**Department of Computer Science and Engineering**

|  |  |
| --- | --- |
| **Course Code:CSE422** | **Credits:3** |
| **Course Name: Artificial Intelligence** | **Prerequisite:** CSE111, CSE221 |

**Lab 04**

**Regression**

1. **Lab Overview:**

The students will learn the regression analysis for the machine learning approach.

1. **Learning Objective:**
   1. What is regression analysis?
   2. How to solve problem using linear regression.
   3. How to build a model for classification using logistic regression.
2. **Lesson Fit:**

There is pre-requisite to this lab: CSE111, CSE221. You should have intensive Programming Knowledge and capability to understand algorithms.

1. **Acceptance and Evaluation**

Students will show the output using different datasets and python code. They will be marked according to their lab performance. The main evaluation criteria will be based on project report and demonstration.

1. **Learning Outcome:**

After this lab, the students will be able to:

* 1. Understand the basic regression analysis.
  2. How to apply linear regression for prediction.
  3. How logistic regression can help to build a model for classification.

1. **Activity Detail**

* **Hour: 1-1.5**

**What is Regression?**

Regression analysis is used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. Besides, regression analysis is widely used for prediction, forecasting, and classification where its use has considerable connect with the field of machine learning.

**What is the difference between Linear and Logistic regression?**

Linear and Logistic regression are the most basic form of regression which are commonly used in Machine Learning. The essential difference between these two is that Logistic regression is used when the dependent variable is binary in nature (i.e., discreet valued output of a function). In contrast, linear regression is used when the dependent variable is continuous (i.e., continuous valued output of a function) and nature of the regression line is linear.

**Details about Linear Regression**

* Linear regression is based on Least Square Estimation which says regression coefficients should be chosen in such a way that it minimizes the sum of the squared distances of each observed response to its fitted value.
* A linear model is a sum of weighted variables that predicts a target output value given an input data instance.
* Linear Regression is an example of a linear model

Input feature vector: x

Predicted output:

Parameters to estimate:

**Linear Regression Model with one variable**

Input feature vector: x

Predicted output:

Parameters to estimate:

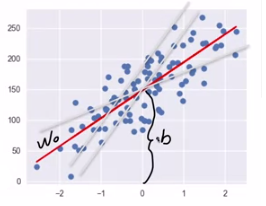


Fig. 1: Linear model

**Linear Regression: Least-Squares**

Find w and b that minimized the mean squared error of the model: the sum squared differences between predicted target and actual target values. But problem is that there is no parameters to control model complexity.

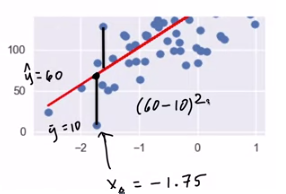


Fig. 2: Linear Regression: Least-Squares

* How are Linear Regression parameters w, b Estimated?
* Parameters are estimated from training data.
* There are many different ways to estimate w and b
  + Different methods correspond to different “fit” criteria and goals and ways to control model complexity
* The Learning algorithm finds the parameters that optimize an objective function, typically to minimize some kind of loss function of the predicted target values Vs. actual target values.

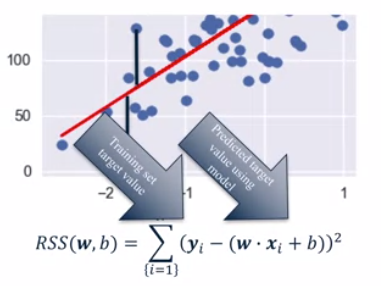
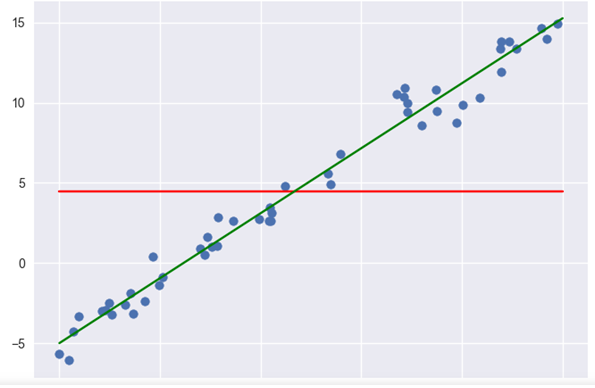


Fig. 3: Optimization of the objective function

Any regression model that we fit is compared to this baseline model to understand it’s goodness of fit.

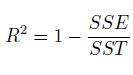
In other words **R-squared simply explains how good is your model when compared to the baseline model**



The red line in the above figure is the baseline model that always predicts mean of observed response of dependent variable Y as the value of Y irrespective of the value of the independent variables.

And the green line is our fitted model which makes use of independent variables to predict the value of dependent variable Y.

R-squared is given by



Where**SSE is the sum of squared errors of our regression model**

https://cdn-images-1.medium.com/max/1600/1*nE6tXrxjU78-y21PrxqrSA.png

And **SST is the sum of squared errors of our baseline model.**

https://cdn-images-1.medium.com/max/1600/1*hF-725hVaU9nZf2WgORoPg.png

So **R-squared can take value between 0 and 1 where values closer to 0 represent a poor fit while values closer to 1 represent a perfect fit .**

* **Hour: 1.5-2.0**

**Details about Logistic Regression**

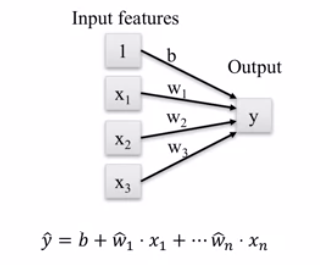
****

Fig. 4: Linear Regression

**Linear model for classification: Logistic Regression**

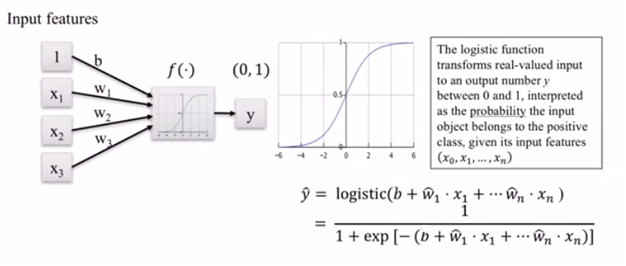
****

Fig. 5: Linear model for classification: Logistic Regression

**Logistic Regression for Binary Classification**

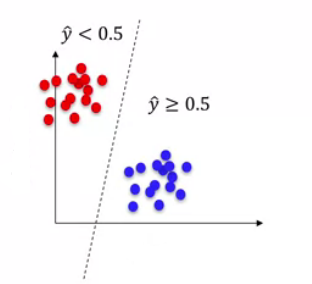


Fig. 6: Logistic Regression for Binary Classification

* **Hour: 2.0-3.0**

(It is Not a Group Task, Try Individually)

**Task 01:** Mark 5 **Time: 30 minutes**

The code for linear regression is given in the TSR. Consider the code for calculation

1. Coefficient value (w)
2. Intercept value (b)

**Evaluation Process (VIVA):** You have to explain your program to the Lab Instructor

**Task 02:** Mark 5 **Time: 30 minutes**

The code for logistic regression is give in the TSR where you will get the classification output for sepal length and sepal width. Your task is to consider petal length and petal width and then classify them.

**Evaluation Process (VIVA):** You have to explain your program to the Lab Instructor