Linear Regression

Saurabh Burewar (B18CSE050)

Dataset

Link to the dataset: https://www.kaggle.com/c/titanic/data

The dataset shows different attributes of the passengers of Titanic and whether they survived or not. I have only used "train.csv" and divided this file itself into train and test sets with a 70:30 size ratio. It contains the following features -

Features	**Definition**	**Key**
Pclass	Ticket class	1=Upper, 2=Middle, 3=Lower
Sex	Sex	male, female
Age	Age in years	
Sibsp	# of siblings/spouses aboard the Titanic	
Parch	# of parents/children aboard the Titanic	
Ticket	Ticket number	
Fare	Passenger fare	
Cabin	Cabin number	

Embarked	Port of embarkation	C = Cherbourg,
		Q = Queenstown,
		S = Southampton

Survived (Target)	Survived or not	0 = No, 1 = Yes
sex_factor (added by me)	Factorize the feature "Sex" to convert it to numerical	0 = male, 1 = female
em_factor (added by me)	Factorize the feature	0 = S = Southampton,

	“Embarked” to convert it to numerical	1 = C = Cherbourg, 2 = Q = Queenstown

🟨 Features used in the training of the models

🟥 Target variable

Dataset Processing

- The feature “cabin” has a lot of missing values (687 out of 891), so I have dropped this column from the dataset.
- The feature “Age” also has missing values so I have dropped all rows with missing values in them.
- I factorized the column “Sex” to create another column “sex_factor” which gives numerical representation of the column “Sex”.
- I factorized the column “Embarked” to create another column “em_factor” which gives numerical representation of the column “Embarked”.

Features used

I have used the features marked yellow in the above table. So, the features used are -

X = 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare', 'sex_factor', 'em_factor'
Y = ‘Survived’

Classification

The task here is to predict whether a passenger will survive based on the features. So, the target variable is “Survived” which shows 0 for not survived and 1 for survived. Therefore, it is a case of binary classification.

Performance

Linear Regression using Pseudo Inverse

Accuracy of the model: 59.813%

Linear Regression using Gradient descent

Accuracy of the model: 57.944%