Error based Learning (Regression & SVM) Lab Sheet

0. Review notebooks `06 Regression; 06 Gradient Descent; 06 SVM`

1. You have been hired by the ESA 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 heart rate throughout the work. The regression model is:

   `OxyCon = w[0] + w[1] x Age + w[2] x HeartRate`

   This table shows data collected for this task:

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

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

   (b) Calculate the sum of squared errors for this set of predictions.

   (c) Assuming a learning rate of 0.000002, calculate the weights at the next iteration of the gradient descent algorithm and the loss at this point in the algorithm. Is it better or worse?

2. A multivariate logistic regression model has been built to predict the propensity of shoppers to repeat purchase a free gift they are given. The features used by the model are the customer age, the socioeconomic band they belong to (either a, b or c), the average amount of money they spend on

each visit and the average number of visits per week they make to the shop. The model is being used by the marketing department to determine who should be given a free gift.
The weights in the trained model are shown in the following table:

Feature	Weight
Intercept (w[0])	-3.83398
Age	-0.02990
Socioeconomic Band B	-0.09089
Socioeconomic Band C	-0.19558
Average Spend	0.02999
Frequency	0.74572

(a) Explain how there are 6 feature weights in the model given that there are only four original features used.
(b) Use this model to make predictions for the following query instances, assume that the positive class is '*yes*':

Query	Age	Socioeconomic Band	Average Spend	Frequency
Q1	56	b	109.32	1.6
Q2	37	c	170.65	0.72
Q3	32	a	165.39	1.08

3 A support vector machine has been built to predict whether a patient is at risk of cardiovascular disease. In the dataset used to train the model, the two target levels are *high risk* (the positive class) and *low risk* (the negative class). The support vectors are shown in the table below where all descriptive variables have been standardised.

In the model the value of w0 is -0.022, and the values of the α parameters are {1.68, 0.23, 0.20, 1.71}

SV	Age	BMI	Blood Pressure	Risk
s1	-0.45	0.01	0.22	low risk
s2	-0.28	-0.52	0.36	low risk
s3	0.37	0.09	-1.08	high risk
s4	0.55	0.22	0.21	high risk

What predictions would this model make for the following query instances?

ID	Age	BMI	Blood Pressure
q1	-0.89	-0.34	0.55
q2	0.45	0.49	-0.47

4. Assess the performance of Linear Regression models on the `bike_sharing.csv` dataset that use:

1. `temp` as the only input variable
   (`X = bikes_df[['temp']].values`)
2. `hum` as the only input variable
3. all features except `casual`, `registered`, `instant` and `dteday`
   (as set up in notebook `06 Regression`)

Use a holdout test set.

You may use LinearRegression or SGDRegressor.

Score performance using R2, MSE, MAE and RMSE.