BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
WORK INTEGRATED LEARNING PROGRAMMES DIVISION
POST GRADUATE PROGRAMME IN ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING
COMPREHENSIVE EXAM – REGRESSION REGULAR

Course Code	**: PCAM ZC211**	**Course Title**	**: REGRESSION**
Nature of Exam	**: Closed Book**	**Marks**	**: 30**
Duration	**: 2 Hours**		
Date of Exam	**: 08/08/2020 (FN)**		

Q1 [2+2 = 4 Marks]

Give practical example of applying:

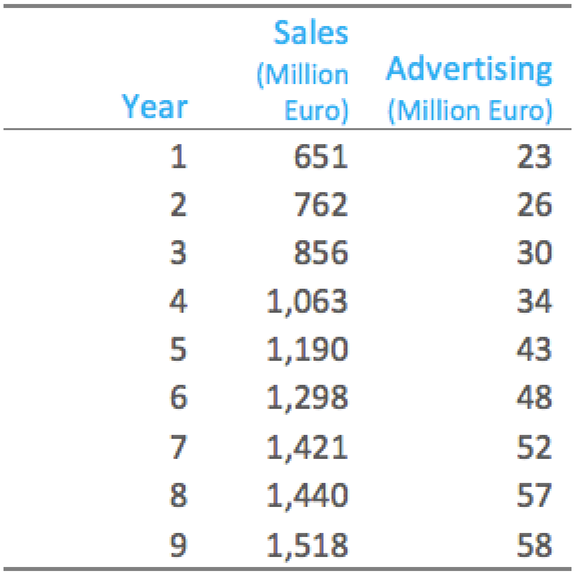
- simple or univariate linear regression
- multiple linear regression

Explain with formula, input and output variables.

Answer

Simple or Univariate Linear Regression:

The table below shows some data from the early days of the Italian clothing company Benetton. Each row in the table shows Benetton’s sales for a year and the amount spent on advertising that year. In this case, our outcome of interest is sales—it is what we want to predict. If we use advertising as the predictor variable, linear regression estimates that Sales = 168 + 23 Advertising. That is, if advertising expenditure is increased by one million Euro, then sales will be expected to increase by 23 million Euros, and if there was no advertising we would expect sales of 168 million Euros.



Year	Sales (Million Euro)	Advertising (Million Euro)
1	651	23
2	762	26
3	856	30
4	1,063	34
5	1,190	43
6	1,298	48
7	1,421	52
8	1,440	57
9	1,518	58

Multiple Linear Regression:

Returning to the above example, we can include year variable in the regression, which gives the result that Sales = 323 + 14 Advertising + 47 Year. The interpretation of this equation is that every extra million Euro of advertising expenditure will lead to an extra 14 million Euro of sales and that sales will grow due to non-advertising factors by 47 million Euro per year.

Q 2. [8 Marks]

Given a dataset with 1 billion points, will vanilla gradient descent or Mini Batch (a form of SGD) create a better model for us with respect to time and iterations?
Explain mathematically

Ans:

Mini-Batch gradient descent (MBGD) leads to faster convergence. It does not require storing all training data in memory (good for large training data)

Batch gradient also works well for large training data and has better model parameter estimation compared to SGD.

Vanilla gradient descent might work well for lower training data.

For a million samples in your dataset, so if you use a typical Gradient Descent optimization technique, you will have to use all of the one million samples for completing one iteration while performing the Gradient Descent, and it has to be done for every iteration until the minima is reached. Hence, it becomes computationally very expensive to perform.

This problem is solved by Mini-Batch Gradient Descent. In MBGD, it does not use all of the training samples. it uses more than one sample which can be computationally useful and faster than SGD.

Q 3. [6 Marks]

Let us assume that we build 2 models on the same data with some different variables in both the models.

Here are the findings below

Metric	Model 1	Model 2
MAE	900	1200
R Squared	68	71
RMSE	5000	5690
F statistic	23	27

Explain which model is the better and why? [6 marks]

Answer

Model 2 is better because of the following reasons:

1) MAE and RMSE only work when the data on which the model is built satisfies all the assumptions of Linear regression. We do not know about the assumptions and data so these metrics cannot be trusted
2) R squared of model 2 is better hence it becomes an automatic choice along with F statistic supporting the same

Q 4. [6 Marks]

Body weight, calorie intake, fat intake and age have an influence on the blood cholesterol level in various subjects. The linear regression analysis can then show whether the independent variables have an effect on the blood cholesterol level (dependent variable).
A.) Write down the steps to perform Forward Feature Selection on the given data. [3 Marks]
B.) Write down the steps to perform Backward Feature Selection on the given data. [3 Marks]

Answer

a.) Forward Feature Selection

1. Let M_0 denote the null model , which contains no predictors.
   This model simply predicts the sample mean for each observation.

2. For k = 0, 1, 2, . . .D-1:

   (a) Consider all D − k models that augment the predictors in Mk with one additional predictor.

   (b) Choose the best among these D − k models, and call it M_K+1.
   Here best is defined as having smallest RSS on training dataset (or highest R_2).

3. Select a single best model from among M_0, M_1 , . . . , M_D having the smallest RSS on testing error (or equivalently largest R_2).

b.) Backward Feature Selection

1. Let M_D denote the full model, which contains all p predictors.
2. For k = D, D − 1, . . . , 1:
    (a) Consider all k models that contain all but one of the predictors in M_K, for a total of k − 1 predictors.

    (b) Choose the best among these k models, and call it M_K−1. Here best is defined as having smallest RSS on training dataset (or highest R_2).

3. Select a single best model from among M_0, M_1 . . ,M_D having smallest RSS on testing error (or equivalently largest R_2).

Q 5. [6 Marks]

Dataset with n records. Let the independent variables be x with 10 features and target variable as y. Suppose we use a linear regression method to model this data. Explain mathematically

A) Ridge Regression
B) Lasso Regression

Answer

a.) Ridge Regression

(x1y1, x2y2,…………. xnyn)

$$\min \frac{1}{2} \sum_{n=1}^{N} [(w_0 + w_1 x\frac{1}{n} + w_2 x\frac{2}{n} + \ldots .. + w_{10} x\frac{10}{n}) - y_n]^2$$

Such that $\sum_{m=1}^{10} {w_m}^2 \leq s$

Above constrained optimization problem can be converted to unconstrained optimization problem using regularization parameter λ

$$\min\{ \frac{1}{2} \sum_{n=1}^{N} [(w_0 + w_1 x\frac{1}{n} + w_2 x\frac{2}{n} + \ldots .. + w_{10} x\frac{10}{n}) - y_n]^2 + \lambda \sum_{m=1}^{10} {w_m}^2 \}$$

Typically 0 < λ < 1
Pickup λ where the test error is minimum

b.) Lasso Regression

(x1y1, x2y2,………… xnyn)

$$\min \frac{1}{2} \sum_{n=1}^{N} [(w_0 + w_1 x \frac{1}{n} + w_2 x \frac{2}{n} + \ldots\ldots + w_{10} x \frac{10}{n}) - y_n]^2$$

Such that $\sum_{m=1}^{10} |w_m| \leq s$

Above constrained optimization problem can be converted to unconstrained optimization problem using regularization parameter λ

$$\min\{ \frac{1}{2} \sum_{n=1}^{N} [(w_0 + w_1 x \frac{1}{n} + w_2 x \frac{2}{n} + \ldots\ldots + w_{10} x \frac{10}{n}) - y_n]^2 + \lambda \sum_{m=1}^{10} |w_m| \}$$

****