{(n)} = -\sum_{k=1}^{K} t_k \ln y_k^{(L)},\ \ y_k^{(L)} = \frac{\exp(\boldsymbol{w}_k^{(L)\top}\boldsymbol{y}^{(L-1)} + b_k^{(L)})}{\sum_{j=1}^{K}\exp(\boldsymbol{w}_j^{(L)\top}\boldsymbol{y}^{(L-1)} + b_j^{(L)})}$$

where $\boldsymbol{t}$ is target of the form $(0, 0, \ldots, 1, 0, 0)^T$

Except $E^{(n)}$, for clarity, we omit the superscript (n) on x, t, u, y etc. for each input sample.

Weight adjustment

- Weight adjustment

$$w_{ji}^{(l)} = w_{ji}^{(l)} - \alpha \frac{\partial E}{\partial w_{ji}^{(l)}} \qquad b_j^{(l)} = b_j^{(l)} - \overset{\text{Learning rate}}{\alpha} \frac{\partial E}{\partial b_j^{(l)}}$$

- Weight decay is often used on ${w_{ji}}^{(l)}$ (not necessary on ${b_j}^{(l)}$) which amounts to adding an additional term on the cost function

$$J = E + \frac{\lambda}{2} \sum_{i,j,l} (w_{ji}^{(l)})^2$$

- Weight adjustment on w is changed to

$$w_{ji}^{(l)} = w_{ji}^{(l)} - \alpha \frac{\partial J}{\partial w_{ji}^{(l)}} = w_{ji}^{(l)} - \alpha \frac{\partial E}{\partial w_{ji}^{(l)}} - \alpha \lambda w_{ji}^{(l)}$$

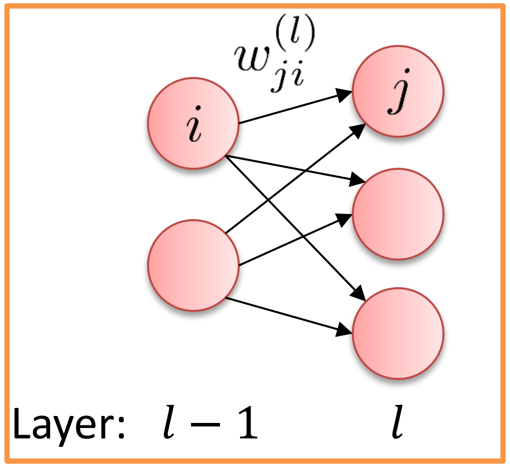
Gradient and local sensitivity

- Define local sensitivity $\delta_i^{(l)} = \frac{\partial E^{(n)}}{\partial u_i^{(l)}}$
- Then for $1 \leq l \leq L$

$$\frac{\partial E^{(n)}}{\partial w_{ji}^{(l)}} = \delta_j^{(l)} \frac{\partial u_j^{(l)}}{\partial w_{ji}^{(l)}} = \delta_j^{(l)} f(u_i^{(l-1)})$$

$$\frac{\partial E^{(n)}}{\partial b_j^{(l)}} = \delta_j^{(l)},$$

since $u_j{}^{(l)} = \sum_i w_{ji}^{(l)} f(u_i^{(l-1)}) + b_j^{(l)}$, where f is the activation function and $f(u_i^{(0)}) = x_i$.



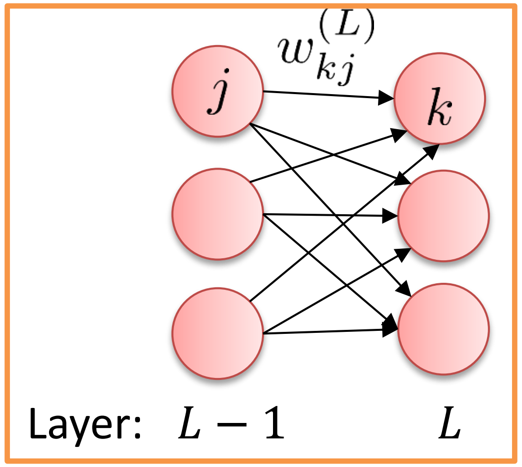
Computing the gradients amounts to computing the local sensitivity in each layer!

Recall: Local sensitivity for MSE layer

- If the squared error is used then the output of the last layer units of MLP are

$$y_k^{(L)} = f(u_k^{(L)}) = f(\boldsymbol{w}_k^{(L)\top} \boxed{\boldsymbol{y}^{(L-1)}} + b_k^{(L)})$$

Output of the units in the (L-1)-th layer



where the activation function f can be

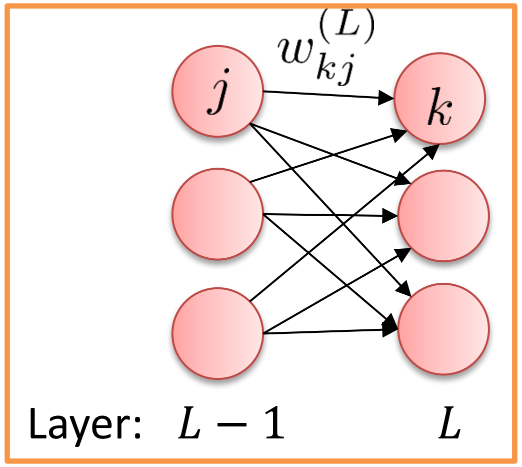
✓ logistic sigmoid ✓ tanh ✓ ReLU

- Recall the error for each sample $E^{(n)} = \frac{1}{2}\sum_{k=1}^{K}(t_k - y_k^{(L)})^2,$
- Local sensitivity

$$\delta_k^{(L)} \triangleq \frac{\partial E^{(n)}}{\partial u_k^{(L)}} = \left(y_k^{(L)} - t_k\right) f'(u_k^{(L)})$$

Recall: local sensitivity for softmax layer

- If the softmax regression is used in the last layer of an MLP, the probabilistic function becomes ($\boldsymbol{\theta}$ is replaced with $\boldsymbol{w}^{(L-1)}$ and $b^{(L-1)}$)



Output of the units in the (L-1)-th layer

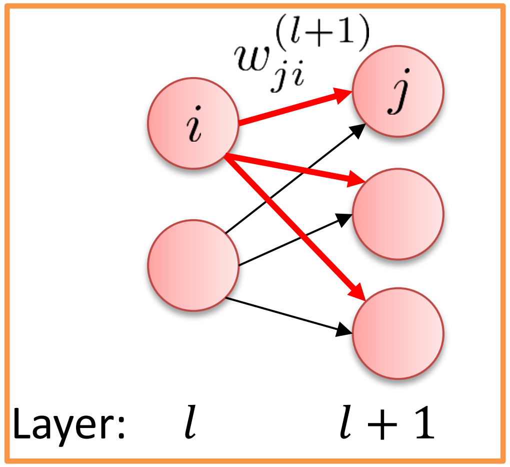
$$y_k^{(L)} \triangleq P(t_k = 1 | \boxed{\boldsymbol{y}^{(L-1)}}) = \frac{\exp(\boldsymbol{w}_k^{(L)\top} \boldsymbol{y}^{(L-1)} + b_k^{(L)})}{\sum_{i=1}^{K} \exp(\boldsymbol{w}_i^{(L)\top} \boldsymbol{y}^{(L-1)} + b_i^{(L)})}$$

- Local sensitivity

$$\delta_k^{(L)} \triangleq \frac{\partial E^{(n)}}{\partial u_k^{(L)}} = y_k^{(L)} - t_k$$

Local sensitivity for other layers

- Define local sensitivity $\delta_i^{(l)} = \frac{\partial E^{(n)}}{\partial u_i^{(l)}}$
- If $1 \leq l < L$, i.e., neuron i is a hidden neuron, it has an effect on all neurons in the next layer, therefore its local sensitivity is



$$\delta_i^{(l)} = \frac{\partial E^{(n)}}{\partial u_i^{(l)}} = \sum_j \frac{\partial E^{(n)}}{\partial u_j^{(l+1)}} \frac{\partial u_j^{(l+1)}}{\partial y_i^{(l)}} \frac{\partial y_i^{(l)}}{\partial u_i^{(l)}} = \sum_j \delta_j^{(l+1)} w_{ji}^{(l+1)} f'(u_i^{(l)})$$

$$u_j^{(l+1)} = \sum_i w_{ji}^{(l+1)} y_i^{(l)} + b_j^{(l+1)} \qquad y_i^{(l)} = f(u_i^{(l)})$$

where f can be any activation function

Therefore we compute $\delta_i^{(l)}$ backward, from $l = L, L-1, \ldots, 1$, and in the sequel $\partial E / \partial W^{(l)}$ and $\partial E / \partial b^{(l)}$ backward

Backpropagation in vector-matrix form

- Local sensitivity $$\boldsymbol{\delta}^{(l)} = \left(\frac{\partial E^{(n)}}{\partial u_1^{(l)}}, \frac{\partial E^{(n)}}{\partial u_2^{(l)}}, \ldots\right)^T$$
- For the output layer L

MSE: $\boldsymbol{\delta}^{(L)} = (\boldsymbol{y}^{(L)} - \boldsymbol{t}) \odot \boldsymbol{f}'(\boldsymbol{u}^{(L)})$ Cross-entropy Err: $\boldsymbol{\delta}^{(L)} = \boldsymbol{y}^{(L)} - \boldsymbol{t}$

where $\odot$ denotes element-wise multiplication

- For the hidden layer $1 \le l < L$

$$\boldsymbol{\delta}^{(l)} = (\boldsymbol{W}^{(l+1)})^\top \boldsymbol{\delta}^{(l+1)} \odot \boldsymbol{f}'(\boldsymbol{u}^{(l)})$$

- Calculate the gradients $1 \le l \le L$

Same dim as $\boldsymbol{W}^{(l)}$

$$\frac{\partial E^{(n)}}{\partial \boldsymbol{W}^{(l)}} = \boldsymbol{\delta}^{(l)}(\boldsymbol{f}(\boldsymbol{u}^{(l-1)}))^\top, \quad \frac{\partial E^{(n)}}{\partial \boldsymbol{b}^{(l)}} = \boldsymbol{\delta}^{(l)}$$

for each sample n

- Update weights

$\longrightarrow$ avg over n

$$\boldsymbol{W}^{(l)} = \boldsymbol{W}^{(l)} - \frac{\alpha}{N}\sum_n \frac{\partial E^{(n)}}{\partial \boldsymbol{W}^{(l)}} - \alpha\lambda \boldsymbol{W}^{(l)}, \quad \boldsymbol{b}^{(l)} = \boldsymbol{b}^{(l)} - \frac{\alpha}{N}\sum_n \frac{\partial E^{(n)}}{\partial \boldsymbol{b}^{(l)}}$$

- The definition of $\boldsymbol{W}$ matrix

$$\boldsymbol{W} = \begin{pmatrix} w_{11} & \cdots & w_{1N} \\ \vdots & \vdots & \vdots \\ w_{M1} & \cdots & w_{MN} \end{pmatrix}$$

where M is the number of neurons in the current layer and N is the number of neurons in the previous layer

- The definition of matrix derivative

$$\frac{\partial E}{\partial \boldsymbol{W}} = \begin{pmatrix} \partial E/\partial w_{11} & \cdots & \partial E/\partial w_{1N} \\ \vdots & \vdots & \vdots \\ \partial E/\partial w_{M1} & \cdots & \partial E/\partial w_{MN} \end{pmatrix}$$

Gradient vanishing

- Note that for the hidden layer $1 \leq l < L$

$$\boldsymbol{\delta}^{(l)} = (\boldsymbol{W}^{(l+1)})^\top \boldsymbol{\delta}^{(l+1)} \odot \boldsymbol{f}'(\boldsymbol{u}^{(l)})$$

  - For logistic function

$$f(z) = \frac{1}{1 + \exp(-z)}$$

    we have $f'(z) = f(z)\big(1 - f(z)\big)$ <1

  - For the tanh function

$$f(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

    we have $f'(z) = 1 - \tanh^2(z) < 1$

- For these two sigmoid functions, $\boldsymbol{\delta}^{(l)}$ is smaller and smaller from L to 1. The gradient approaches zero in lower layers

$$\frac{\partial E^{(n)}}{\partial \boldsymbol{W}^{(l)}} = \boldsymbol{\delta}^{(l)}(\boldsymbol{f}(\boldsymbol{u}^{(l-1)}))^\top, \qquad \frac{\partial E^{(n)}}{\partial \boldsymbol{b}^{(l)}} = \boldsymbol{\delta}^{(l)}$$

ReLU function alleviates this effect

Implementation

- Run forward process
  - Calculate $\boldsymbol{f}\left(\boldsymbol{u}^{(l)}\right)$ and $\boldsymbol{f}'\left(\boldsymbol{u}^{(l)}\right)$ for $l = 1,2,\dots,L$
- Run backward process
  - Calculate $\boldsymbol{\delta}^{(l)}$ and $\partial E/\partial \boldsymbol{W}^{(l)}$,$\partial E/\partial \boldsymbol{b}^{(l)}$ for $l = L, L-1, \dots, 1$
- Update $\boldsymbol{W}^{(l)}$ and $\boldsymbol{b}^{(l)}$ for $l = 1,2,\dots,L$
- Modular programming
  - Implement the layer as a class and provide functions for forward calculation and backward calculation, respectively
  - The forward functions and backward functions differ according to the type of the layer, e.g., input layer, hidden layer, output layer, etc.
  - Then you can design different structures of MLP by specifying the layer modules in a main file

Summary so far

	Feedforward	Backpropagation
MSE output layer	$\boldsymbol{y}^{(L)} = \boldsymbol{f}\left(\boldsymbol{W}^{(L)}\boldsymbol{y}^{(L-1)} + \boldsymbol{b}^{(L)}\right)$, where $\boldsymbol{f}$ is the act function	$\boldsymbol{\delta}^{(L)} = (\boldsymbol{y}^{(L)} - \boldsymbol{t}) \odot \boldsymbol{f}'(\boldsymbol{u}^{(L)})$ $\frac{\partial E^{(n)}}{\partial \boldsymbol{W}^{(l)}} = \boldsymbol{\delta}^{(l)}(\boldsymbol{f}(\boldsymbol{u}^{(l-1)}))^{\top}$
CE output layer	$\boldsymbol{y}^{(l)} = \boldsymbol{f}\left(\boldsymbol{W}^{(l)}\boldsymbol{y}^{(l-1)} + \boldsymbol{b}^{(l)}\right)$, where $\boldsymbol{f}$ is the softmax function	$\boldsymbol{\delta}^{(L)} = \boldsymbol{y}^{(L)} - \boldsymbol{t}$ $\frac{\partial E^{(n)}}{\partial \boldsymbol{W}^{(l)}} = \boldsymbol{\delta}^{(l)}(\boldsymbol{f}(\boldsymbol{u}^{(l-1)}))^{\top}$
Hidden layer	$\boldsymbol{y}^{(l)} = \boldsymbol{f}\left(\boldsymbol{W}^{(l)}\boldsymbol{y}^{(l-1)} + \boldsymbol{b}^{(l)}\right)$, where $\boldsymbol{f}$ is the act function	$\boldsymbol{\delta}^{(l)} = (\boldsymbol{W}^{(l+1)})^{\top}\boldsymbol{\delta}^{(l+1)} \odot \boldsymbol{f}'(\boldsymbol{u}^{(l)})$ $\frac{\partial E^{(n)}}{\partial \boldsymbol{W}^{(l)}} = \boldsymbol{\delta}^{(l)}(\boldsymbol{f}(\boldsymbol{u}^{(l-1)}))^{\top}$

Activation functions: sigmoid, tanh, ReLU

Outline

- Regression and classification
- Multi-layer perceptron
- Backpropagation
- Optimization techniques
- Coding considerations

Weight initialization

W inputting to a neuron is drawn from a distribution:

- Gaussian
  - a Gaussian distribution with zero mean and fixed std, e.g., 0.01
- Xavier
  - a distribution with zero mean and a specific std $1/\sqrt{n_{\mathrm{in}}}$ where n_{in} is the number of neurons feeding into the neuron
  - Gaussian distribution or uniform distribution is often used
- MSRA
  - a Gaussian distribution with zero mean and a specific std $2/\sqrt{n_{\mathrm{in}}}$

Learning rate

- In SGD the learning rate α is typically much smaller than a corresponding learning rate in batch gradient descent because there is much more variance in the update.
- Choosing the proper schedule
  - One standard method is to use a small enough constant learning rate that gives stable convergence in the initial epoch (full pass through the training set) or two of training and then halve the value of the learning rate as convergence slows down.
  - An even better approach is to evaluate a held out set after each epoch and anneal the learning rate when the change in objective between epochs is below a small threshold.
  - Another commonly used schedule is to anneal the learning rate at each iteration t as $\frac{a}{b+t}$ where a and b dictate the initial learning rate and when the annealing begins respectively.

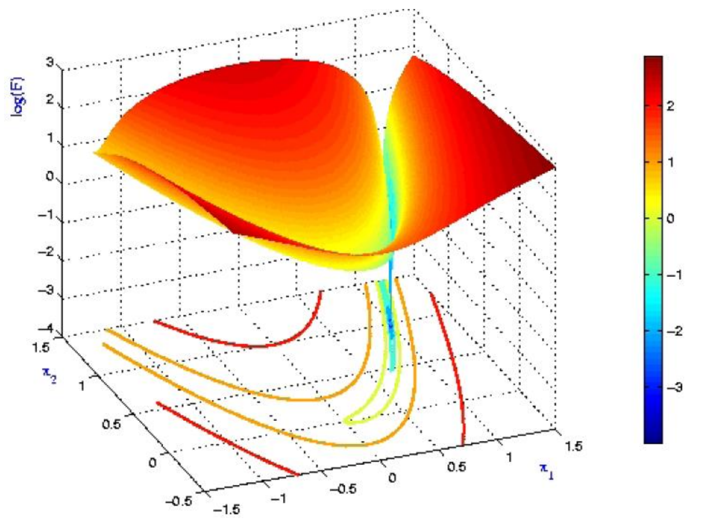
Order of training samples

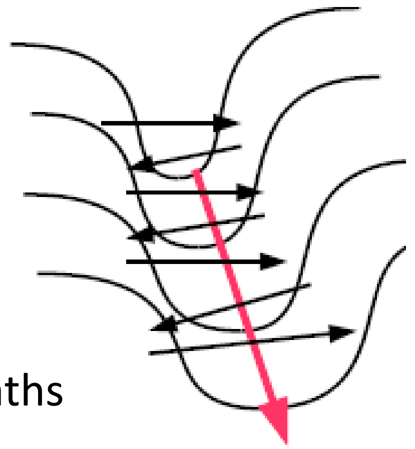
- If the data is given in some meaningful order, this can bias the gradient and lead to poor convergence
- Generally a good method to avoid this is to randomly shuffle the data prior to each epoch of training.

Pathological curvature

- The objective has the form of a long shallow ravine leading to the optimum and steep walls on the sides
    - as seen in the well-known Rosenbrock function
- The objectives of deep architectures have this form near local optima and thus standard SGD tends to oscillate across the narrow ravine

$$f(x,y) = (1-x)^2 + 100(y-x^2)^2$$





Black arrows: gradient descent paths

Momentum

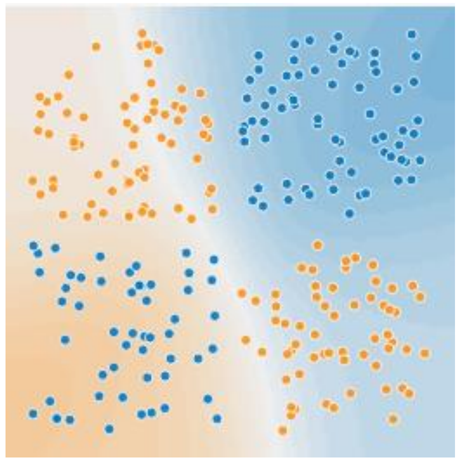
- Momentum is one method for pushing the objective more quickly along the shallow ravine
- The momentum update is given by,

$$\boldsymbol{v} = \gamma \boldsymbol{v} - \alpha \nabla_{\boldsymbol{\theta}} J\left(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, \boldsymbol{t}^{(i)}\right)$$
$$\boldsymbol{\theta} = \boldsymbol{\theta} + \boldsymbol{v}$$

  – $\boldsymbol{v}$ is the current velocity vector
  – $\gamma \in (0,1]$ determines for how many iterations the previous gradients are incorporated into the current update.
  – One strategy: γ is set to 0.5 until the initial learning stabilizes and then is increased to 0.9 or higher

Experiment 1: Classification of 2D points

- http://playground.tensorflow.org



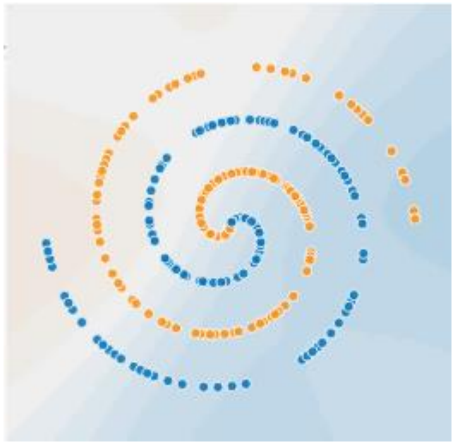
Network setting

- Input: x1, x2
- Hidden layer: 1 layer, 4 neurons
- Use default values for other hyper-parameters

1. Run the training process
2. Change the learning rate to 1 and run
3. Change the learning rate back to 0.03, but use a regularization "L2" with rate 0.01, 0.1, or 1
4. Add a 2nd hidden layer with 2 neurons, and run

Experiment 1: Classification of 2D points

- http://playground.tensorflow.org



Set a suitable setting for solving this problem

- Change the hidden layers, input representation, learning rates, regularization…

Experiment 2: Classification of handwritten digits

https://cs.stanford.edu/people/karpathy/convnetjs/demo/mnist.html

MNIST

- 60,000 training images and 10,000 test images
- 28x28 black and white images

Network setting

```
layer_defs = [];
layer_defs.push({type:'input', out_sx:24,
out_sy:24, out_depth:1});
layer_defs.push({type:'conv', sx:5, filters:8,
stride:1, pad:2, activation:'relu'});
layer_defs.push({type:'pool', sx:2, stride:2});
layer_defs.push({type:'conv', sx:5, filters:16,
stride:1, pad:2, activation:'relu'});
layer_defs.push({type:'pool', sx:3, stride:3});
layer_defs.push({type:'softmax',
num_classes:10});
```

Change the red part to:
layer_defs.push({type:'fc', num_neurons:32, activation:'relu'});

Outline

- Regression and classification
- Multi-layer perceptron
- Backpropagation
- Optimization techniques
- Coding considerations

Motivation

	Feedforward	Backpropagation
Hidden layer	$\boldsymbol{y}^{(l)} = \boldsymbol{f}\left(\boldsymbol{W}^{(l)}\boldsymbol{y}^{(l-1)} + \boldsymbol{b}^{(l)}\right)$, where $\boldsymbol{f}$ is the act function	$\boldsymbol{\delta}^{(l)} = (\boldsymbol{W}^{(l+1)})^\top \boldsymbol{\delta}^{(l+1)} \odot \boldsymbol{f}'(\boldsymbol{u}^{(l)})$ $\frac{\partial E^{(n)}}{\partial \boldsymbol{W}^{(l)}} = \boldsymbol{\delta}^{(l)} (\boldsymbol{f}(\boldsymbol{u}^{(l-1)}))^\top$
MSE output layer	$\boldsymbol{y}^{(L)} = \boldsymbol{f}\left(\boldsymbol{W}^{(L)}\boldsymbol{y}^{(L-1)} + \boldsymbol{b}^{(L)}\right)$, where $\boldsymbol{f}$ is the act function	$\boldsymbol{\delta}^{(L)} = (\boldsymbol{y}^{(L)} - \boldsymbol{t}) \odot \boldsymbol{f}'(\boldsymbol{u}^{(L)})$ $\frac{\partial E^{(n)}}{\partial \boldsymbol{W}^{(l)}} = \boldsymbol{\delta}^{(l)} (\boldsymbol{f}(\boldsymbol{u}^{(l-1)}))^\top$

Everytime when $\boldsymbol{f}$ changes, the feedfoward and backward computation need changes!

More flexible setting

- The input layer or hidden layer

$$y_j{}^{(l)} = f\left(\sum_i w_{ji}^{(l)} y_i^{(l-1)} + b_j^{(l)}\right)$$

  can be decomposed into two layers
  - Fully connected layer: $u_j{}^{(l)} = \sum_i w_{ji}^{(l)} y_i^{(l-1)} + b_j^{(l)}$
  - Activation layer: $y_j^{(l)} = f(u_j{}^{(l)})$
- The squared error layer $E^{(n)} = \frac{1}{2}\left|\left|\boldsymbol{f}\left(\boldsymbol{u}^{(L)}\right) - \boldsymbol{t}\right|\right|_2^2$

  can be decomposed into two layers
  - Activation layer: $y_k^{(L)} = f(u_k{}^{(L)})$, where f can be any function
  - Loss layer: $E^{(n)} = \frac{1}{2}\left|\left|\boldsymbol{y}^{(L)} - \boldsymbol{t}\right|\right|^2$

Question

- Consider the squared error function

$$E^{(n)} = \frac{1}{2}\left|\left|\boldsymbol{f}\left(\boldsymbol{W}^{(L)}\boldsymbol{y}^{(L-1)} + \boldsymbol{b}^{(L)}\right) - \boldsymbol{t}\right|\right|_2^2$$

How many layers can be designed?

More flexible setting

- The cross-entropy error layer $E^{(n)} = -\sum_{k=1}^{K} t_k \ln f\left(u_k^{(L)}\right)$ can be decomposed into two layers
  - Softmax layer: $y_k^{(L)} = f(u_k{}^{(L)})$, where f is the softmax function
  - Loss layer: $E^{(n)} = -\sum_{k=1}^{K} t_k \ln y_k^{(L)}$
  - But this is unnecessary! Why?
- Consider this error

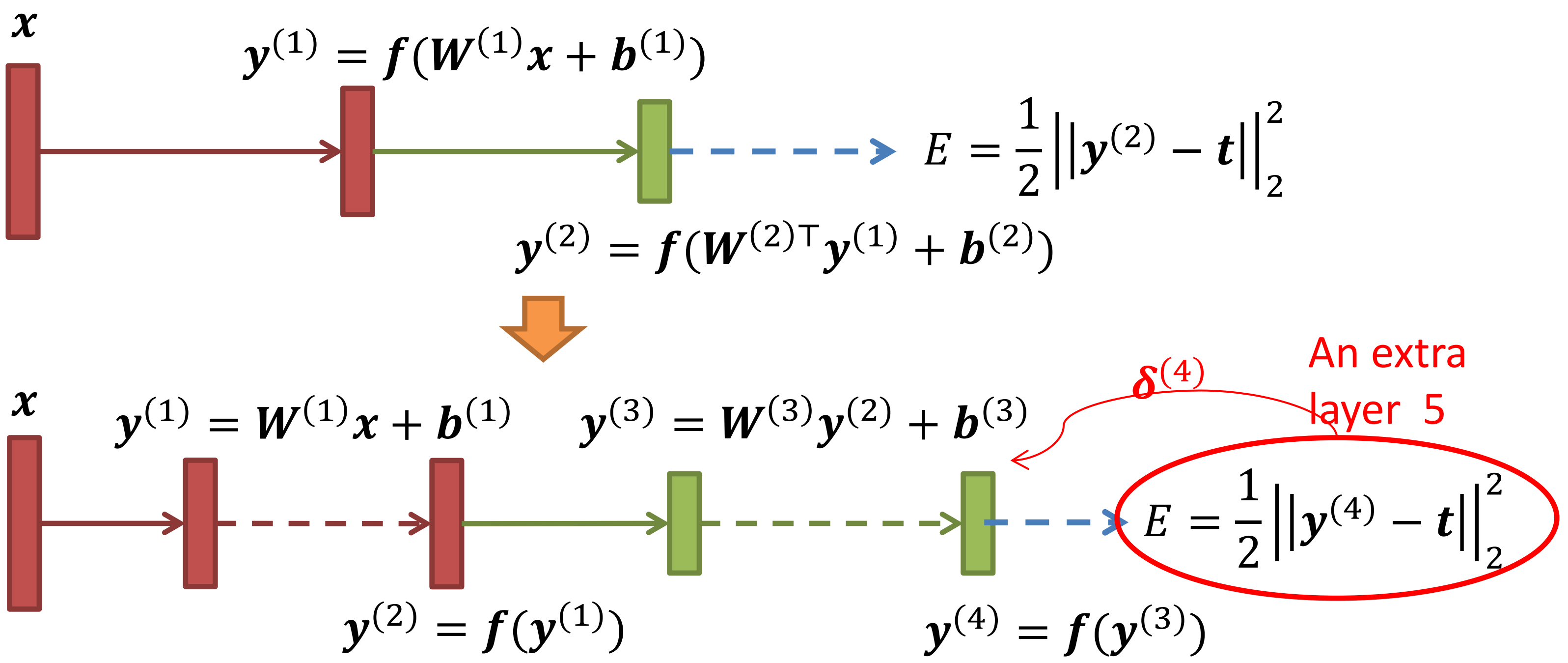
$$E^{(n)} = -\sum_{k=1}^{K} t_k \ln f\left(\sum_i w_{ki}^{(l)} y_i^{(l-1)} + b_k^{(l)}\right)$$

How many layers can be designed?

Example 1

- An MLP with one hidden layer using the MSE loss

Solid arrow: w/ param.
Dashed arrow: w/o param.



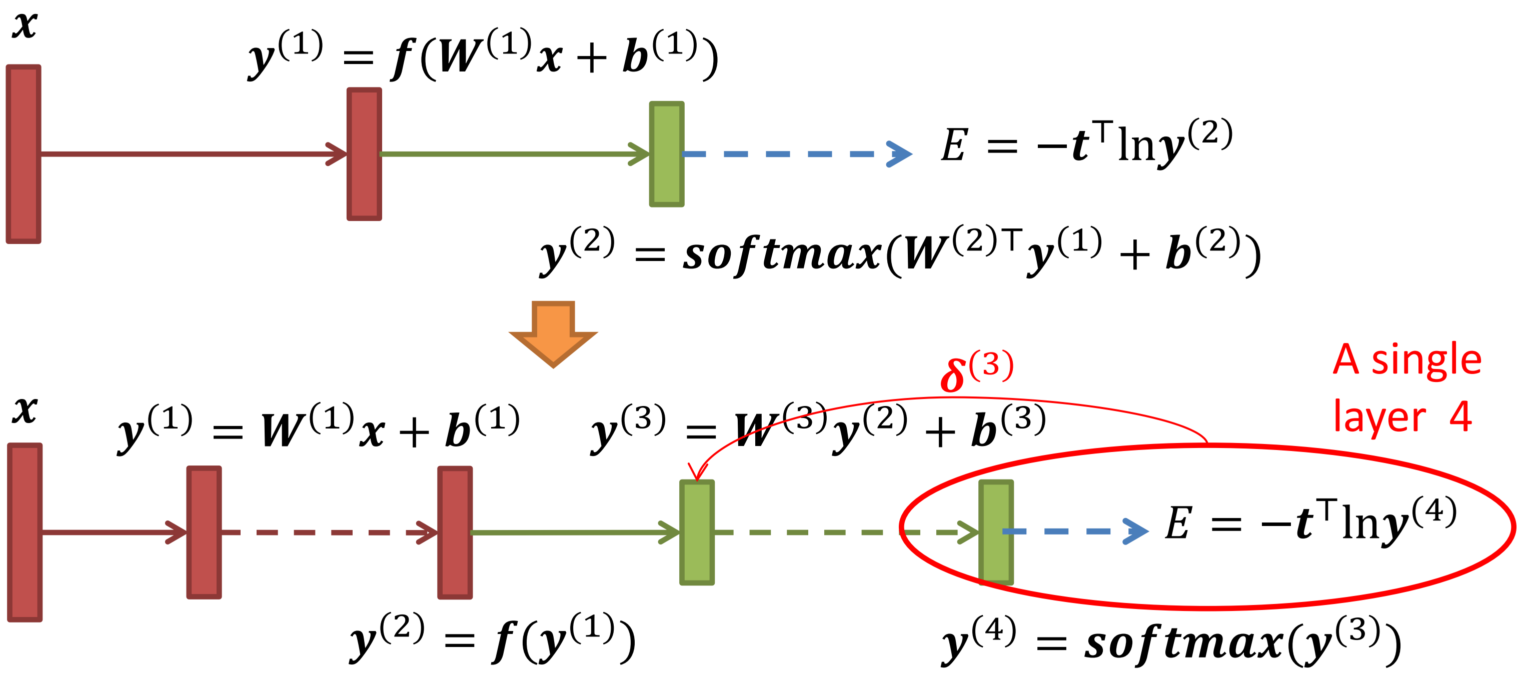
$\boldsymbol{x}$
$\boldsymbol{y}^{(1)} = \boldsymbol{f}(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})$
$E = \frac{1}{2}\left|\left|\boldsymbol{y}^{(2)} - \boldsymbol{t}\right|\right|_2^2$
$\boldsymbol{y}^{(2)} = \boldsymbol{f}(\boldsymbol{W}^{(2)\top}\boldsymbol{y}^{(1)} + \boldsymbol{b}^{(2)})$
$\boldsymbol{x}$
$\boldsymbol{y}^{(1)} = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$
$\boldsymbol{y}^{(3)} = \boldsymbol{W}^{(3)}\boldsymbol{y}^{(2)} + \boldsymbol{b}^{(3)}$
$\boldsymbol{\delta}^{(4)}$
An extra layer 5
$E = \frac{1}{2}\left|\left|\boldsymbol{y}^{(4)} - \boldsymbol{t}\right|\right|_2^2$
$\boldsymbol{y}^{(2)} = \boldsymbol{f}(\boldsymbol{y}^{(1)})$
$\boldsymbol{y}^{(4)} = \boldsymbol{f}(\boldsymbol{y}^{(3)})$

Layer: **fc1** **act1** **fc2** **act2** **loss**

$\boldsymbol{u}^{(l)}$ and $\boldsymbol{y}^{(l)}$ are the same in every layer l

Example 2

- An MLP with one hidden layer using the cross-entropy error loss



Solid arrow: w/ param.
Dashed arrow: w/o param.
$\boldsymbol{x}$
$\boldsymbol{y}^{(1)} = \boldsymbol{f}(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})$
$E = -\boldsymbol{t}^\top \ln \boldsymbol{y}^{(2)}$
$\boldsymbol{y}^{(2)} = \boldsymbol{softmax}(\boldsymbol{W}^{(2)\top}\boldsymbol{y}^{(1)} + \boldsymbol{b}^{(2)})$
$\boldsymbol{\delta}^{(3)}$
A single layer 4
$\boldsymbol{x}$
$\boldsymbol{y}^{(1)} = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$
$\boldsymbol{y}^{(3)} = \boldsymbol{W}^{(3)}\boldsymbol{y}^{(2)} + \boldsymbol{b}^{(3)}$
$E = -\boldsymbol{t}^\top \ln \boldsymbol{y}^{(4)}$
$\boldsymbol{y}^{(2)} = \boldsymbol{f}(\boldsymbol{y}^{(1)})$
$\boldsymbol{y}^{(4)} = \boldsymbol{softmax}(\boldsymbol{y}^{(3)})$

Layer: **fc1** **act1** **fc2** **act2** **loss**

$\boldsymbol{u}^{(l)}$ and $\boldsymbol{y}^{(l)}$ are the same in every layer l

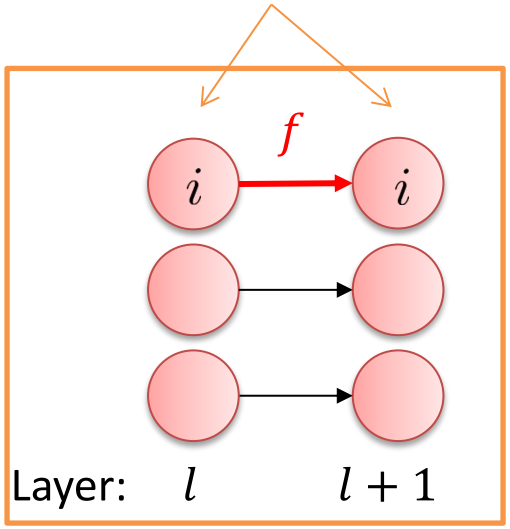
Exercise

- Derive the local sensitivity $\boldsymbol{\delta}$ and gradient $\partial E/\partial \boldsymbol{W}$ and $\partial E/\partial \boldsymbol{b}$ where applicable for
  - Euclidean loss layer: $E^{(n)} = \frac{1}{2}\left|\left|\boldsymbol{y}^{(L)} - \boldsymbol{t}\right|\right|^2$
    - Note that here we calculate $\boldsymbol{\delta}^{(L)} = \partial E^{(n)}/\partial \boldsymbol{y}^{(L)}$
  - Softmax-cross-entropy error layer $E^{(n)} = -\sum_{k=1}^{K} t_k \ln f\left(y_k^{(L)}\right)$
    - Note that here we calculate $\boldsymbol{\delta}^{(L-1)} = \partial E^{(n)}/\partial \boldsymbol{y}^{(L-1)}$
  - Fully connected layer: ${y_j}^{(l)} = \sum_i w_{ji}^{(l)} y_i^{(l-1)} + b_j^{(l)}$
  - Sigmoid layer: $y_j^{(l)} = f({y_j}^{(l-1)})$, where f is a sigmoid function
  - ReLU layer: $y_j^{(l)} = f({y_j}^{(l-1)})$, where f is a ReLU function

These layers are shown in the previous slides

Hint

No act fun anymore



- Suppose the $(l+1)$-th layer is a sigmoid activation layer:

$$y_i^{(l+1)} = f\left(y_i^{(l)}\right)$$

where f is the sigmoid function

- Neuron i in the l-th layer only affects neuron i in the $(l+1)$-th layer, therefore

$$\delta_i^{(l)} = \frac{\partial E^{(n)}}{\partial u_i^{(l)}} = \frac{\partial E^{(n)}}{\partial y_i^{(l)}} = \frac{\partial E^{(n)}}{\partial y_i^{(l+1)}} \frac{\partial y_i^{(l+1)}}{\partial y_i^{(l)}} = \delta_i^{(l+1)} f'(y_i^{(l)})$$

Note that this layer doesn't have w and b

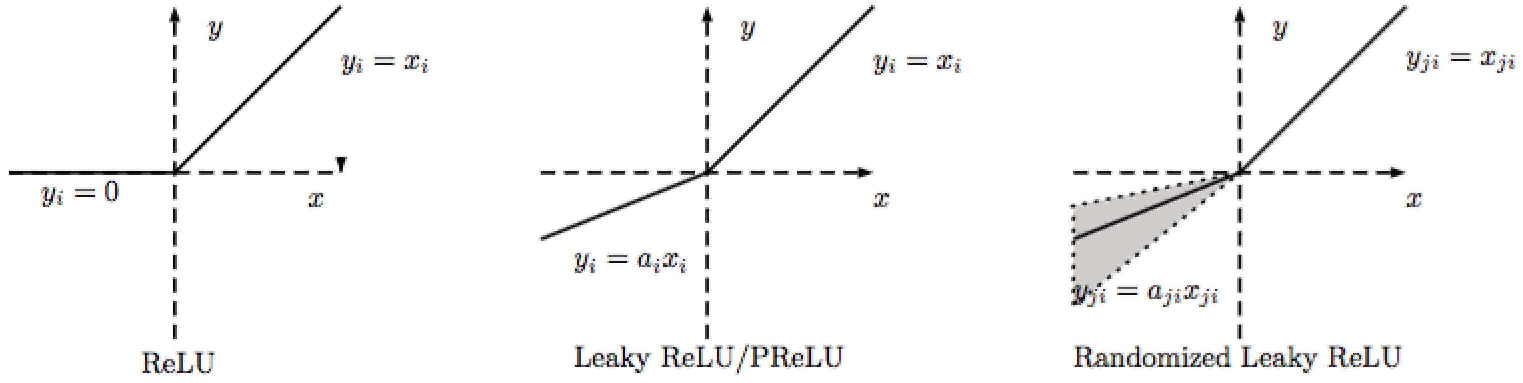
- Similarly, you can derive the results for other layers

Summary

- Regression and classification
  - Linear and nonlinear regression
  - Logistic regression
  - Softmax regression
- MLP
  - All to all connections
  - The last layer is for regression or classification
  - Activation functions: sigmoid, tanh, ReLU
- Backpropagation
  - Local sensitivity
  - Gradient vanishing
- Optimization techniques
  - Weight initialization, learning rate, order of training samples, momentum
- Coding considerations
  - Euclidean loss layer
  - Softmax-cross-entropy error layer
  - Fully connected layer
  - Activation layer

Further reading

- Variants of ReLU activation function



  - Xu, Wang, Chen, Li, Empirical Evaluation of Rectified Activations in Convolution Network, arXiv:1505.00853v2

- Other types of activation functions
  - Softplus: $f(x) = \log(e^x + 1)$
  - Softsign: $f(x) = \frac{x}{|x|+1}$