Homework 1 – Deep Learning (CS/DS541, Murai, Fall 2022)

(Adapted from Prof. Whitehill)

Rohin Siddhartha Palaniappan Venkateswaran

Solutions

1 Python and Numpy Warm-up Exercises

This section of the homework is written in the python file titled "hw1_cs541_dl.py".Before running the code, please follow the instructions given below:

After downloading the file, navigate to the folder where the file was downloaded, open terminal and make the script executable by typing the command given below

```
sudo chmod +x "hw1_cs541_dl.py"
```

Now open "hw1_cs541_dl.py" and edit the file paths to load the data for Linear Regression as well as Probability distributions.

For Linear Regression, edit the file paths as follows **in quotes** inside **np.load()**:
Line 103: Type the full path for training data
Line 104: Type the full path for output labels for training data
Line 105: Type the full path for testing data
Line 106: Type the full path for output labels for testing data

For Probability distribution:
Line 227: Type the full path to the PoissonX.npy data **in quotes** inside **np.load()**

Now in order to run the code, navigate to the directory where the code was downloaded, and type the following command given below

```
python3 hw1_cs541_dl.py
```

The outputs of section **1**, **2** and **3** should be visible consecutively.

Note: Another alternative way to run the code is to open this Colab notebook and run the cells from top to bottom. The data for linear regression is automatically loaded in the Colab notebook. The PoissonX.npy has to be uploaded into the run-time from the local system.

2 Linear Regression via Analytical Solution

Training error: 50.667115870146546
Testing error: 272.21887728091906

3 Probability Distributions

(a) **Estimating the Parameters of a Probability Distribution:**

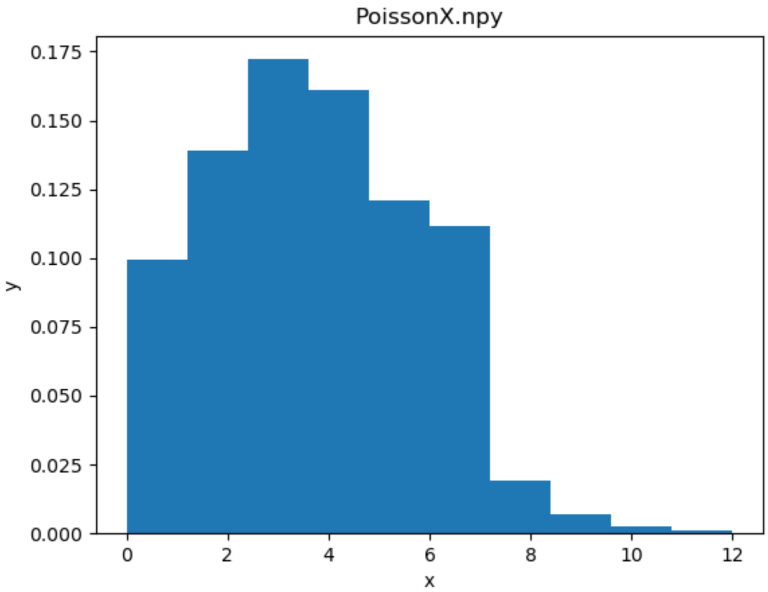

Figure 1: Empirical probability distribution of data in PoissonX.npy

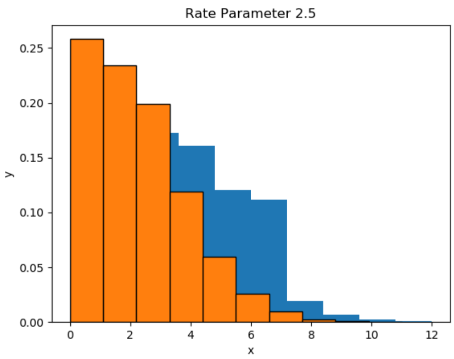

Figure 2: Rate parameter: 2.5

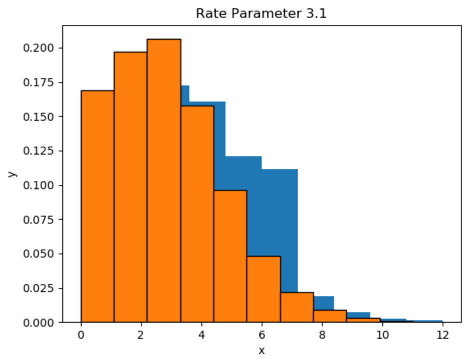

Figure 3: Rate parameter: 3.1

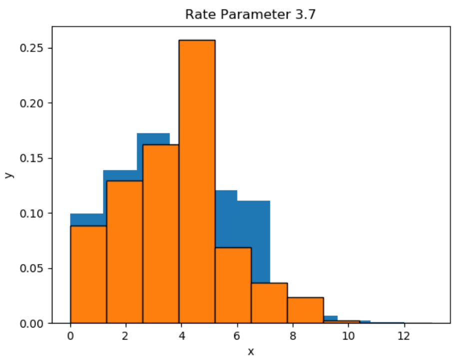

Figure 4: Rate parameter: 3.7

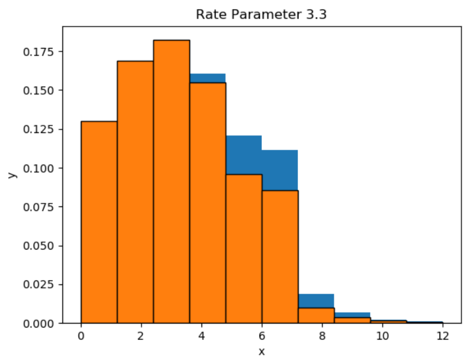

Figure 5: Rate parameter: 3.3

Hence, by visual inspection, the parameter values which are most consistent with the given data are **3.1 and 3.3**.
(b) **Conditional Probability Distributions to Represent the Uncertainty of Functions:**
i. For $\mathbf{x}$ **values with large magnitude**, the corresponding value of y tends to be larger.
ii. For $\mathbf{x}$ **values with small magnitude**, the uncertainty in the corresponding value of y tend to be larger.

4 Proofs/Derivations

(a) Let $\nabla_{\mathbf{x}} f(\mathbf{x})$ represent the column vector containing all the partial derivatives of f w.r.t. $\mathbf{x}$, i.e.,

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

For any two column vectors $\mathbf{x}$,$\mathbf{a} \in \mathbb{R}^n$, prove that

$$\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{a}) = \nabla_{\mathbf{x}}(\mathbf{a}^\top \mathbf{x}) = \mathbf{a}$$

Solution:

$$\begin{aligned} \nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{a}) &= \nabla_{\mathbf{x}} \sum_{i=1}^{n} a_i x_i \\ &= \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{bmatrix} \sum_{i=1}^{n} a_i x_i \end{aligned}$$

Consider the first derivative calculation,

$$\frac{\partial}{\partial x_1} \sum_{i=1}^{n} a_i x_i = \frac{\partial}{\partial x_1} \big(a_1 x_1 + \dots a_n x_n \big) = a_1$$

Repeating the same procedure for all the other n terms,

$$\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{a}) = \nabla_{\mathbf{x}}(\mathbf{a}^\top \mathbf{x}) = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = a$$

(b) Prove that $\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A} \mathbf{x}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$ for any column vector $\mathbf{x} \in \mathbb{R}^n$ and any $n \times n$ matrix $\mathbf{A}$.

Solution:

$$\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A} \mathbf{x}) = \nabla_{\mathbf{x}} \left(\begin{bmatrix} x_1 \dots x_n \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \dots & \mathbf{A}_{1n} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{n1} & \dots & \mathbf{A}_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \right)$$

$$= \nabla_{\mathbf{x}} \left(\begin{bmatrix} x_1\mathbf{A}_{11} + x_2\mathbf{A}_{21} + \ldots x_n\mathbf{A}_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \right)$$

$$= \nabla_{\mathbf{x}} \left(\textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} x_i A_{ij} x_j \right)$$

$$= \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{bmatrix} \left(\textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} x_i A_{ij} x_j \right)$$

After differentiating the term above, we get

$$= \sum_{j} x_i A_{ij} + \sum_{j} x_j A_{ji}$$

$$= \mathbf{A} \left(x_1 \ldots x_n \right) + \mathbf{A}^\top \left(x_1 \ldots x_n \right)$$

$$= \begin{bmatrix} \mathbf{A}_1 \left(x_1 \ldots x_n \right) + \mathbf{A}_1^\top \left(x_1 \ldots x_n \right) \\ \vdots \\ \mathbf{A}_n \left(x_1 \ldots x_n \right) + \mathbf{A}_n^\top \left(x_1 \ldots x_n \right) \end{bmatrix}$$

$$= (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$$

(c) Based on the theorem above, prove that $\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}\mathbf{x}) = 2\mathbf{A}\mathbf{x}$

Solution: From previous derivation, we know that

$$\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}\mathbf{x}) = \nabla_{\mathbf{x}} \left(\textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} x_i A_{ij} x_j \right)$$

$$= \sum_{j=1}^{n} x_j A_{ij} + \sum_{j=1}^{n} x_j A_{ji}$$
$$= \sum_{j=1}^{n} 2A_{ij} x_j$$

(Since A is a symmetric matrix, $\mathbf{A}_{ij} = \mathbf{A}_{ji}$)

$$= 2\mathbf{A}\mathbf{x}$$

(d) Based on the theorems above, prove that $\nabla_{\mathbf{x}}[(\mathbf{Ax}+\mathbf{b})^\top(\mathbf{Ax}+\mathbf{b})] = 2\mathbf{A}^\top(\mathbf{Ax}+\mathbf{b})$ for any column vector $\mathbf{x} \in \mathbb{R}^n$, any symmetric $n \times n$ matrix $\mathbf{A}$, and any constant column vector $\mathbf{b} \in \mathbb{R}^n$

Solution:

$$\begin{aligned}
\nabla_{\mathbf{x}}[(\mathbf{Ax}+\mathbf{b})^\top(\mathbf{Ax}+\mathbf{b})] &= \nabla_{\mathbf{x}}[(\mathbf{x}^\top\mathbf{A}^\top+\mathbf{b}^\top)(\mathbf{Ax}+\mathbf{b})] \\
&= \nabla_{\mathbf{x}}(\mathbf{x}^\top\mathbf{A}^\top\mathbf{Ax}+\mathbf{x}^\top\mathbf{A}^\top\mathbf{b}+\mathbf{b}^\top\mathbf{Ax}+\mathbf{b}^\top\mathbf{b})
\end{aligned}$$

Since $\mathbf{A}^\top\mathbf{A}, \mathbf{b}^\top\mathbf{b}$ are symmetric matrices, after differentiation, we get

$$\begin{aligned}
&= 2\mathbf{A}^\top\mathbf{Ax}+2\mathbf{A}^\top\mathbf{b} \\
&= 2\mathbf{A}^\top(\mathbf{Ax}+\mathbf{b})
\end{aligned}$$