CS 446 / ECE 449 Homework 2

Naman Shukla

TOTAL POINTS

8 / 8

QUESTION 1

1 Linear Regression Basics 6 / 6

✓ **- 0 pts Correct**

- 1 pts 1.(f) not closed form

- 1 pts 1.(a) Incorrect

- 0.5 pts 1.(a) Missing summation

- 0.5 pts 1.(a) Undefined notation, N (or n, K) is used but undefined

- 0.5 pts 1.(d) Inconsistent with the loss in (a)

- 0.5 pts 1.(d) Not in vector form

- 0.5 pts 1.(d) index of y is missing

- 1 pts 1.(d) Incorrect

- 0.5 pts 1.(e) Inconsistent with the loss in (a)

- 1 pts 1.(e) Incorrect

- 1 pts 1.(f) Incorrect

- 0 pts Please select pages for answers

- 0.5 pts 1.(f) u should be a column vector

- 0.5 pts 1.(f) undefined notation

QUESTION 2

2 Linear Regression Probabilistic Interpretation 2 / 2

✓ **+ 1 pts (a) Correct**

+ 0.5 pts (a) Undefined Notation

+ 0 pts (a) Incorrect

✓ **+ 1 pts (b) Correct**

+ 0.5 pts (b) Undefined Notation

+ 0.5 pts (b) Incorrect but consistent with (a)

+ 0 pts (b) Incorrect

+ 0.5 pts (b) Constant term dropped/not specified

+ 0.5 pts (a) Typo

University of Illinois Spring 2018

CS 446: Machine Learning
Homework 2

Due on Tuesday, January 30, 2018, 11:59 a.m. Central Time

1. **[6 points]** Linear Regression Basics
Consider a linear model of the form $\hat{y}^{(i)} = \mathbf{w}^\intercal \mathbf{x}^{(i)} + b$, where $\mathbf{w}, \mathbf{x} \in \mathbb{R}^K$ and $b \in \mathbb{R}$. Next, we are given a training dataset, $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}$ denoting the corresponding input-target example pairs.

   (a) What is the loss function, $\mathcal{L}$, for training a linear regression model? (Don't forget the $\frac{1}{2}$)

   Your answer: Given $\hat{y}^{(i)} = \mathbf{w}^\intercal \mathbf{x}^{(i)} + b$, where $\mathbf{w}, \mathbf{x} \in \mathbb{R}^K$ and $b \in \mathbb{R}$.
   The loss function is given by :

   $$\mathcal{L} = \frac{1}{2} \sum_{i=1}^{|\mathcal{D}|} |\hat{y}^{(i)} - y^{(i)}|^2$$

   Where $\hat{y}$ is given by

   $$\hat{y} = [\mathbf{w_1 w_2} \cdots \mathbf{w}_{|\mathcal{D}|}]^\intercal [\mathbf{x_1 x_2} \cdots \mathbf{x}_{|\mathcal{D}|}] + b$$

   $$\hat{y} = [\mathbf{w} \;\; b]^\intercal [\mathbf{x} \;\; 1]$$

   Therefore,

   $$\mathcal{L} = \frac{1}{2} \|[\mathbf{w} \;\; b]^\intercal [\mathbf{x} \;\; 1] - \mathbf{y}\|_2^2$$

   (b) Compute $\frac{\partial \mathcal{L}}{\partial \hat{y}^{(i)}}$.

   Your answer:

   $$\mathcal{L} = \frac{1}{2} \sum_{i=1}^{|\mathcal{D}|} |(w_1 x_1 + w_2 x_2 \cdots w_K x_K + b) - y^{(i)}|^2$$

   $$\mathcal{L} = \frac{1}{2} \sum_{i=1}^{|\mathcal{D}|} |\hat{y}^{(i)} - y^{(i)}|^2$$

   Now taking partial derivative w.r.t $\hat{y}^{(i)}$ and applying chain rule, we get :

   $$\frac{\partial \mathcal{L}}{\partial \hat{y}^{(i)}} = |(1)|(\hat{y}^{(i)} - y^{(i)})$$

   $$\frac{\partial \mathcal{L}}{\partial \hat{y}^{(i)}} = (\hat{y}^{(i)} - y^{(i)})$$

(c) Compute $\frac{\partial \hat{y}^{(i)}}{\partial \mathbf{w_k}}$, where $\mathbf{w_k}$ denotes the k^{th} element of $\mathbf{w}$.

Your answer:

$$\hat{y}^{(i)} = (w_1 x_1 + w_2 x_2 \cdots w_K x_K + b)$$

taking derivative, we get:

$$\frac{\partial \hat{y}^{(i)}}{\partial \mathbf{w}_k} = {\mathbf{x}_k}^{(i)}$$

(d) Putting the previous parts together, what is $\nabla_{\mathbf{w}} \mathcal{L}$?

Your answer:

$$\mathcal{L} = \frac{1}{2} \sum_{i=1}^{|\mathcal{D}|} |\hat{y}^{(i)} - y^{(i)}|^2$$

$$\mathcal{L} = \frac{1}{2} \sum_{i=1}^{|\mathcal{D}|} |(w_1 x_1 + w_2 x_2 \cdots w_K x_K + b) - y^{(i)}|^2$$

Now combining above two equations:

$$\frac{\partial \mathcal{L}}{\partial w_k} = \frac{\partial \mathcal{L}}{\partial \hat{y}^{(i)}} \times \frac{\partial \hat{y}^{(i)}}{\partial \mathbf{w}_k}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w_k}} = ({\mathbf{x}_k}^{(i)}) \sum_{i=1}^{|\mathcal{D}|} (\hat{y}^{(i)} - y^{(i)})$$

Generalizing,

$$\nabla_{\mathbf{w}} \mathcal{L} = \mathbf{x}^{\intercal}([\mathbf{w} \;\; b]^{\intercal}[\mathbf{x} \;\; 1] - \mathbf{y})$$

(e) Compute $\frac{\partial \mathcal{L}}{\partial b}$.

Your answer:

$$\mathcal{L} = \frac{1}{2} \sum_{i=1}^{|\mathcal{D}|} |(w_1 x_1 + w_2 x_2 \cdots w_K x_K + b) - y^{(i)}|^2$$

Now taking partial derivative w.r.t b and applying chain rule, we get :

$$\frac{\partial \mathcal{L}}{\partial b} = |(1)| \sum_{i=1}^{|\mathcal{D}|} (\hat{y}^{(i)} - y^{(i)})$$

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{|\mathcal{D}|} (\hat{y}^{(i)} - y^{(i)})$$

(f) For convenience, we group $\mathbf{w}$ and b together into $\mathbf{u}$, then we denote $\mathbf{z} = [\mathbf{x} \;\; 1]$. (*i.e.* $\hat{y} = \mathbf{u}^{\intercal}[x, 1] = \mathbf{w}^{\intercal} x + b$). What are the optimal parameters $\mathbf{u}^* = [\mathbf{w}^*, b^*]$? Use the notation $\mathbf{Z} \in \mathbb{R}^{|D| \times (K+1)}$ and $\mathbf{y} \in \mathbb{R}^{|D|}$ in the answer. Where, each row of $\mathbf{Z}, \mathbf{y}$ denotes an example input-target pair in the dataset.

Your answer: Given that : $\mathbf{z} = [\mathbf{x} \;\; 1]$ and $\mathbf{u} = [\mathbf{w} \;\; b]$
Using what we have established in part(d), and using the fact that,
$\nabla_{\mathbf{w}}\mathcal{L} = \nabla_{\mathbf{u}}\mathcal{L}$ as $\mathbf{u}$ is affine function of $\mathbf{w}$ with 1 and b as the parameters.

We get,

$$\nabla_{\mathbf{u}}\mathcal{L} = \mathbf{z}^\intercal(\mathbf{u}^\intercal\mathbf{z} - \mathbf{y})$$

To find the optimal parameter value for $\mathbf{u}$ (i.e. $\mathbf{u}^*$), we set $\nabla_{\mathbf{u}}\mathcal{L}$ to 0, we get,

$$0 = \mathbf{z}^\intercal(\mathbf{u}^\intercal\mathbf{z} - \mathbf{y})$$

thus we get,

$$\mathbf{u}^* = (\mathbf{z}^\intercal\mathbf{z})^{-1}\mathbf{z}^\intercal\mathbf{y}$$

2. **[2 points]** Linear Regression Probabilistic Interpretation
Consider that the input $x^{(i)} \in \mathbb{R}$ and target variable $y^{(i)} \in \mathbb{R}$ to have to following relationship.

$$y^{(i)} = w \cdot x^{(i)} + \epsilon^{(i)}$$

where, ϵ is independently and identically distributed according to a Gaussian distribution with zero mean and unit variance.

(a) What is the conditional probability $p(y^{(i)}|x^{(i)}, w)$.

Your answer:

$$p(y^{(i)}|x^{(i)}, w) = \frac{1}{|\sigma|\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y^{(i)} - \mu}{|\sigma|}\right)^2\right]$$

Now given is ϵ is distributed as $\mathcal{N}(0, 1)$
Therefore, our model equation would be as follows :

$$\mu = w \cdot \mathbf{x} + 0$$

$$\sigma = 1$$

Hence, we get :

$$p(y^{(i)}|x^{(i)}, w) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y^{(i)} - w \cdot x^{(i)}}{1}\right)^2\right]$$

(b) Given a dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}$, what is the negative log likelihood of the dataset according to our model? (Simplify.)

1 Linear Regression Basics 6 / 6

✓ - 0 pts **Correct**

- 1 pts 1.(f) not closed form
- 1 pts 1.(a) Incorrect
- 0.5 pts 1.(a) Missing summation
- 0.5 pts 1.(a) Undefined notation, N (or n, K) is used but undefined
- 0.5 pts 1.(d) Inconsistent with the loss in (a)
- 0.5 pts 1.(d) Not in vector form
- 0.5 pts 1.(d) index of y is missing
- 1 pts 1.(d) Incorrect
- 0.5 pts 1.(e) Inconsistent with the loss in (a)
- 1 pts 1.(e) Incorrect
- 1 pts 1.(f) Incorrect
- 0 pts Please select pages for answers
- 0.5 pts 1.(f) u should be a column vector
- 0.5 pts 1.(f) undefined notation

Your answer: Given that : $\mathbf{z} = [\mathbf{x} \;\; 1]$ and $\mathbf{u} = [\mathbf{w} \;\; b]$
Using what we have established in part(d), and using the fact that,
$\nabla_{\mathbf{w}}\mathcal{L} = \nabla_{\mathbf{u}}\mathcal{L}$ as $\mathbf{u}$ is affine function of $\mathbf{w}$ with 1 and b as the parameters.

We get,

$$\nabla_{\mathbf{u}}\mathcal{L} = \mathbf{z}^\intercal(\mathbf{u}^\intercal\mathbf{z} - \mathbf{y})$$

To find the optimal parameter value for $\mathbf{u}$ (i.e. $\mathbf{u}^*$), we set $\nabla_{\mathbf{u}}\mathcal{L}$ to 0, we get,

$$0 = \mathbf{z}^\intercal(\mathbf{u}^\intercal\mathbf{z} - \mathbf{y})$$

thus we get,

$$\mathbf{u}^* = (\mathbf{z}^\intercal\mathbf{z})^{-1}\mathbf{z}^\intercal\mathbf{y}$$

2. **[2 points]** Linear Regression Probabilistic Interpretation
Consider that the input $x^{(i)} \in \mathbb{R}$ and target variable $y^{(i)} \in \mathbb{R}$ to have to following relationship.

$$y^{(i)} = w \cdot x^{(i)} + \epsilon^{(i)}$$

where, ϵ is independently and identically distributed according to a Gaussian distribution with zero mean and unit variance.

(a) What is the conditional probability $p(y^{(i)}|x^{(i)}, w)$.

Your answer:

$$p(y^{(i)}|x^{(i)}, w) = \frac{1}{|\sigma|\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y^{(i)} - \mu}{|\sigma|}\right)^2\right]$$

Now given is ϵ is distributed as $\mathcal{N}(0, 1)$
Therefore, our model equation would be as follows :

$$\mu = w \cdot \mathbf{x} + 0$$

$$\sigma = 1$$

Hence, we get :

$$p(y^{(i)}|x^{(i)}, w) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y^{(i)} - w \cdot x^{(i)}}{1}\right)^2\right]$$

(b) Given a dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}$, what is the negative log likelihood of the dataset according to our model? (Simplify.)

Your answer:
The log-likelihood (for i.i.d) is defined as,

$$l(w) : \log p(\mathcal{D}|w) = \sum_{i=1}^{|\mathcal{D}|} \log p(y^{(i)}|x^{(i)}, w)$$

Therefore, the negative log-likelihood is,

$$NLL(w) = -\sum_{i=1}^{|\mathcal{D}|} \log p(y^{(i)}|x^{(i)}, w)$$

using part(a),

$$NLL(w) = -\sum_{i=1}^{|\mathcal{D}|} \log \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y^{(i)} - w \cdot x^{(i)}}{1}\right)^2\right]$$

Upon simplification,

$$NLL(w) = \frac{1}{2} RSS(\mathbf{w}) + \frac{|\mathcal{D}|}{2} \log(2\pi)$$

where,

$$RSS(w) = \sum_{i=1}^{|\mathcal{D}|} (y^{(i)} - w \cdot x^{(i)})^2$$

2 Linear Regression Probabilistic Interpretation 2 / 2

- ✓ + 1 pts **(a) Correct**
  - + 0.5 pts (a) Undefined Notation
  - + 0 pts (a) Incorrect
- ✓ + 1 pts **(b) Correct**
  - + 0.5 pts (b) Undefined Notation
  - + 0.5 pts (b) Incorrect but consistent with (a)
  - + 0 pts (b) Incorrect
  - + 0.5 pts (b) Constant term dropped/not specified
  - + 0.5 pts (a) Typo