Assignment 1 Notes

mn

Contents

1 Definitions

- row vectors:
- 1 layer network:
  - Input dimension: $\mathrm{D_x}$ ($\boldsymbol{x} \in [N, D_x]$)
  - Hidden units: H ($\mathbf{W}_1 \in [D_x, H],\ \boldsymbol{b}_1 \in [H]$)
  - Ouput dimension: $\mathrm{D_y}$ ($\mathbf{W}_2 \in [H, D_y],\ b_2 \in [D_y]$)

2 Assigment 1

Org latex export is not that great yet, and one still has to work on the alignment.

2.1 1.a) Softmax function

Softmax function: $\mathrm{softmax}(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$

$$\mathrm{softmax}(x+c)_i = \frac{e^{x_i+c}}{\sum_j e^{x_j+c}} = \frac{e^c e^{x_i}}{e^c \sum_j e^{x_j}} = \frac{e^{x_i}}{\sum_j e^{x_j}} = \mathrm{softmax}(x)_i$$

2.2 2.a) Derivative of the sigmoid function

Sigmoid function: $\sigma(x) = \frac{1}{1+e^{-x}}$

$$\frac{\partial}{\partial x}\sigma(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}\,\frac{e^{-x}}{1+e^{-x}} = \frac{1}{1+e^{-x}}\,\frac{(1+e^{-x})-1}{1+e^{-x}} = \sigma(x)(1-\sigma(x))$$

2.3 2.b) Derivative of the Cross Entropy Loss

Cross Entropy Loss: $CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum_i y_i \log(\hat{y}_i)$

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\theta)$$

$$\frac{\partial}{\partial \theta_j} CE(\boldsymbol{y}, \hat{\boldsymbol{y}})_j = -\frac{\partial}{\partial \theta_j}\sum_i y_i \log \hat{y}_i = -\frac{\partial}{\partial \theta_j}\sum_i y_i \log(e^{\theta_i}/\sum_l e^{\theta_l}) = -\frac{\partial}{\partial \theta_j}\sum_i y_i(\theta_i - \log\sum_l e^{\theta_l})$$

$\boldsymbol{y}$ is a *one-hot* label vector with a non-zero value (=1) at index k.

Given that

$$\frac{\partial}{\partial \theta_j} \log \sum_l e^{\theta_l} = \frac{e_j^{\theta}}{\sum_l e^{\theta_l}} = \hat{y}_j$$

if $j \neq k$:

$$\frac{\partial}{\partial \theta_j} CE(\boldsymbol{y}, \hat{\boldsymbol{y}})_j = \frac{\partial}{\partial \theta_j} \log \sum_l e^{\theta_l} = \hat{y}_j$$

if $j = k$:

$$\frac{\partial}{\partial \theta_j} CE(\boldsymbol{y}, \hat{\boldsymbol{y}})_j = \hat{y}_j - 1$$

Combining (vectorizing) these results:

$$\frac{\partial}{\partial \boldsymbol{\theta}} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \hat{\boldsymbol{y}} - \boldsymbol{y}$$

2.4 2.c) Gradient of the one-layer network

$$\frac{\partial J(x)}{\partial x} = \frac{\partial}{\partial x} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\frac{\partial}{\partial x} \sum_i y_i \log(\text{softmax}(\sigma(\boldsymbol{x}\mathbf{W}_1 + \boldsymbol{b}_1)\mathbf{W}_2 + \boldsymbol{b}_2))$$

With $\hat{\boldsymbol{y}} = \text{softmax}(\theta)$, $\theta = \sigma(\boldsymbol{h})\mathbf{W}_2 + \boldsymbol{b}_2$ and $\boldsymbol{h} = \boldsymbol{x}\mathbf{W}_1 + \boldsymbol{b}_1$

$$\frac{\partial J(x)}{\partial x} = \sum_j \frac{\partial J}{\partial \theta_j} \frac{\partial \theta_j}{\partial x} = \sum_j \sigma(\boldsymbol{h})(1 - \sigma(\boldsymbol{h}))\mathbf{W}_1 \mathbf{W}_{2,j} \times \begin{cases} \hat{y}_j & \text{if } j \neq k \\ \hat{y}_j - 1 & \text{if } j = k \end{cases}$$

2.4.1 Alternative (more layer-systematic) derivation

- Forward pass:

$$J(x) = CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum_i y_i \log(\text{softmax}(\sigma(\boldsymbol{x}\mathbf{W}_1 + \boldsymbol{b}_1)\mathbf{W}_2 + \boldsymbol{b}_2))$$

  - Input: $\boldsymbol{x}$
  - L1 affine transformation: $\boldsymbol{z}_1 = \boldsymbol{x}\mathbf{W}_1 + \boldsymbol{b}_1$
  - L1 activation: $\boldsymbol{h} = \sigma(\boldsymbol{z}_1)$
  - L2 affine transformation: $\boldsymbol{z}_2 = \boldsymbol{h}\mathbf{W}_2 + \boldsymbol{b}_2$
  - Scores computation: $\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{z}_2)$
  - Cross-Entropy Loss: $CE = \sum_i \boldsymbol{y} \log \hat{\boldsymbol{y}}$

- Backward pass (gradient derivation):

$$\frac{\partial J}{\partial x} = \frac{\partial J}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_2} \frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}_1} \frac{\partial \boldsymbol{z}_1}{\partial x} = \frac{\partial J}{\partial x} = (\hat{\boldsymbol{y}} - \boldsymbol{y})\mathbf{W}_2^T \sigma(\boldsymbol{z}_1)(1 - \sigma(z_1))\mathbf{W}_1^T$$

  - $dz_2 = \frac{\partial J}{\partial \boldsymbol{z}_2} = \hat{\boldsymbol{y}} - \boldsymbol{y}$
  - $dh_1 = \frac{\partial J}{\partial \boldsymbol{h}} = dz_2 \frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{h}} = dz_2 \mathbf{W}_2^T$
  - $dz_1 = \frac{\partial J}{\partial \boldsymbol{z}_1} = dh_1 \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}_1} = dh_1 \circ \frac{\partial \sigma(\boldsymbol{z}_1)}{\partial \boldsymbol{z}_1}$
  - $dx = \frac{\partial J}{\partial \boldsymbol{x}} = dz_1 \frac{\partial \boldsymbol{z}_1}{\partial \boldsymbol{x}} = dz_1 \mathbf{W}_1^T$

- Other gradients:
  - dW2 $= \frac{\partial J}{\partial W_2} = \frac{\partial J}{\partial z_2} \frac{\partial z_2}{\partial \mathbf{W}_2} = \boldsymbol{h}^T dz_2$
  - db2 $= \frac{\partial J}{\partial b_2} = \frac{\partial J}{\partial z_2} \frac{\partial z_2}{\partial \boldsymbol{b}_2} = dz_2$ (summed over the first dimension)
  - dW1 $= \frac{\partial J}{\partial W_1} = \frac{\partial J}{\partial z_1} \frac{\partial z_1}{\partial \mathbf{W}_1} = \boldsymbol{x}^T dz_1$
  - db1 $= \frac{\partial J}{\partial b_1} = \frac{\partial J}{\partial z_1} \frac{\partial z_1}{\partial \boldsymbol{b}_1} = dz_1$ (summed over the first dimension)

2.5 2.d) Number of parameters

Number of parameters: $(D_x + 1)H + (H + 1)D_y$