Report for Deep Learning (CS541)_Homework1:

1g, -1
1l, -1 avoid using loop
1n, -1

Report By:
Swapneel Dhananjay Wagholikar (WPI ID: 257598983)
Anuj Pradeep Pai Raikar (WPI ID: 181758784)

Solution to Question 1:

All the codes for this Python and Numpy Warm-up Exercises are given in python file. Respective comments are also written in the code. Please check.

Solution to Question 2(a):

Required age regressor is trained using Linear Regression. **(Find codes and comments in python file)**

Below formula is used for finding weights w in linear regression (X_tr, Y_tr) function.

$$\mathbf{w} = \left(\mathbf{X}\mathbf{X}^\top\right)^{-1}\mathbf{X}\mathbf{y} \qquad \text{where } \mathbf{X} = \begin{bmatrix} | & & | \\ \mathbf{x}^{(1)} & \dots & \mathbf{x}^{(n)} \\ | & & | \end{bmatrix}$$

Below formula is used for finding mean square error.

$$f_{\text{MSE}}(\mathbf{w}) = \frac{1}{2n}\sum_{i=1}^{n}(\hat{y}^{(i)} - y^{(i)})^2$$

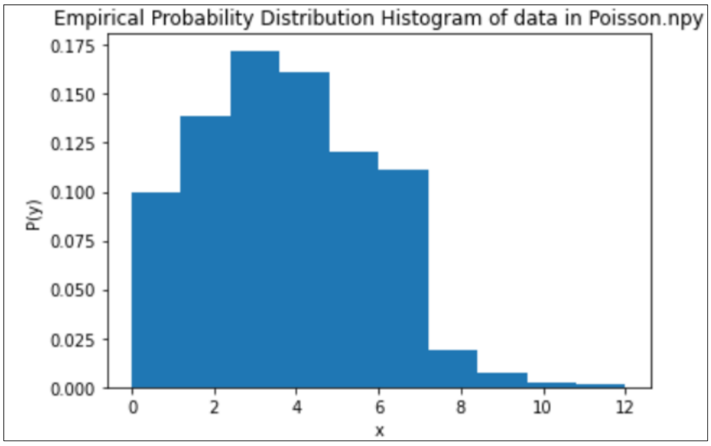
Mean Square Error (fMSE) is computed for both training set and testing set separately. The results are:

a. cost fMSE on the training set Dtr = 50.465284361134664
b. cost fMSE on the testing set Dte = 268.7927887195447

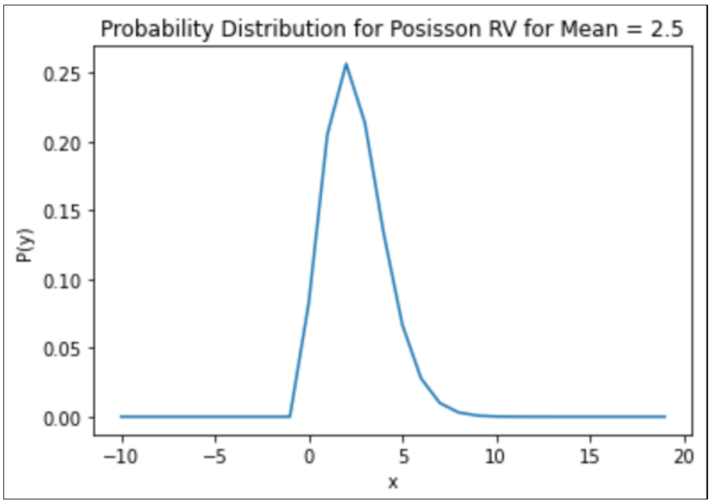
Solution to Question 3(a):

Required Probability Distributions are plotted using functions matplotlib.pyplot.hist with density=True and scipy.stats.poisson. Please check the below graphs.

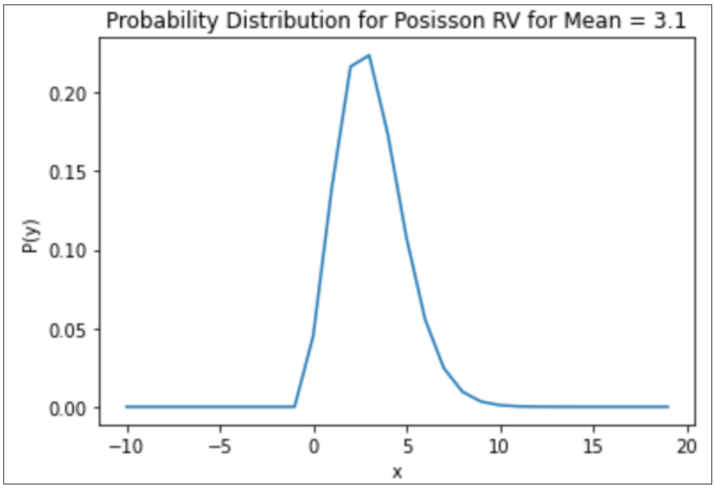
Empirical Probability Distribution of the Data included in PoissonX.npy:



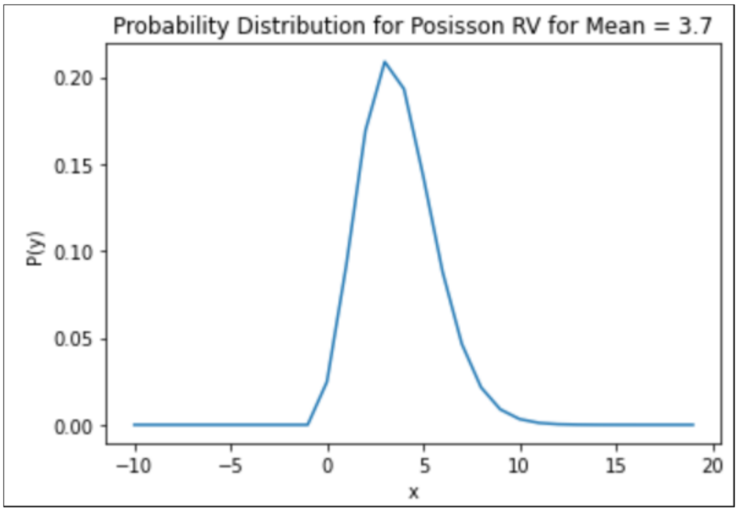
Probability Distributions of a Poisson random variable with rate parameter 2.5:



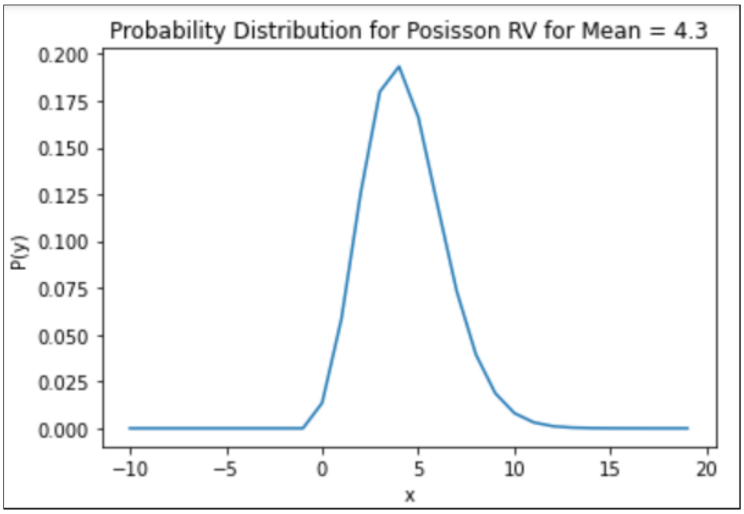
Probability Distributions of a Poisson random variable with rate parameter 3.1:



Probability Distributions of a Poisson random variable with rate parameter 3.7:



Probability Distributions of a Poisson random variable with rate parameter 4.3:



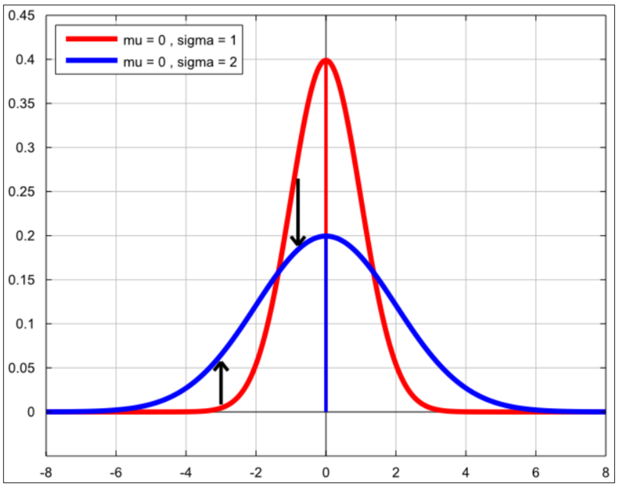
By visually inspecting above plots, we can infer that probability distribution of a Poisson random variable with rate parameter 3.7 is most consistent with the empirical distribution of the data.

Solution to Question 3(b):

Given function for normal distribution is:

$$P(y \mid x) = \mathcal{N}\left(\mu = x^2, \sigma^2 = \left(2 - \frac{1}{1 + e^{-x^2}}\right)^2\right)$$

In general, the graph of normal distribution is:



(i) By looking at above graph, we can tell that corresponding value of y will be larger if the standard deviation is less. Therefore, to have the larger value of y, we need our variance (σ^2) to be smaller.

As per our given normal distribution function, we can see that if we increase the value of x, the value of variance is decreasing. (i.e. x↑, σ^2↓)

Explanation: When x↑, (1 + e−x2) ↓

Therefore, (1/(1 + e−x2)) ↑

Therefore, ((2- (1/(1 + e−x2)))^2) ↓

(i.e. x↑, σ^2↓)

Hence, with the large magnitude of x, corresponding value of y tend be larger.

(ii) By looking at above graph, we can tell that the graph is more bulky with large value of variance. Therefore, uncertainty in the value of y is directly correlated to variance. Hence, to have the larger value of uncertainty, we need our variance (σ^2) to be larger.

As per our given normal distribution function, we can see that if we decrease the value of x, the value of variance is increasing. (i.e. x↓, σ^2↑)

Explanation: When x↓, (1 + e−x2) ↑

Therefore, (1/(1 + e−x2)) ↓

Therefore, ((2- (1/(1 + e−x2)))^2)↑

(i.e. x↓, σ^2↑)

Hence, with the small magnitude of x, uncertainty in the corresponding value of y tend be larger.

Solution to Question 4:

Please find all the required 4 proofs in below part of PDF (Scanned copy in legible handwriting)

Q.4 (a) To prove:

$$\nabla_x (x^T a) = \nabla_x (a^T x) = a$$

where, $\nabla_x = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$, $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, $y = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$

Proof:

$$\nabla_x (x^T a)$$

$$= \nabla_x [x_1 \ x_2 \ \ldots, \ x_n] \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} -$$

$$= \nabla_x [x_1 a_1 + x_2 a_2 + \ldots x_n a_n]$$

... (which will be the same result for $\nabla_x (a^T x)$)

$$= \begin{bmatrix} \frac{\partial [x_1 a_1 + x_2 a_2 + \ldots + x_n a_n]}{\partial x_1} \\ \frac{\partial [x_1 a_1 + x_2 a_2 + \ldots + x_n a_n]}{\partial x_2} \\ \vdots \\ \frac{\partial [x_1 a_1 + x_2 a_2 + \ldots + x_n a_n]}{\partial x_n} \end{bmatrix}$$

$$= \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = a$$

$$\boxed{\therefore \nabla_x (x^T a) = \nabla_x (a^T x) = a}$$

Hence, proved.

Q.4] (b) To prove:

$$\nabla_x (x^T A x) = (A + A^T) x$$

where, $\nabla_x = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$, $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, $A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} \\ \vdots \\ a_{n1} & \cdots & \cdots & a_{nn} \end{bmatrix}$

Proof :

$$\nabla_x \left(x^T A x\right)$$

$$= \nabla_x \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \nabla_x \begin{bmatrix} \sum_{i=0}^{n} x_i a_{i1} & \sum_{i=1}^{n} x_i a_{i2} & \cdots & \sum_{i=1}^{n} x_i a_{in} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \nabla_x \left[x_1 \sum_{i=0}^{n} x_i a_{i1} + x_2 \sum_{i=1}^{n} x_i a_{i2} + \cdots + x_n \sum_{i=1}^{n} x_i a_{in} \right]$$

$$= \begin{bmatrix} 2a_{11}x_1 + (a_{12}+a_{21})x_2 & \cdots\cdots & +(a_{1n}+a_{n1})x_n \\ (a_{12}+a_{21})x_1 + 2a_{22}\,x_2 & \cdots\cdots & + \\ \vdots & \vdots & \\ \vdots & \vdots & \\ (a_{1n}+a_{n1})x_1 + & \cdots\cdots & +2a_{nn}x_n \end{bmatrix}$$

$$= \begin{bmatrix} 2a_{11} & (a_{12}+a_{21}) & \cdots\cdots & (a_{1n}+a_{n1}) \\ (a_{12}+a_{21}) & 2a_{22} & \cdots & \vdots \\ \vdots & \vdots & & \vdots \\ (a_{1n}+a_{n1}) & \cdots\cdots & & 2a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \left(\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & \vdots \\ \vdots & & & \\ a_{n1} & \cdots & \cdots & a_{nn} \end{bmatrix} + \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & & & \vdots \\ \vdots & & & \\ a_{1n} & & & a_{nn} \end{bmatrix} \right) \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \left(A + A^T\right) x$$

$$\boxed{\therefore\ \nabla_x \left(x^T A x\right) = \left(A + A^T\right) x}$$ ... Hence proved.

Q.4] (c) To prove:

$$\nabla_x (x^T A x) = 2Ax$$

where A is any $n \times n$ symmetrix Matrix

Proof:

As per the proof of Q.4](b), we know that

$$\nabla_x (x^T A x) = (A + A^T) x$$

But, when A is a symmetric matrix,

$\Rightarrow A = A^T$

$$\therefore \nabla_x (x^T A x) = (A + A^T) x = 2Ax$$

$$\Rightarrow \boxed{\nabla_x (x^T A x) = 2Ax}$$

... Hence, proved.

Q.4] (d) To prove:

$$\nabla_x \left[(Ax+b)^T (Ax+b) \right] = 2A^T (Ax+b)$$

where, x is column vector

b is const. column vector

A is $n \times n$ symmetric matrix

Proof:

$$\therefore \nabla_x \left[(Ax+b)^T (Ax+b) \right]$$

$$= \nabla_x \left[x^T A^T (Ax+b) + b^T (Ax+b) \right]$$

$$= \nabla_x \left[x^T A^T A x + x^T A^T b + b^T A x + b^T b \right]$$

$$= \nabla_x (x^T (A^T A) x) + \nabla (x^T (Ab)) + \nabla ((b^T A^T) x)$$

.... ($\because A = A^T$)

∴ By using the proofs in Q.4 a, b, c :

$$= 2A^TAx + Ab + Ab + 0$$

$$= 2A(Ax+b)$$

$$= 2A^T(Ax+b) \quad \text{.... (as } A = A^T\text{)}$$

$$\boxed{\therefore \nabla_x\left[(Ax+b)^T(Ax+b)\right] = 2A^T(Ax+b)}$$

... Hence, proved