Unit 14 - Week 11 - L1 regularization

Course outline

Assignment 11

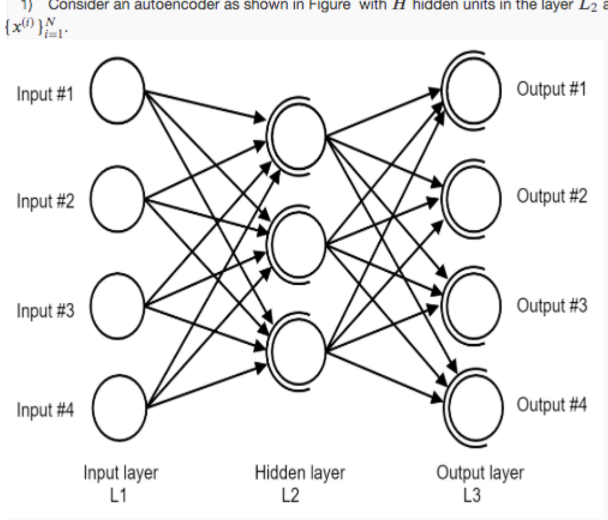
The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2019-10-17, 23:59 IST.

Instructions:

1. **Attempt all questions.**
2. **Submission deadline**: 16th October 2019 23:59 IST
3. **Solutions to be posted**: 17th October 2019
4. Older browsers might show **unnecessary vertical bars** at the end of math equations.

1) Consider an autoencoder as shown in Figure with H hidden units in the layer L_2 and input to the autoencoder is a set of P-dimensional unlabeled data $\{x^{(i)}\}_{i=1}^{N}$. ***1 point***



We use the following notation for this autoencoder:

- W^e denotes the $P \times H$ weight matrix between L_1 and L_2.
- W^d denotes the $H \times P$ weight matrix between L_2 and L_3.
- σ denotes the activation function for L_2 and L_3.
- $s_j^{(i)} = \sum_{k=1}^{P} W_{kj}^e x_k^{(i)}$.
- $z_j^{(i)} = \sigma\left(\sum_{k=1}^{P} W_{kj}^e x_k^{(i)}\right)$.
- $t_j^{(i)} = \sum_{k=1}^{H} W_{kj}^d z_k^{(i)}$.
- $\hat{x}_j^{(i)} = \sigma\left(\sum_{k=1}^{H} W_{kj}^d z_k^{(i)}\right)$
- $J(W^e, W^d)^{(i)} = \|x^{(i)} - \hat{x}^{(i)}\|^2 = \sum_{j=1}^{P}\left(x_j^{(i)} - \hat{x}_j^{(i)}\right)^2$ is the reconstruction error for example $x^{(i)}$.
- $J(W^e, W^d) = \sum_{i=1}^{N} J(W^e, W^d)^{(i)}$ is the total reconstruction error.
- We add 1 to the input layer and hidden layer so that no bias term has to be considered.

The above notations and details are to be used for questions 1-4.

Compute $\frac{\partial J^{(i)}}{\partial W_{kl}^d}$

- $\sum_{j=1}^{P}\left(\left(\hat{x}_j^{(i)} + x_j^{(i)}\right)\frac{\partial \hat{x}_j^{(i)}}{\partial W_{kl}^d}\right)$
- $\sum_{j=1}^{P}\left(2\left(\hat{x}_j^{(i)} + x_j^{(i)}\right)\frac{\partial \hat{x}_j^{(i)}}{\partial W_{kl}^d}\right)$
- $\sum_{j=1}^{P}\left(\left(\hat{x}_j^{(i)} - x_j^{(i)}\right)\frac{\partial \hat{x}_j^{(i)}}{\partial W_{kl}^d}\right)$
- $\sum_{j=1}^{P}\left(2\left(\hat{x}_j^{(i)} - x_j^{(i)}\right)\frac{\partial \hat{x}_j^{(i)}}{\partial W_{kl}^d}\right)$

No, the answer is incorrect.
Score: 0
Accepted Answers:
$\sum_{j=1}^{P}\left(2\left(\hat{x}_j^{(i)} - x_j^{(i)}\right)\frac{\partial \hat{x}_j^{(i)}}{\partial W_{kl}^d}\right)$

2) Compute $\frac{\partial J^{(i)}}{\partial W_{kl}^e}$ ***2 points***

- $\frac{\partial J^{(i)}}{\partial s_j^{(i)}} W_{kj}^d$
- $\frac{\partial J^{(i)}}{\partial s_j^{(i)}} W_{kj}^e$
- $\frac{\partial J^{(i)}}{\partial s_j^{(i)}} z_k^{(i)}$
- $\frac{\partial J^{(i)}}{\partial s_j^{(i)}} x_k^{(i)}$

No, the answer is incorrect.
Score: 0
Accepted Answers:
$\frac{\partial J^{(i)}}{\partial s_j^{(i)}} x_k^{(i)}$

3) Compute $\frac{\partial J^{(i)}}{\partial s_j^{(i)}}$ ***2 points***

- $\sum_{j=1}^{H} \frac{\partial J^{(i)}}{\partial t_k^{(i)}} W_{jk}^e \sigma\left(s_j^{(i)}\right)$
- $\sum_{j=1}^{H} \frac{\partial J^{(i)}}{\partial t_k^{(i)}} W_{jk}^e \sigma'\left(s_j^{(i)}\right)$
- $\sum_{j=1}^{H} \frac{\partial J^{(i)}}{\partial t_k^{(i)}} W_{jk}^d \sigma\left(s_j^{(i)}\right)$
- $\sum_{j=1}^{H} \frac{\partial J^{(i)}}{\partial t_k^{(i)}} W_{jk}^d \sigma'\left(s_j^{(i)}\right)$

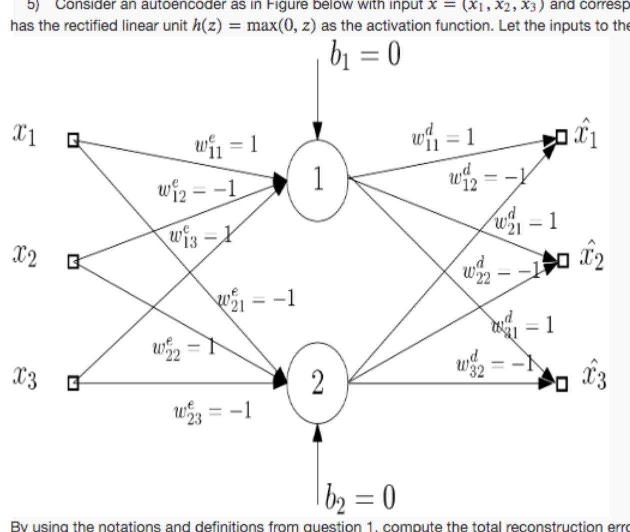
No, the answer is incorrect.
Score: 0
Accepted Answers:
$\sum_{j=1}^{H} \frac{\partial J^{(i)}}{\partial t_k^{(i)}} W_{jk}^d \sigma'\left(s_j^{(i)}\right)$

4) Compute $\frac{\partial J^{(i)}}{\partial s_j^{(i)}}$. ***2 points***

- $\sum_{j=1}^{H} \frac{\partial J^{(i)}}{\partial t_k^{(i)}} W_{jk}^e \sigma\left(s_j^{(i)}\right)$
- $\sum_{j=1}^{H} \frac{\partial J^{(i)}}{\partial t_k^{(i)}} W_{jk}^e \sigma'\left(s_j^{(i)}\right)$
- $\sum_{j=1}^{H} \frac{\partial J^{(i)}}{\partial t_k^{(i)}} W_{jk}^d \sigma\left(s_j^{(i)}\right)$
- $\sum_{j=1}^{H} \frac{\partial J^{(i)}}{\partial t_k^{(i)}} W_{jk}^d \sigma'\left(s_j^{(i)}\right)$

No, the answer is incorrect.
Score: 0
Accepted Answers:
$\sum_{j=1}^{H} \frac{\partial J^{(i)}}{\partial t_k^{(i)}} W_{jk}^d \sigma'\left(s_j^{(i)}\right)$

5) Consider an autoencoder as in Figure below with input $x = (x_1, x_2, x_3)$ and corresponding desired output as $\hat{x} = (x_1, x_2, x_3)$. Each hidden neuron has the rectified linear unit $h(z) = \max(0, z)$ as the activation function. Let the inputs to the network be $.x = \{(1,0,0), (1,1,1)\}$. ***2 points***



By using the notations and definitions from question 1, compute the total reconstruction error after the first forward pass.

- $J = 2$
- $J = 4$
- $J = 5$
- $J = 3$

No, the answer is incorrect.
Score: 0
Accepted Answers:
$J = 5$

6) In continuation with question 5, compute $\frac{\partial J^{(1)}}{\partial w_{11}^d} =$

No, the answer is incorrect.
Score: 0
Accepted Answers:
(Type: Numeric) 0

1 point

7) Let us consider the quadratic set up under L_1 regularity constraints. Let ***1 point***
$\hat{\underline{w}} = \arg\min_{\underline{w}} \|\mathbf{X}\underline{w} - \underline{y}\|^2$
such that $\|\underline{w}\|_1 \leq t$. Assume $\underline{w} \in \mathbb{R}^2$. Expressing the regularity constraints in the form $A\underline{w} \leq \underline{b}$ we get

- $\begin{bmatrix} -1 & 1 \\ 1 & -1 \\ -1 & -1 \\ 1 & -1 \end{bmatrix}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \leq \begin{pmatrix} -t \\ -t \\ -t \\ -t \end{pmatrix}$
- $\begin{bmatrix} 1 & 1 \\ -1 & -1 \\ -1 & 1 \\ 1 & -1 \end{bmatrix}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \leq \begin{pmatrix} -t \\ -t \\ -t \\ t \end{pmatrix}$
- $\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ -1 & 1 \\ 1 & -1 \end{bmatrix}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \leq \begin{pmatrix} -t \\ t \\ -t \\ t \end{pmatrix}$
- $\begin{bmatrix} 1 & -1 \\ 1 & 1 \\ -1 & 1 \\ -1 & -1 \end{bmatrix}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \leq \begin{pmatrix} t \\ t \\ t \\ t \end{pmatrix}$

No, the answer is incorrect.
Score: 0
Accepted Answers:
$\begin{bmatrix} 1 & -1 \\ 1 & 1 \\ -1 & 1 \\ -1 & -1 \end{bmatrix}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \leq \begin{pmatrix} t \\ t \\ t \\ t \end{pmatrix}$

8) In continuation with question 7, setting up the optimization problem in the matrix form we get ***2 points***

- $\min_{\underline{w}} \ \underline{w}^T H \underline{w} + 2\underline{f}\underline{w}$
  s.t. $A\underline{w} \leq \begin{pmatrix} -t \\ t \\ t \\ t \end{pmatrix}$
  where $\underline{f} = -\underline{y}^T X$
- $\min_{\underline{w}} \ \underline{w}^T H \underline{w} - 2\underline{f}\underline{w}$
  s.t. $A\underline{w} \leq \begin{pmatrix} -t \\ t \\ t \\ t \end{pmatrix}$
  where $\underline{f} = -\underline{y}^T X$
- $\min_{\underline{w}} \ \underline{w}^T H \underline{w} - 2\underline{f}\underline{w}$
  s.t. $A\underline{w} \leq \begin{pmatrix} t \\ t \\ t \\ t \end{pmatrix}$
  where $\underline{f} = -\underline{y}^T X$
- $\min_{\underline{w}} \ \underline{w}^T H \underline{w} + 2\underline{f}\underline{w}$
  s.t. $A\underline{w} \leq \begin{pmatrix} t \\ t \\ t \\ t \end{pmatrix}$
  where $\underline{f} = -\underline{y}^T X$

No, the answer is incorrect.
Score: 0
Accepted Answers:
$\min_{\underline{w}} \ \underline{w}^T H \underline{w} + 2\underline{f}\underline{w}$
s.t. $A\underline{w} \leq \begin{pmatrix} t \\ t \\ t \\ t \end{pmatrix}$
where $\underline{f} = -\underline{y}^T X$

9) In continuation with questions 7 and 8, suppose we need to reduce from 2^d constraints to $2d$ constraints where d is the dimensionality of the weight vector $\underline{w}$. This can be achieved ***2 points***

- By expressing each $\{w_i\}_{i=1}^{d}$ as a sum of two negative variables
- By expressing each $\{w_i\}_{i=1}^{d}$ as a sum of two non-negative variables
- By expressing each $\{w_i\}_{i=1}^{d}$ as a difference of two negative variables
- By expressing each $\{w_i\}_{i=1}^{d}$ as a difference of two non-negative variables

No, the answer is incorrect.
Score: 0
Accepted Answers:
By expressing each $\{w_i\}_{i=1}^{d}$ as a difference of two non-negative variables

10) Consider the objective function $J(\underline{w}) = R(\underline{w}) + \lambda\|\underline{w}\|_1$. ***1 point***
(True/False) The given objective is differentiable $\forall \underline{w}$.

- True
- False

No, the answer is incorrect.
Score: 0
Accepted Answers:
False

11) For the function $f(x) = |x|$, where x is a scalar, define the subdifferential as ***1 point***
$\partial f(x) = \begin{cases} -1 & \text{if } x < 0, \\ [-1, 1] & \text{if } x = 0, \\ +1 & \text{if } x > 0. \end{cases}$
what can you say about $\partial_{w_j} J(\underline{w})$ with the definition of $J(\underline{w})$ as in question 10.

- $\frac{\partial J(w)}{\partial w_j} = \begin{cases} \nabla_j R - \lambda & \text{if } w_j < 0, \\ [\nabla_j R - \lambda, \nabla_j R + \lambda] & \text{if } w_j = 0, \\ \nabla_j R + \lambda & \text{if } w_j > 0. \end{cases}$
- $\frac{\partial J(w)}{\partial w_j} = \begin{cases} \nabla_j R - \lambda & \text{if } w_j < 0, \\ [\nabla_j R + \lambda, \nabla_j R - \lambda] & \text{if } w_j = 0, \\ \nabla_j R + \lambda & \text{if } w_j > 0. \end{cases}$
- $\frac{\partial J(w)}{\partial w_j} = \begin{cases} \nabla_j R - \lambda & \text{if } w_j < 0, \\ [\nabla_j R - \lambda, \nabla_j R + \lambda] & \text{if } w_j = 0, \\ \nabla_j R - \lambda & \text{if } w_j > 0. \end{cases}$
- $\frac{\partial J(w)}{\partial w_j} = \begin{cases} \nabla_j R + \lambda & \text{if } w_j < 0, \\ [\nabla_j R - \lambda, \nabla_j R + \lambda] & \text{if } w_j = 0, \\ \nabla_j R + \lambda & \text{if } w_j > 0. \end{cases}$

No, the answer is incorrect.
Score: 0
Accepted Answers:
$\frac{\partial J(w)}{\partial w_j} = \begin{cases} \nabla_j R - \lambda & \text{if } w_j < 0, \\ [\nabla_j R - \lambda, \nabla_j R + \lambda] & \text{if } w_j = 0, \\ \nabla_j R + \lambda & \text{if } w_j > 0. \end{cases}$

12) In continuation with question 11, the conditions for the optima are ***2 points***

- $\begin{cases} \nabla_j R(\underline{w}) + \lambda sign(\underline{w}) = 0 & \text{if } |w_j| > 0, \\ |\nabla_j R(\underline{w})| \leq \lambda & \text{if } w_j = 0. \end{cases}$
- $\begin{cases} \nabla_j R(\underline{w}) - \lambda sign(\underline{w}) = 0 & \text{if } |w_j| > 0, \\ |\nabla_j R(\underline{w})| \leq \lambda & \text{if } w_j = 0. \end{cases}$
- $\begin{cases} \nabla_j R(\underline{w}) + \lambda sign(\underline{w}) = 0 & \text{if } |w_j| > 0, \\ |\nabla_j R(\underline{w})| \geq \lambda & \text{if } w_j = 0. \end{cases}$
- $\begin{cases} \nabla_j R(\underline{w}) - \lambda sign(\underline{w}) = 0 & \text{if } |w_j| > 0, \\ |\nabla_j R(\underline{w})| \geq \lambda & \text{if } w_j = 0. \end{cases}$

No, the answer is incorrect.
Score: 0
Accepted Answers:
$\begin{cases} \nabla_j R(\underline{w}) + \lambda sign(\underline{w}) = 0 & \text{if } |w_j| > 0, \\ |\nabla_j R(\underline{w})| \leq \lambda & \text{if } w_j = 0. \end{cases}$

13) (True/False) In continuation with question 12, the equations for optimality are linear. ***1 point***

- True
- False

No, the answer is incorrect.
Score: 0
Accepted Answers:
False