CSC 322: Machine Learning

Homework 3

Question 1 [8 points]:

A multivariate linear regression model has been built to predict the **heating load** in a residential building on the basis of a set of descriptive features describing the characteristics of the building. Heating load is the amount of heat energy required to keep a building at a specified temperature, usually 65° Fahrenheit during the winter regardless of outside temperature. The descriptive features used are the overall surface area of the building, the height of the building, the area of the building's roof, and the percentage of wall area in the building that is glazed. This kind of model would be useful to architects or engineers when designing a new building.[1] The trained model is

$$\begin{aligned}\text{HEATING LOAD} = &-26.030 + 0.0497 \times \text{SURFACE AREA} \\ &+ 4.942 \times \text{HEIGHT} - 0.090 \times \text{ROOF AREA} \\ &+ 20.523 \times \text{GLAZING AREA}\end{aligned}$$

Use this model to make predictions for each of the query instances shown in the following table.

ID	SURFACE AREA	HEIGHT	ROOF AREA	GLAZING AREA
1	784.0	3.5	220.5	0.25
2	710.5	3.0	210.5	0.10
3	563.5	7.0	122.5	0.40
4	637.0	6.0	147.0	0.60

Question 2 [42 points (12+12+6+12)]:

You have been hired by the European Space Agency to build a model that predicts the amount of oxygen that an astronaut consumes when performing five minutes of intense physical work. The descriptive features for the model will be the age of the astronaut and their average heart rate throughout the work. The regression model is

$$\text{OXYCON} = \mathbf{w}[0] + \mathbf{w}[1] \times \text{AGE} + \mathbf{w}[2] \times \text{HEARTRATE}$$

The table that follows shows a historical dataset that has been collected for this task.

ID	OXYCON	AGE	HEART RATE	ID	OXYCON	AGE	HEART RATE
1	37.99	41	138	7	44.72	43	158
2	47.34	42	153	8	36.42	46	143
3	44.38	37	151	9	31.21	37	138
4	28.17	46	133	10	54.85	38	158
5	27.07	48	126	11	39.84	43	143
6	37.85	44	145	12	30.83	43	138

(a) Assuming that the current weights in a multivariate linear regression model are $\mathbf{w}[0] = -59.50$, $\mathbf{w}[1] = -0.15$, and $\mathbf{w}[2] = 0.60$, make a prediction for each training instance using this model.

(b) Calculate the sum of squared errors for the set of predictions generated in Part (a).

(c) Assuming a learning rate of 0.000002, calculate the weights at the next iteration of the gradient descent algorithm.

(d) Calculate the sum of squared errors for a set of predictions generated using the new set of weights calculated in Part (c).

Question 3 [15 points]:

A multivariate logistic regression model has been built to predict the propensity of shoppers to perform a repeat purchase of a free gift that they are given. The descriptive features used by the model are the age of the customer, the socioeconomic band to which the customer belongs (a, b, or c), the average amount of money the customer spends on each visit to the shop, and the average number of visits the customer makes to the shop per week. This model is being used by the marketing department to determine who should be given the free gift. The weights in the trained model are shown in the following table.

Feature	Weight
Intercept ($\mathbf{w}$ [0])	-3.82398
AGE	-0.02990
SOCIOECONOMIC BAND B	-0.09089
SOCIOECONOMIC BAND C	-0.19558
SHOP VALUE	0.02999
SHOP FREQUENCY	0.74572

Use this model to make predictions for each of the following query instances.

ID	AGE	SOCIOECONOMIC BAND	SHOP FREQUENCY	SHOP VALUE
1	56	b	1.60	109.32
2	21	c	4.92	11.28
3	48	b	1.21	161.19
4	37	c	0.72	170.65
5	32	a	1.08	165.39

Question 4: [35 points (21+14)]

A multivariate logistic regression model has been built to diagnose breast cancer in patients on the basis of features extracted from tissue samples extracted by biopsy.[3] The model uses three descriptive features—MITOSES, a measure of how fast cells are growing; CLUMPTHICKNESS, a measure of the amount of layering in cells; and BLANDCHROMATIN, a measure of the texture of cell nuclei—and predicts the status of a biopsy as either *benign* or *malignant*. The weights in the trained model are shown in the following table.

Feature	Weight
Intercept ($\mathbf{w}[0]$)	−13.92
MITOSES	3.09
CLUMPTHICKNESS	0.63
BLANDCHROMATIN	1.11

(a) Use this model to make predictions for each of the following query instances.

ID	MITOSES	CLUMP THICKNESS	BLAND CHROMATIN
1	7	4	3
2	3	5	1
3	3	3	3
4	5	3	1
5	7	4	4
6	10	4	1
7	5	2	1

(b) The following are the ground truth labels for the query instances from Part (a).

$\mathbf{d}_1$	$\mathbf{d}_2$	$\mathbf{d}_3$	$\mathbf{d}_4$	$\mathbf{d}_5$	$\mathbf{d}_6$	$\mathbf{d}_7$
benign	*benign*	*malignant*	*benign*	*malignant*	*malignant*	*benign*

i. Using the ground truth labels, calculate the **squared error** loss for each query instance (assume that *benign* = 0 and *malignant* = 1).