



M902

Βασικές Μαθηματικές Έννοιες στη Γλωσσική Τεχνολογία

Project Recap

Κυλάφη Χριστίνα-Θεανώ

LT1200012

January, 2021

TABLE OF CONTENTS

Question 1 _____ 3

Question 2 _____ 5

Question 3 _____ 7

Question 4 _____ 8

Question 5 _____ 9

Question 1

1. Let $X_{1 \times n} = [x_1, x_2, \dots, x_n]$ the input feature values of the neural network model and $W_{1 \times n} = [w_1, w_2, \dots, w_n]$, the learnt weights during the training process (and w_0 a bias to be added in the weighted sum).

Then the weighted sum is calculated as follows:

$$\begin{aligned} z_{1 \times 1} &= X W^T = [X]_{1 \times n} [W^T]_{n \times 1} = [x_1 \ x_2 \ \dots \ x_{n-1} \ x_n] \begin{bmatrix} w_1 \\ w_2 \\ \dots \\ w_{n-1} \\ w_n \end{bmatrix} \\ &= [x_1 w_1 + x_2 w_2 + \dots + x_n w_n] \\ &= \left[\sum_{i=1}^n x_i w_i \right] \end{aligned}$$

where W^T the transpose of the weight matrix. The previous output could also be calculated by

transposing the input values matrix and the formula $z_{1 \times 1} = W X^T = \sum_{i=1}^n w_i x_i = \sum_{i=1}^n x_i w_i$.

2. Let σ the sigmoid function, where $\sigma(x) = \frac{1}{1 + e^{-x}}$.

A bias w_0 is added to the weighted sum z :

$$z' \rightarrow \text{bias} + z = w_0 + \sum_{i=1}^n x_i w_i = w_0 + x_1 w_1 + x_2 w_2 + \dots + x_n w_n$$

Then, in order to calculate the derivative of the output, we convert z' into a function of the $f(x) = ax + c$ form :

$$f(w_1) = z'(w_1) = w_0 + x_1 w_1 + x_2 w_2 + \dots + x_n w_n = x_1 w_1 + c,$$

where:

$$x_1, c = w_0 + x_2 w_2 + \dots + x_n w_n, \text{ constants.}$$

Then:

$$\begin{aligned}\sigma(f(w_1)) &= \frac{1}{1 + e^{-f(w_1)}} = \frac{1}{1 + e^{-(x_1 w_1 + c)}} = \frac{1}{1 + \frac{1}{e^{(x_1 w_1 + c)}}} = \frac{1}{\frac{e^{(x_1 w_1 + c)} + 1}{e^{(x_1 w_1 + c)}}} \\ &= \frac{e^{(x_1 w_1 + c)}}{e^{(x_1 w_1 + c)} + 1} = \frac{e^{x_1 w_1} e^c}{e^{x_1 w_1} e^c + 1} \quad (1)\end{aligned}$$

$\sigma(f(w_1))$ is differentiable in $Dom(\sigma)$, as a composition of differentiable functions σ , f , with its derivative $\sigma'(f(w_1))$ as follows:

$$\begin{aligned}\frac{d(\sigma(f(w_1)))}{dw_1} &= \sigma(f(w_1)) (1 - \sigma(f(w_1))) = \sigma'(f(w_1)) \\ &= \left(\frac{1}{1 + e^{-f(w_1)}} \right)' \stackrel{(1)}{=} \left(\frac{e^{x_1 w_1} e^c}{e^{x_1 w_1} e^c + 1} \right)' \\ &= \frac{(e^{x_1 w_1} e^c)'(e^{x_1 w_1} e^c + 1) - (e^{x_1 w_1} e^c)(e^{x_1 w_1} e^c + 1)'}{(e^{x_1 w_1} e^c + 1)^2} \\ &= \frac{(x_1 e^{x_1 w_1} e^c)(e^{x_1 w_1} e^c + 1) - (e^{x_1 w_1} e^c)(x_1 e^{x_1 w_1} e^c)}{(e^{x_1 w_1} e^c + 1)^2} \\ &= \frac{(x_1 e^{x_1 w_1} e^c)(e^{x_1 w_1} e^c + 1 - e^{x_1 w_1} e^c)}{(e^{x_1 w_1} e^c + 1)^2} \\ &= \frac{x_1 e^{x_1 w_1 + c}}{(e^{x_1 w_1} e^c + 1)^2} \rightarrow \text{derivative of the network output } z' \text{ (} z \text{ with added bias } w_0 \text{)} \\ &\quad \text{with respect to the parameter } w_1\end{aligned}$$

Question 2

Let

$$D = \{ (x_1, y_1), (x_2, y_2), \dots, (x_N, y_N) \}$$
$$y_i \in \{ 0, 1 \}, i \in [1, 2, \dots, N]$$

1. In order to perform density estimation on the given data, we use the Maximum Likelihood Estimation method (**MLE**). Assuming that the observations are independent and identically distributed - **I.I.D.** (following the Bernoulli distribution $Bern(Y = y; \theta)$, where $\theta \in [0, 1]$), the likelihood of all the data is the product of the respective **likelihood** of each point (x_i, y_i) , $i \in [1, \dots, N]$:

$$\begin{aligned} \mathcal{L}(\theta) &= P(x_1 \rightarrow y_1, x_2 \rightarrow y_2, \dots, x_n \rightarrow y_n | \theta) \\ &= \prod_{i=1}^n P(x_i \rightarrow y_i | \theta) = \prod_{i=1}^n \hat{y}_i^{y_i} (1 - \hat{y}_i)^{1-y_i} \\ &= \prod_{i=1}^n f(x_i | \theta)^{y_i} (1 - f(x_i | \theta))^{1-y_i} \quad (1) \end{aligned}$$

2. The expression above is **not** part of the **optimal** solution for the task. First of all, it is not so easy to differentiate (during the maximisation process). Also, the sequential multiplications can lead to considerably **small** values, a situation that involves the risk of **underflow** (running out of floating point precision which means lack of enough bits to represent those really small numbers, thus rounding them to ~ 0). Therefore, the natural logarithm of this expression is preferred (transition from **product** to **sum**). Summing is less expensive than multiplication and the logarithm is a monotonically increasing function, ensuring that the maximum value occurs at the same point as the original probability function. For the aforementioned reasons, we alter (1) using the logarithm function and calculate the log-likelihood as follows:

$$\mathcal{LL}(\theta) = \log(\mathcal{L}(\theta))$$

$$\mathcal{LL}(\theta) = \log(\mathcal{L}(\theta)) = \log\left(\prod_{i=1}^n f(x_i | \theta)^{y_i} (1 - f(x_i | \theta))^{1-y_i}\right)$$

$$\begin{aligned}
&= \sum_{i=1}^n \log \left(f(x_i | \theta)^{y_i} (1 - f(x_i | \theta))^{1-y_i} \right) \\
&= \sum_{i=1}^n (y_i \log(f(x_i | \theta)) + (1 - y_i) \log(1 - f(x_i | \theta))) \\
&= \sum_{i=1}^n y_i \log(f(x_i | \theta)) + \sum_{i=1}^n (1 - y_i) \log(1 - f(x_i | \theta)) \\
&= \left(\sum_{i=1}^n y_i \right) \log(f(x_i | \theta)) + \left(\sum_{i=1}^n (1 - y_i) \right) \log(1 - f(x_i | \theta)) \\
&= Y \log(f(x_i | \theta)) + (n - Y) \log(1 - f(x_i | \theta)),
\end{aligned}$$

where

$$Y = \sum_{i=1}^n y_i$$

Question 3

1. The set $\{-4, -3, -2, -1, 0, 1, 2, 3, \dots\}$ can also be notated as:

$b \rightarrow \{x \in \mathbb{Z} : -4 \leq x\}$, which means that the set consists of integer numbers that are greater or equal than the integer -4 .

2. Let

$A = \{\text{Greek articles}\} = \{\text{οι, η, τους, ένας, των, ...}\}$

$B = \{\text{Greek vowels}\} = \{\alpha, \epsilon, \eta, \iota, \omicron, \upsilon, \omega\}$

Then:

$A \cap B = \{\text{Greek articles}\} \cap \{\text{Greek vowels}\}$

$= \{\text{οι, η, τους, ένας, των, ...}\} \cap \{\alpha, \epsilon, \eta, \iota, \omicron, \upsilon, \omega\}$

$= \{\alpha, \epsilon, \eta, \iota, \omicron, \upsilon, \omega\} = \{\text{Greek vowels}\}$

$= B$

3. Let

$A = \{\text{Words in Document A}\}$

$B = \{\text{Words in Document B}\}$

$A \subset B$

Then:

a. $B - A$: The set consisting of words contained in Document B that are not contained in Document A .

b. $A \subset B \implies |A| \leq |B|$, where $|A|$, $|B|$ the respective cardinalities of sets A , B .

c. Sets A and B , carry information about the unique words that occur in the respective documents (a set cannot contain duplicates by definition). No information is given about the frequency of each word, thus the total words comprising the documents. Therefore, no safe conclusion can be made concerning the lengths of the two Documents A and B .

Question 4

Let **2, 4, 15, 8, 7, 7, 6, 3, 3, 3, 40** observations (x_i , $i = 1, 2, \dots, 11$) of a random variable X .

Then the mean and median values are calculated as follows:

a. Mean value:

$$\frac{\sum_{i=1}^{11} x_i}{11} = \frac{2 + 4 + 15 + 8 + 7 + 7 + 6 + 3 + 3 + 3 + 40}{11} = 8.91$$

Median value:

2, 3, 3, 3, 4, **6**, 7, 7, 8, 15, 40

b. The **mean** value is the most commonly used statistical measure of central tendency with regard to an observation list. However, sometimes due to the structure of the observations' distribution, the **median** is more representative and a more appropriate metric of the data. For example, **not normally distributed** values or the existence of **outliers**, considerably affect the value of the distribution's mean. In our case, there is an **outlier** of value **40** in the dataset, so even though the two values, mean and median, are not too different from one another (**8.91** VS **6**), the **mean** value exhibits **sensitivity** to the **outlier**, **proving** that the **median** is a **better** choice for getting a proper **idea** of the **data distribution** given above.

Question 5

A **statistical measure** for the evaluation of a **word's** importance in a document, is **TF-IDF**, that stands for term frequency-inverse document frequency. It can successfully been used for the tasks of relevance scoring and ranking (recommendation/information retrieval/search) as well as filtering (text cleaning) to name a few.

The calculated value represents the weight of each word, consisting of two parts that measure the importance of the word in a particular document with consideration to both the exact document and the corpus (total collection of documents) factors. This way, the significance of the word is estimated in a more accurate manner.

Specifically, the TF-IDF of a word $w_i \in W$ in a document $d_i \in D$, where W the total number of unique words contained in all the corpus comprising D documents, is calculated as follows:

$$TF(w_i, d_i) = \frac{occ_{w_i d_i}}{\sum_{i=1}^W occ_{w_i d_i}}$$

where

$occ_{w_i d_i}$, the occurrences of the word w_i in document d_i

$\sum_{i=1}^W occ_{w_i d_i}$, the sum of all the words in document d_i (length of d_i)

$$TF(w_i, d_i) = \frac{occ_{w_i d_i}}{\sum_{i=1}^W occ_{w_i d_i}}$$

