CS 446 / ECE 449 Homework 3

Naman Shukla

TOTAL POINTS

15 / 15

QUESTION 1

1 a **2 / 2**

✓ **- 0 pts** Correct

- 2 pts Incorrect

QUESTION 2

2 b **3 / 3**

✓ **- 0 pts** Correct

- 3 pts Incorrect

QUESTION 3

3 c **3 / 3**

✓ **- 0 pts** Correct

- 1 pts Negative gradient

- 1 pts Mixed exponential and sigmoid expression

- 2 pts Not expressed as a function of g(a)

- 3 pts Incorrect

QUESTION 4

4 d **5 / 5**

✓ **- 0 pts** Correct

- 5 pts Incorrect

QUESTION 5

5 e **2 / 2**

✓ **- 0 pts** Correct

- 1 pts Partially correct, the assumption should be independent and identically distributed

- 2 pts Incorrect, the assumption should be independent and identically distributed

University of Illinois Spring 2018

CS 446: Machine Learning
Homework 3: Binary Classification

Due on Tuesday, Feb 06, 2018, 11:59 a.m. Central Time

1. **[15 points]** Binary Classifiers

(a) In order to use a linear regression model for binary classification, how do we map the regression output $\mathbf{w}^\top\mathbf{x}$ to the class labels $y \in \{-1, 1\}$?

Your answer:

$$y^{(i)} \in \{-1, 1\}$$

$$\hat{y}^{(i)} = sign(\mathbf{w}^\top\mathbf{x})$$

(b) In logistic regression, the activation function $\mathrm{g}(a) = \frac{1}{1+e^{-a}}$ is called sigmoid. Then how do we map the sigmoid output $\mathrm{g}(\mathbf{w}^\top\mathbf{x})$ to binary class labels $y \in \{-1, 1\}$?

Your answer:

$$\mathrm{g}(\mathbf{w}^\top\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^\top\mathbf{x})}}$$

$$\mathrm{p}(y^{(i)} = 1|x^{(i)}) = \mathrm{g}(a) = \frac{1}{1 + e^{-(\mathbf{w}^\top\mathbf{x})}}$$

For classification, critical value = 0.5, we have

$y = 1$ when

$$\mathrm{p} \geq 0.5$$

and
$y = -1$ when

$$\mathrm{p} < 0.5$$

(c) Is it possible to write the derivative of the sigmoid function g w.r.t a, i.e. $\frac{\partial g}{\partial a}$, as a simple function of itself g? If so, how?

1 a 2 / 2

✓ - 0 pts Correct

- 2 pts Incorrect

University of Illinois Spring 2018

CS 446: Machine Learning

Homework 3: Binary Classification

Due on Tuesday, Feb 06, 2018, 11:59 a.m. Central Time

1. **[15 points]** Binary Classifiers

   (a) In order to use a linear regression model for binary classification, how do we map the regression output $\mathbf{w}^\top\mathbf{x}$ to the class labels $y \in \{-1, 1\}$?

   Your answer:

   $$y^{(i)} \in \{-1, 1\}$$

   $$\hat{y}^{(i)} = sign(\mathbf{w}^\top\mathbf{x})$$

   (b) In logistic regression, the activation function $\mathrm{g}(a) = \frac{1}{1+e^{-a}}$ is called sigmoid. Then how do we map the sigmoid output $\mathrm{g}(\mathbf{w}^\top\mathbf{x})$ to binary class labels $y \in \{-1, 1\}$?

   Your answer:

   $$\mathrm{g}(\mathbf{w}^\top\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^\top\mathbf{x})}}$$

   $$\mathrm{p}(y^{(i)} = 1|x^{(i)}) = \mathrm{g}(a) = \frac{1}{1 + e^{-(\mathbf{w}^\top\mathbf{x})}}$$

   For classification, critical value = 0.5, we have

   $y = 1$ when

   $$\mathrm{p} \geq 0.5$$

   and
   $y = -1$ when

   $$\mathrm{p} < 0.5$$

   (c) Is it possible to write the derivative of the sigmoid function g w.r.t a, i.e. $\frac{\partial g}{\partial a}$, as a simple function of itself g? If so, how?

2 b 3 / 3

✓ - 0 pts Correct

- 3 pts Incorrect

Your answer:

$$
\begin{aligned}
\frac{d}{da}\,\mathrm{g}(a) &= \frac{d}{da}\left[\frac{1}{1+e^{-a}}\right] \\
&= \frac{d}{da}\left(1+\mathrm{e}^{-a}\right)^{-1} \\
&= -(1+e^{-a})^{-2}(-e^{-a}) \\
&= \frac{e^{-a}}{(1+e^{-a})^2} \\
&= \frac{1}{1+e^{-a}} \cdot \frac{e^{-a}}{1+e^{-a}} \\
&= \frac{1}{1+e^{-a}} \cdot \frac{(1+e^{-a})-1}{1+e^{-a}} \\
&= \frac{1}{1+e^{-a}} \cdot \left(\frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}\right) \\
&= \frac{1}{1+e^{-a}} \cdot \left(1 - \frac{1}{1+e^{-a}}\right) \\
&= \mathrm{g}(a)\cdot(1-\mathrm{g}(a))
\end{aligned}
$$

(d) Assume quadratic loss is used in the logistic regression together with the sigmoid function. Then the program becomes:

$$\min_{\mathbf{w}} f(\mathbf{w}) := \frac{1}{2}\sum_i \left(y_i - \mathrm{g}(\mathbf{w}^\top \mathbf{x}_i)\right)^2$$

where $y \in \{0,1\}$. To solve it by gradient descent, what would be the $\mathbf{w}$ update equation?

3 C 3 / 3

✓ - 0 pts Correct

- 1 pts Negative gradient
- 1 pts Mixed exponential and sigmoid expression
- 2 pts Not expressed as a function of g(a)
- 3 pts Incorrect

Your answer:

$$loss = \frac{1}{2}\sum_i \left(a_{(i)}\right)^2$$

$$a_{(i)} = \left(y_i - \mathrm{g}(\mathbf{w}^\top \mathbf{x}_i)\right)$$

With step size $= \alpha$

$$\mathbf{w}_{(t+1)} = \mathbf{w}_{(t)} - \alpha \nabla_{\mathbf{w}} f(\mathbf{w})$$

where :

$$\frac{\partial loss}{\partial w_k} = \frac{\partial(\frac{1}{2}\sum_i \left(a_{(i)}\right)^2)}{\partial a_{(i)}} \times \frac{\partial a_{(i)}}{\partial\, \mathrm{g}(\mathbf{w}^\top \mathbf{x}_i)} \times \frac{\partial\, \mathrm{g}(\mathbf{w}^\top \mathbf{x}_i)}{\partial(\mathbf{w}^\top \mathbf{x}_i)} \times \frac{\partial(\mathbf{w}^\top \mathbf{x}_i)}{\partial \mathbf{w}_k}$$

$$\frac{\partial loss}{\partial w_k} = \sum_i \left(y_i - \mathrm{g}(\mathbf{w}^\top \mathbf{x}_i)\right) \times (-1) \times ((\mathrm{g}(\mathbf{w}^\top \mathbf{x}_i)) \cdot (1 - \mathrm{g}(\mathbf{w}^\top \mathbf{x}_i))) \times \mathbf{x}_i^k$$

Upon simplification:

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \sum_i \left(\mathrm{g}(\mathbf{w}^\top \mathbf{x}_i) - y_i\right) ((\mathrm{g}(\mathbf{w}^\top \mathbf{x}_i)) \cdot (1 - \mathrm{g}(\mathbf{w}^\top \mathbf{x}_i))) \cdot \mathbf{x}_i$$

(e) Assume $y \in \{-1, 1\}$. Consider the following program for logistic regression:

$$\min_{\mathbf{w}} f(\mathbf{w}) := \sum_i \log\left(1 + \exp(-y^{(i)} \mathbf{w}^T \phi(x^{(i)}))\right).$$

The above program for binary classification makes an assumption on the samples/data points. What is the assumption?

Your answer: The assumption is that the samples/data points are independent and identically distributed (i.i.d)

4 d 5 / 5

✓ - 0 pts Correct

- 5 pts Incorrect

Your answer:

$$loss = \frac{1}{2}\sum_i \left(a_{(i)}\right)^2$$

$$a_{(i)} = \Big(y_i - \mathrm{g}(\mathbf{w}^\top \mathbf{x}_i)\Big)$$

With step size $= \alpha$

$$\mathbf{w}_{(t+1)} = \mathbf{w}_{(t)} - \alpha \nabla_{\mathbf{w}} f(\mathbf{w})$$

where :

$$\frac{\partial loss}{\partial w_k} = \frac{\partial(\frac{1}{2}\sum_i \left(a_{(i)}\right)^2)}{\partial a_{(i)}} \times \frac{\partial a_{(i)}}{\partial\, \mathrm{g}(\mathbf{w}^\top \mathbf{x}_i)} \times \frac{\partial\, \mathrm{g}(\mathbf{w}^\top \mathbf{x}_i)}{\partial(\mathbf{w}^\top \mathbf{x}_i)} \times \frac{\partial(\mathbf{w}^\top \mathbf{x}_i)}{\partial \mathbf{w}_k}$$

$$\frac{\partial loss}{\partial w_k} = \sum_i \Big(y_i - \mathrm{g}(\mathbf{w}^\top \mathbf{x}_i)\Big) \times (-1) \times ((\mathrm{g}(\mathbf{w}^\top \mathbf{x}_i)) \cdot (1 - \mathrm{g}(\mathbf{w}^\top \mathbf{x}_i))) \times \mathbf{x}_i^k$$

Upon simplification:

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \sum_i \Big(\mathrm{g}(\mathbf{w}^\top \mathbf{x}_i) - y_i\Big) ((\mathrm{g}(\mathbf{w}^\top \mathbf{x}_i)) \cdot (1 - \mathrm{g}(\mathbf{w}^\top \mathbf{x}_i))) \cdot \mathbf{x}_i$$

(e) Assume $y \in \{-1, 1\}$. Consider the following program for logistic regression:

$$\min_{\mathbf{w}} f(\mathbf{w}) := \sum_i \log\Big(1 + \exp(-y^{(i)}\mathbf{w}^T \phi(x^{(i)}))\Big).$$

The above program for binary classification makes an assumption on the samples/data points. What is the assumption?

Your answer: The assumption is that the samples/data points are independent and identically distributed (i.i.d)

5 e 2 / 2

- ✓ **- 0 pts** Correct
- **- 1 pts** Partially correct, the assumption should be independent and identically distributed
- **- 2 pts** Incorrect, the assumption should be independent and identically distributed